Convergence of sampling in protein simulations

Berk Hess
Department of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands
(Received 3 August 2001; published 1 March 2002)

With molecular dynamics protein dynamics can be simulated in atomic detail. Current computers are not fast enough to probe all available conformations, but fluctuations around one conformation can be sampled to a reasonable extent. The motions with the largest fluctuations can be filtered out of a simulation using covariance or principal component analysis. A problem with this analysis is that random diffusion can appear as correlated motion. An analysis is presented of how long a simulation should be to obtain relevant results for global motions. The analysis reveals that the cosine content of the principal components is a good indicator for bad sampling.

DOI: 10.1103/PhysRevE.65.031910 PACS number(s): 87.14.Ee, 87.15.Aa, 87.15.He

I. INTRODUCTION

Proteins are complex objects with motions in a large range of length and time scales. In a classical description of protein dynamics the fluctuations range from bond and angle vibrations of tenths of Ångstroms on the femtosecond time scale to (un)folding of the whole protein on a time scale of seconds. Currently, there are no experimental techniques that can follow the detailed dynamics of proteins in time. This leaves molecular dynamics as the only tool to study this regime. With the current computers simulations of proteins are limited to hundreds of nanoseconds.

In a trajectory of a protein the Cartesian coordinates of the atoms contain a mixture of fast and slow modes of motion. Covariance or principal component analysis, which has also been termed quasiharmonic analysis [1], ''molecule optimal dynamic coordinates'' [2,3] and ''essential dynamics'' [4], can be used to separate these modes of motion based on amplitude. In a protein a few modes contain more than half of the total fluctuation in the system. For a long simulation the first few modes usually describe global, collective motions. But one has to be careful when interpreting the results of such an analysis, since random diffusion can produce patterns that resemble collective behavior. There are several examples in the literature where cosine-shaped principal components have been interpreted as transition of the protein from one state to another. Recently this has been proven that such cosines also emerge from random diffusion without potential [5]. For short protein simulations the first few principal components will always be caused by random diffusion, since the time is too short to reach barriers in the potential. Although this diffusive behavior can be of interest, the final direction and amplitude of the motions cannot be estimated, because of the inherent properties of random diffusion. This paper will assess when effects of the potential become visible in the first few principal components. This determines the minimum simulation length that is required to draw any conclusions on global motions in the protein. First principal component analysis will be described in more detail.

Most degrees of freedom of a protein will be highly constrained due to bonded interactions between the atoms. The protein moves in a N-dimensional space, where N is three times the number of atoms. Only a few of these degrees of freedom will contribute significantly to the global fluctuations of the protein. Mass-weighted covariance or principal component analysis can be used to find these degrees of freedom. This is the equivalent of normal mode analysis at nonzero temperatures [6,7]. The analysis can be applied on any high-dimensional set of coordinates $\mathbf{x}(t)$. After a translational and rotational fit of all structures to a reference structure, the mass-weighted covariance matrix C is built and diagonalized

$$C_{ij} = M_{ii}^{1/2} \overline{(x_i(t) - \overline{x_i(t)}) M_{jj}^{1/2} (x_j(t) - \overline{x_j(t)})}, \quad (1)$$

$$C = R \,\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) R^T, \quad (2)$$

where an overline denotes averaging over time, M is a diagonal matrix containing the masses of the particles and N is three times the number of particles. The ith column of the rotation matrix R is the eigenvector or principal mode corresponding to eigenvalue λ_i. The eigenvalue is the mean square fluctuation in the direction of the principal mode. The projections of $\mathbf{x}$ on the eigenvectors are the principal components,

$$\mathbf{p}(t) = R^T M^{1/2} \mathbf{x}(t). \quad (3)$$

Principal component analysis is just a rotation of space, where $\mathbf{p}(t)$ are the new, massweighted, rotated coordinates. In the following sections the analysis will be applied to four molecular dynamics trajectories of proteins and a model system.

II. MOLECULAR DYNAMICS SIMULATIONS

Four molecular dynamics simulations of 40 ns each were performed on two different proteins in explicit solvent. The proteins are HPr, Histidine containing phosphocarrier protein, of 85 residues and $T4$-lysozyme of 164 residues. The simulations were performed with the Gromacs package [8], using the Gromos96 force field [9]. In all simulations the angle vibrations of the hydrogens in the protein were removed, by replacing the hydrogen atoms with interaction sites [10]. The hydrogen charges were fixed at the position of the minimum of the angle potentials. This procedure, together with the LINCS algorithm for constraining bonds [11]

enables the use of a time step of 4 fs. The temperature was coupled to 300 K, the pressure to 1 bar, using a Berendsen thermostat and barostat [12], with coupling times of 0.1 and 1 ps, respectively.

The starting structure for the HPr simulations was taken from protein data bank (PDB) entry 1poh [13]. The protein with 89 crystal waters was solvated in a truncated octahedron with a nearest image distance of 5.5 nm. The total number of SPC (simple point charge) water molecules [14] was 3841. After energy minimization two simulations of 40 ns were performed using different random initial velocities. A twin-range cutoff of 1.0/1.4 nm was used. Forces below 1.0 nm were updated every step, the forces between 1.0 and 1.4 nm and the neighbor list were updated every five steps.

The starting structure for the *T*4-lysozyme simulations was taken from PDB entry 2lzm [15]. The protein with 118 crystal waters was solvated in a rhombic dodecahedron with a nearest image distance of 7 nm. After energy minimization the system was neutralized by replacing eight waters with eight chlorine ions. The ions were inserted at the water oxygen position with the most favorable electrostatic potential, the potential was recalculated after every ion insertion. The total number of water molecules was 7156. After another energy minimization two simulations of 40 ns were performed, using different random initial velocities. A reaction field with a dielectric constant of 80 was used to prevent accumulation of ions at the cutoff. A twin-range cutoff of 1.0/1.5 nm was used.

For HPr the average root mean square deviation (RMSD) of C_α atoms over 5–40 ns with the pdb structure is 0.29 nm for the first and 0.23 nm for the second trajectory. After diverging initially, the two trajectories come close together. The RMSD between the final structures is 0.24 nm. The fold and most of the secondary structure stays intact during the simulations, the RMSD is mainly caused by a slight reorientation of the secondary structure elements with respect to each other. For lysozyme the average RMSD of C_α atoms over 5–40 ns with the pdb structure is 0.28 nm for the first and 0.32 nm for the second trajectory. The main motion is hinge bending of the two domains. In the first simulation the protein closes with respect to the pdb structure, in the second one it opens. The program DYNDOM [16] was used to quantify the rotation between the two structures from the first and second simulation that have the highest RMSD with respect to each other. DYNDOM reports a 71° rotation between two domains, which consist of residues 14-58,65-78 and residues 3-13,79-159, respectively. In Figs. 1 and 2 RMSD matrices are shown in which transitions between conformations can be seen easily. It is impossible to cluster the sampled conformations in a unique way. But roughly one could say that each simulation samples three conformations with a time span between 4 and 22 ns. The average RMSD between structures within each conformation is between 0.1 and 0.15 nm. The RMSD between structures from different conformations ranges from 0.2 to 0.4 nm.

A search was performed on all simulations for time intervals in which the protein seems to move around only one structure. Principal component analysis on the C_α atoms was used as a tool to find such intervals. Since all C_α atoms have

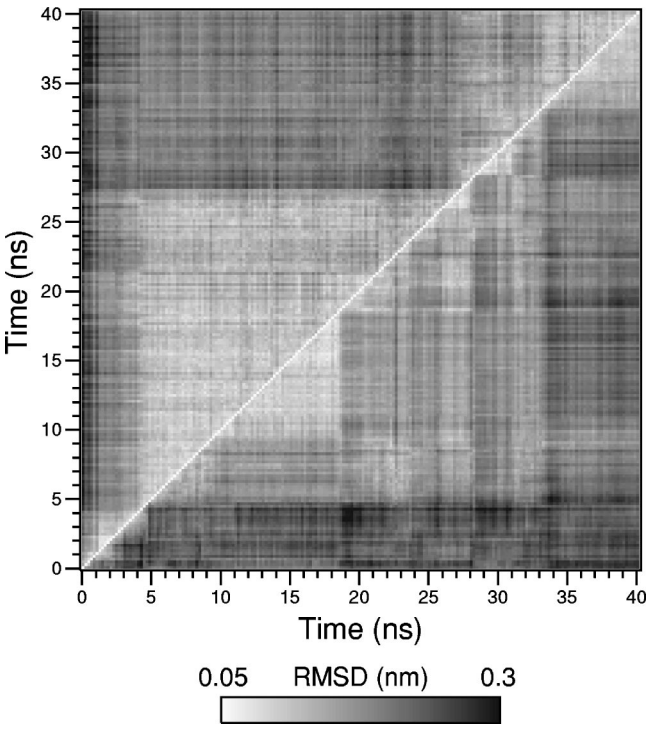

FIG. 1. Root mean square deviation (RMSD) matrix, showing the RMSD of the C_α atoms of each pair of structures in HPr simulation 1 (upper-left triangle) and HPr simulation 2 (lower-right triangle).

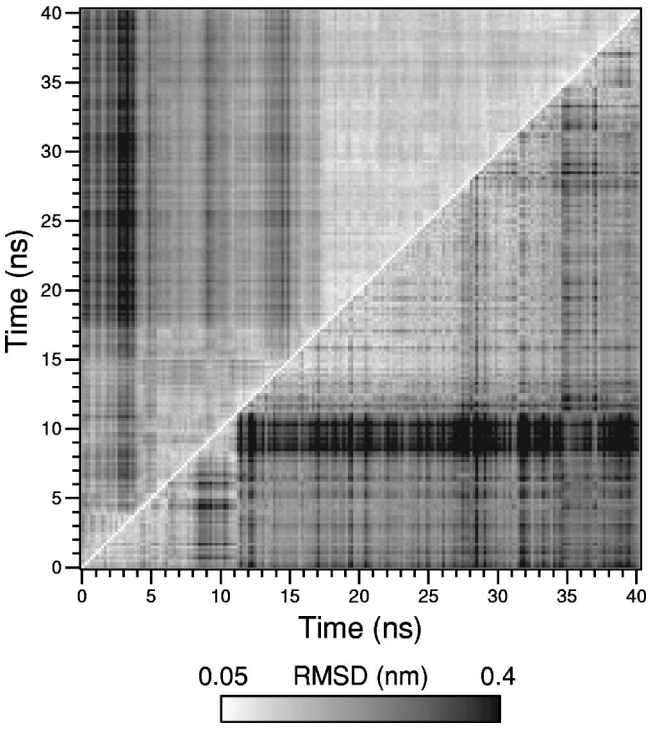

FIG. 2. Root mean square deviation (RMSD) matrix, showing the RMSD of the C_α atoms of each pair of structures in lysozyme simulation 1 (upper-left triangle) and lysozyme simulation 2 (lower-right triangle).

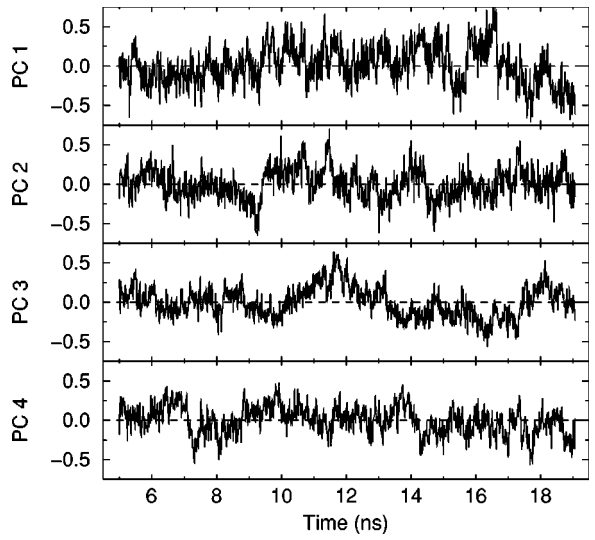

FIG. 3. First four principal components for the interval of HPr simulation 1. Unit: nm.

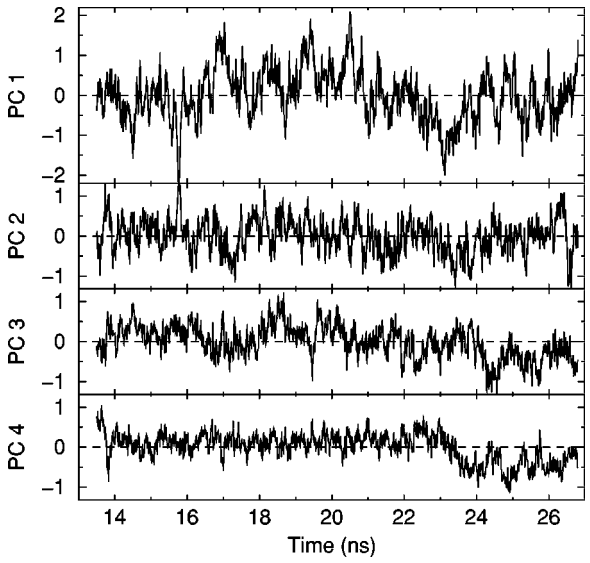

FIG. 4. First four principal components for the interval of lysozyme simulation 2. Unit: nm.

the same mass, a non-mass-weighted analysis was performed. The criterion for sampling around one structure was that the distribution of the principal components (PC's) was close to Gaussian. The relative deviation of the third and fourth central moments from those of a Gaussian distribution with the same mean and variance was used as a measure for similarity to a Gaussian distribution. A number of candidate intervals were identified by visual inspection of the RMSD matrices. From this set of intervals the two where the PC's were closest to Gaussian were selected. In the first HPr simulation interval 5–19.1 ns was found, the average deviation from Gaussian over all PC's, weighted with the widths of the distributions, is 7% and 3% for the third and fourth central moment, respectively. In the second lysozyme simulation interval 13.3–26.8 ns was found, the deviations are 11% and 7%. All structures were fitted to the first structure of the interval.

The eigenvalues depend on the principal component index as a power law with an exponent of $-4/3$ (not shown), except for the first nine eigenvalues of HPr, which have an exponent of about $-2/3$. The first four PC's for the time intervals are shown in Figs. 3 and 4. All PC's exhibit rapid fluctuations of the order of tens of picoseconds and slower fluctuations of the order of hundreds of picoseconds. Only the fourth PC of the lysozyme simulation shows significant ''nonrandom'' behavior with a jump between two states at 23 ns. Inertia effects are negligible, since the velocity autocorrelation functions of the PC's (not shown) have a negative minimum around 1 ps, which is an order of magnitude shorter than the correlation times of the PC's. The overdamped dynamics together with the Gaussian distribution of the PC's suggests the approximation of diffusion in a high-dimensional harmonic potential. In this approximation, the force constant for the harmonic potential of PC i is given by

$$k_i = \frac{k_B T}{\lambda_i}, \tag{4}$$

where k_B is the Boltzmann's constant, T is the temperature, and λ_i is the eigenvalue or mean square fluctuation of PC i. The friction coefficient is given by

$$\zeta_i = k_i \tau_i, \tag{5}$$

where τ_i is a correlation time. The autocorrelation functions of the PC's can be fitted well with a sum of two exponentials

$$f(t) = \lambda \left[(1-\beta) \exp\left(-\frac{t}{\tau_f} \right) + \beta \exp\left(-\frac{t}{\tau_s} \right) \right], \tag{6}$$

where τ_f and τ_s are the fast and slow correlation time, respectively. This implies two stochastic processes with white noise and time independent friction constants. For the first 30 PC's of HPr β is 0.51 on average, for lysozyme this is 0.50. Some of the β's of the first few PC's differ significantly from the average. A separate harmonic force constant and friction constant can be calculated for the fast and slow fluctuations of each PC. The obtained friction constants for the first 30 PC's of HPr and lysozyme are shown in Fig. 5. Although the friction constants vary by an order of magnitude, they show little systematic dependence on the PC index. Only the last 17 friction constants for the slow fluctuations of HPr are significantly higher than first 13. When the high friction constant of PC number 15 of HPr is discarded, the average ratio of the slow over the fast friction constant is 33 for both HPr and lysozyme. This analysis suggests that the global dynamics of both proteins is governed by a fast and a slow diffusion process, for which the diffusion constants are independent of the direction. In the harmonic approximation, the high-dimensional energy landscapes for both processes are almost identical. The harmonic force constants are close to linear with PC index.

To study the convergence of the sampling, both intervals were divided in up to 256 subintervals, with steps of a factor of 2. Principal component analysis was performed on each of these subintervals. This gives good statistics for the short

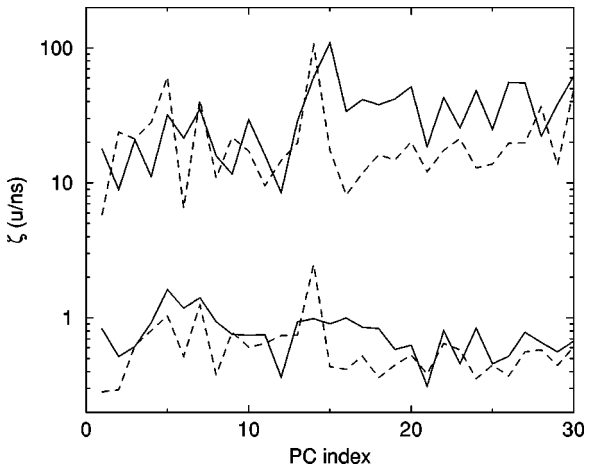

FIG. 5. Estimated friction constants for the fast and slow fluctuations of the first 30 PC's of HPr (solid lines) and lysozyme (dashed lines).

intervals and bad statistics for longer times. The increase of fluctuations with time can be seen in the root mean square fluctuation (Fig. 6), which is equal to the square root of the sum of eigenvalues divided by the number of atoms. The fluctuations of HPr seem to have leveled off at 14 ns, the fluctuations of lysozyme still increase at 14 ns, but with a smaller slope than at shorter times.

The overlap of the fluctuations can be used as a measure for the convergence of the sampled space. This can be done in terms of covariance matrices. The elements of the covariance matrix are proportional to the square of the displacement, so the square root of the matrix is required to examine the extent of sampling. The square root can be calculated from the eigenvalues λ_i and the eigenvectors, which are the columns $\mathbf{R}_i$ of the rotation matrix R. For a symmetric and diagonally dominant matrix A of size $N\times N$ the square root can be calculated as

$$A^{1/2}=R\,\mathrm{diag}(\lambda_1^{1/2},\lambda_2^{1/2},\ldots,\lambda_N^{1/2})R^T, \tag{7}$$

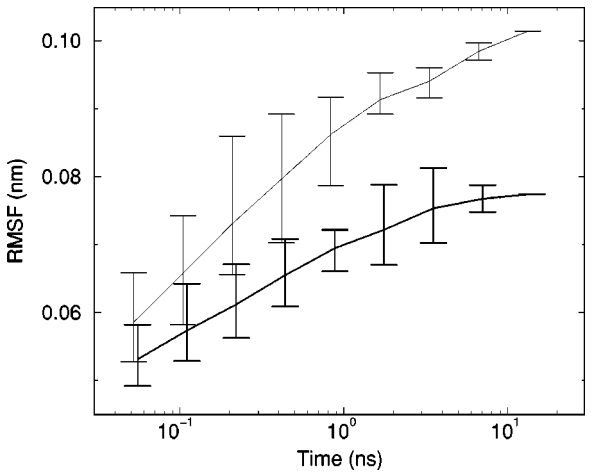

FIG. 6. Root mean square fluctuation (RMSF) of the C_α atoms as a function of the length of the subinterval for HPr (thick line) and lysozyme (thin line). The lines are averages over all subintervals, the error bars indicate the intervals containing 90% of the points.

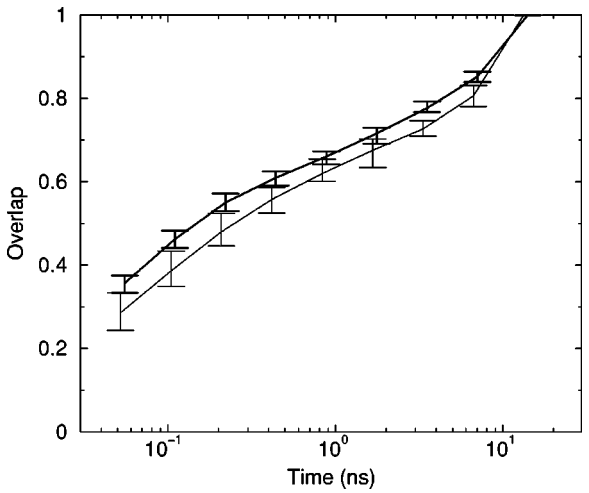

FIG. 7. Overlap [expression (11)] of the sampling of each subinterval with the sampling after 14 ns for HPr (thick line) and lysozyme (thin line). The lines are averages over all subintervals, the error bars indicate the intervals containing 90% of the points.

where diag() is a diagonal matrix. It can be verified easily that the product of this matrix with itself gives A. Now a difference d between covariance matrices A and B can be defined as follows:

$$d(A,B)=\sqrt{\mathrm{tr}[(A^{1/2}-B^{1/2})^2]} \tag{8}$$

$$=\sqrt{\mathrm{tr}(A+B-2A^{1/2}B^{1/2})} \tag{9}$$

$$=\left[\sum_{i=1}^{N}(\lambda_i^A+\lambda_i^B)-2\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\lambda_i^A\lambda_j^B}(\mathbf{R}_i^A\cdot\mathbf{R}_j^B)^2\right]^{1/2}, \tag{10}$$

where tr is the trace of a matrix. The overlap s as can now be defined as

$$s(A,B)=1-\frac{d(A,B)}{\sqrt{\mathrm{tr}\,A+\mathrm{tr}\,B}}. \tag{11}$$

The overlap is one if and only if matrices A and B are identical. It is zero when the sampled subspaces are completely orthogonal. This measure has several advantages over the commonly used subspace overlap, which is the overlap between the subspaces of the first n_A and n_B eigenvectors of matrix A and B. The subspace overlap depends strongly on n_A and n_B. Also, it ignores the eigenvalues. Thus, differences between eigenvectors with small and large eigenvalues contribute equally. But more importantly, (nearly) degenerate subspaces are treated incorrectly. When two or more eigenvalues are equal, the orientation of the corresponding eigenvectors within the subspace is random. This will cause a random difference in the subspace overlap number, whereas for the covariance matrix overlap measure, these identical subspaces do not contribute to the difference.

The covariance matrices for each time were compared with the matrices over the whole intervals (Fig. 7). The overlap is not an exact measure of the convergence, since the

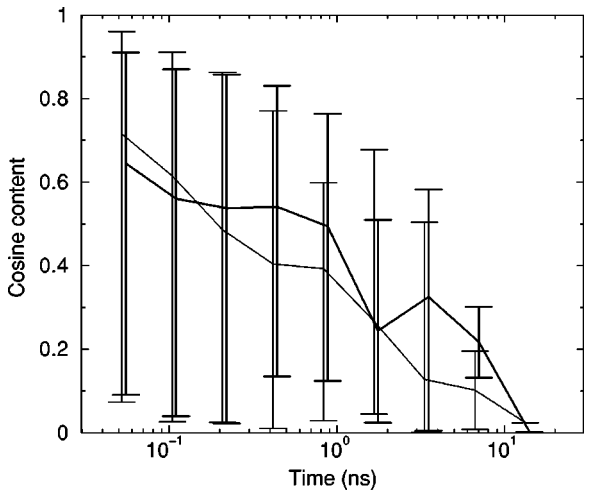

FIG. 8. Cosine content [expression (12)] of the first principal component as a function of the length of the subinterval for HPr (thick line) and lysozyme (thin line). The lines are averages over all subintervals, the error bars indicate the intervals containing 90% of the points.

covariances have not converged after 14 ns. It should, however, give a good indication, especially for the HPr interval, since the total fluctuation seems to be converged.

A possible measure for the sampling of a simulation could be the cosine content of the PC's. The first few principal components of random diffusion, without potential, are cosines with the number of periods equal to half the principal component index, as was proven in Ref. [5]. A measure for similarity to random diffusion is the cosine content

$$c_i = \frac{2}{T}\left(\int_0^T \cos(k\pi t) p_i(t)\mathrm{d}t\right)^2 \left(\int_0^T p_i^2(t)dt\right)^{-1}. \quad (12)$$

The cosine content can take values between zero (no cosine) and 1 (a perfect cosine). It is an absolute measure, which can be extracted from one covariance analysis, in contrast to many other convergence measures, which require comparisons of quantities between different analysis intervals. The cosine content for the time intervals is shown in Fig. 8. The average cosine content decreases from 70% at 50 ps to almost 0 at 14 ns. This would make the cosine content a good indicator for convergence, but unfortunately the deviations from the average are of the size of the average itself.

III. ONE-DIMENSIONAL DIFFUSION

Diffusion in a harmonic potential is described by a stochastic differential equation

$$\frac{dx}{dt} = -ax + r(t). \quad (13)$$

The stochastic term $r(t)$ is δ correlated

$$\langle r(t)\rangle = 0, \quad (14)$$

$$\langle r(t) r(t+\tau)\rangle = 2D\delta(\tau), \quad (15)$$

where $\delta(\tau)$ is the Dirac δ function. The parameter a, which is the inverse correlation time and the diffusion constant D are determined by the force constant k of the harmonic potential, the friction constant ζ, the temperature T, and the Boltzmann's constant k_B,

$$a = \frac{k}{\zeta}, \quad (16)$$

$$D = \frac{k_B \mathrm{T}}{\zeta}. \quad (17)$$

The solution of the differential equation can be written written as the sum of an exponent times the position at time zero and an integral over the stochastic term

$$x(t) = e^{-at}\left(x(0) + \int_0^t e^{av} r(v) dv\right). \quad (18)$$

When simulating a complex system it is generally not known where the minimum of the potential is located. The best estimate for the center is a time average over the simulation (assuming it started from an equilibrated conformation). For the model system, the deviation of the average position from zero can be calculated, when starting from position X_0,

$$\langle \overline{x^2}\rangle_{x(0)=X_0} = \left\langle \left(\frac{1}{T}\int_0^T x(t)dt\right)^2 \right\rangle_{x(0)=X_0} \quad (19)$$

$$= \frac{1}{(aT)^2}\left[(1-e^{-aT})^2 X_0^2 + \frac{D}{a}(-3+2aT + 4e^{-aT} - e^{-2aT})\right], \quad (20)$$

here $\langle\ \rangle$ denotes ensemble averaging and an overline denotes time averaging, the full derivation is given in the Appendix. When also the ensemble average over the starting positions, which are Gaussian distributed with variance D/a, is taken, the expression simplifies to

$$\langle \overline{x^2}\rangle = \frac{D}{a}\frac{-2+2aT+2e^{-aT}}{(aT)^2}. \quad (21)$$

The deviation from the average can be expanded for aT or $1/aT$ small,

$$\langle \overline{x^2}\rangle = \frac{D}{a}\left(1 - \frac{aT}{3}\right) + O((aT)^2), \quad (22)$$

$$\langle \overline{x^2}\rangle = \frac{D}{a}\frac{2}{aT} + O\left(\frac{1}{(aT)^2}\right). \quad (23)$$

The average position converges to zero as $1/\sqrt{T}$, this reflects the fact that for long times the positions are uncorrelated.

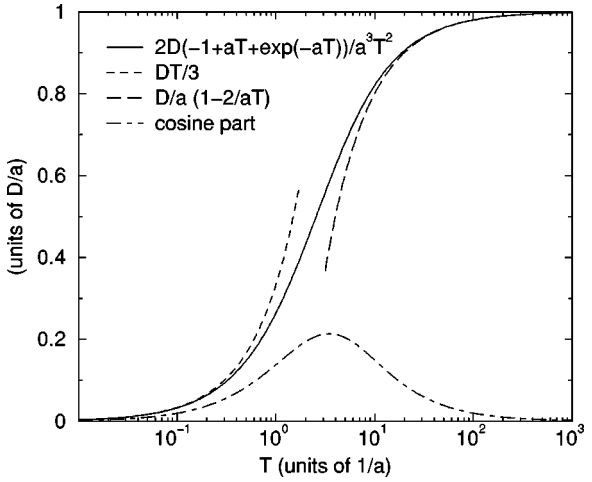

FIG. 9. Ensemble average of the variance of diffusion in a harmonic potential in one dimension [expression (28)], the dashed lines are the approximations for aT small and large. The dot-dashed line is the expectation of the part of the variance that is caused by a cosine of half a period, expression (A8) with $k=1$.

The $1/\sqrt{T}$ convergence is equivalent to the $1/\sqrt{n}$ convergence of the average over n independent draws from a distribution.

When a minimum has been sampled to a reasonable extent, the variance over the simulation is a good indication for the size of the energy well. In the model system the convergence of the variance can be calculated, when starting from position X_0,

$$\langle \overline{(x-\bar{x})^2} \rangle_{x(0)=X_0} \tag{24}$$

$$= \langle \overline{x^2} - \overline{x}^2 \rangle_{x(0)=X_0} \tag{25}$$

$$= \left\langle \frac{1}{T}\int_0^T x(t)^2 dt \right\rangle_{x(0)=X_0} - \langle \overline{x^2} \rangle_{x(0)=X_0} \tag{26}$$

$$= \frac{1}{2aT}\left[(1-e^{-2aT})X_0^2 + \frac{D}{a} \times (-1+2aT+e^{-2aT}) \right] - \langle \overline{x^2} \rangle_{x(0)=X_0}, \tag{27}$$

or when starting from an ensemble average,

$$\langle \overline{(x-\bar{x})^2} \rangle = \frac{D}{a} - \langle \overline{x^2} \rangle. \tag{28}$$

The variance can be expanded for aT small, using Eq. (22)

$$\langle \overline{(x-\bar{x})^2} \rangle = \frac{1}{3} DT + O((aT)^2). \tag{29}$$

This shows that on short time scales the system behaves purely diffusive. The variance and the expansions for short and long times are plotted in Fig. 9. It is not possible to calculate the expectation of the cosine content analytically, because of the stochastic term in the denominator. The expectation of the numerator only can be calculated [see the Appendix, Eq. (A8)], it is also plotted in Fig. 9.

IV. DIFFUSION WITH TWO TIME SCALES

In the protein simulations each principal component has two correlation times, which differ more than an order of magnitude. To model this the position of the minimum in Eq. (13) needs to diffuse in a harmonic potential according to the same equation, but with a longer correlation time. The dynamics of x is now described by two coupled stochastic differential equation

$$\frac{dx}{dt} = -a_x(x-y) + r_x(t), \tag{30}$$

$$\frac{dy}{dt} = -a_y y + r_y(t), \tag{31}$$

where $a_x \gg a_y$. The stochastic terms have expectation zero and the variances are γ correlated:

$$\langle r_x(t) r_x(t+\tau) \rangle = 2D_x \delta(\tau), \tag{32}$$

$$\langle r_y(t) r_y(t+\tau) \rangle = 2D_y \delta(\tau). \tag{33}$$

When variance of x is λ and the diffusion of the minimum contributes a fraction of β to the variance, the diffusion constants are given by

$$D_x = (1-\beta)\lambda a_x, \tag{34}$$

$$D_y = \beta \lambda a_y. \tag{35}$$

Because a_x is much larger than a_y,y in Eq. (30) can be considered as a parameter and the two stochastic differential equations can be treated separately. In this approximation the expectation of the variance of x over an interval of length T is

$$\langle \overline{(x-\bar{x})^2} \rangle = \lambda \left(1 - (1-\beta)\frac{-2+2a_xT+2e^{-a_xT}}{(a_xT)^2} - \beta \frac{-2+2a_yT+2e^{-a_yT}}{(a_yT)^2} \right). \tag{36}$$

V. HIGH-DIMENSIONAL DIFFUSION

The one-dimensional model with two time scales can be extended to N dimensions by simply combining N uncorrelated one-dimensional models,

$$\frac{dx_i}{dt} = -a_{x,i}(x_i - y_i) + r_{x,i}(t), \quad i=1,\ldots,N, \tag{37}$$

$$\frac{dy_i}{dt} = -a_{y,i} y_i + r_{y,i}(t), \quad i=1,\ldots,N. \tag{38}$$

The stochastic terms have expectation zero and the variances are γ correlated:

$$\langle r_{x,i}(t) r_{x,j}(t+\tau)\rangle = 2D_{x,i}\delta_{ij}\delta(\tau), \tag{39}$$

$$\langle r_{y,i}(t) r_{y,j}(t+\tau)\rangle = 2D_{y,i}\delta_{ij}\delta(\tau). \tag{40}$$

This model describes diffusion in an N-dimensional harmonic potential, with two correlation times for each dimension. To mimic protein dynamics, the force constants were chosen equal for x_i and y_i, proportional to i and the diffusion constants independent of the direction:

$$a_{i,x} = 32ai, \tag{41}$$

$$a_{i,y} = ai, \tag{42}$$

$$D_{i,x} = 0.532\,D, \tag{43}$$

$$D_{i,y} = 0.5D. \tag{44}$$

Although the model is relatively simple, analytically deriving collective properties, such as principal modes, is very difficult, if not impossible. To analyze the collective properties, 100 simulations were performed of the model system with $N=30$. Thus the longest correlation time in the model is $1/a$ and the shortest correlation time is $1/(960a)$. The time step of the Euler integrator was $1/(204\,80a)$ and $1/(204\,800a)$ to collect short time-scale data. Each simulation was started from a different, equilibrated, conformation with a different random seed. The simulations were analyzed for time intervals with 16 different lengths, ranging from $1/(512a)$ to $64/a$.

The only quantity for which the expectation can easily be calculated analytically is the sum of variances $V(T)$,

$$V(T) = \sum_{i=1}^{N} \overline{(x_i - \overline{x_i})^2}, \tag{45}$$

$$\langle V(T)\rangle = \sum_{i=1}^{N} \frac{D}{ai}\left(1 - 0.5\,\frac{-2+64aiT+2e^{-32aiT}}{(32iT)^2} - 0.5\,\frac{-2+2aiT+2e^{-aiT}}{(iT)^2}\right). \tag{46}$$

For the model system $V(\infty)$ is $4.0D/a$. In Fig. 10 the average of $\sqrt{V(T)/V(\infty)}$ over the simulations in plotted, which matches the analytical expression exactly. The transitions range from random diffusion to full fluctuation is larger than in the one-dimensional case, since the time scales in the different dimensions cover a range of 30. The length of the transition range can be scaled by choosing a different exponent for the a_i's in Eqs. (41) and (42). Figure 10 also shows a measure for the convergence of the sampled space s, which is defined in the Appendix [expression (11)].

When the time evolution of a high-dimensional system is known, but the potential is unknown or too complex to analyze, the only method to find the directions with the largest fluctuations is principal component analysis. There are two limiting types of behavior in the model system. On very short time scales, $a_{x,N}T \ll 1$, the systems behaves purely diffusive. This regime was analyzed in Ref. [5]. The principal modes have a random orientation and the eigenvalues of the covariance matrix decrease with the square of the index. The PC's are cosines with the number of periods equal to half the eigenvector index. On very long time scales, $a_{y,1}T \gg 1$, the principal modes are oriented along the coordinate axis. PC i converge to x_i, the eigenvalues to $D/(ai)$. On intermediate time scales the principal modes will be partially oriented in the coordinate directions. The first few PC's will still resemble cosines. This was analyzed qualitatively using the ensemble of 100 trajectories and with an analytical approximation for the first PC. No fitting was used in the principal component analyses.

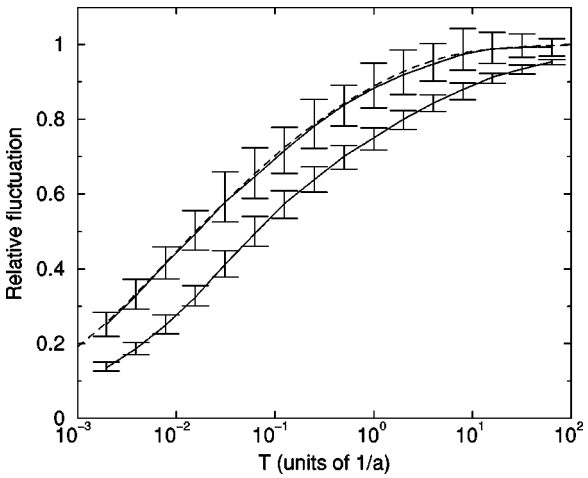

FIG. 10. Top curve is the square root of the sum of variances over all 30 coordinates as a function of time, divided by the value at full sampling. The dashed line is the analytical curve [expression (46)]. The bottom curve is the overlap [expression (11)] of the sampled space with the space at full sampling. All values are averages over 100 simulations, the error bars indicate the intervals containing 90% of the points.

The cosine content of the first principal component (12) is shown in Fig. 11, both for the simulations and for the analytical approximation. As expected, the analytical approximation overestimates the cosine content at intermediate times. Although the average cosine content decreases monotonically in time, it is not a sensitive measure for sampling because of the large fluctuations over the different simulations. When the cosine content is close to 1, one can be sure that the simulation is not converged. When the cosine content is close to zero, one could have full sampling, but it is equally possible that the simulation time is about $2/a$ or less, where the sampling is far from converged and the diffusional motion dominates.

VI. DISCUSSION

In the simulations presented the two proteins jump in a relatively short time from one shallow potential well to an-

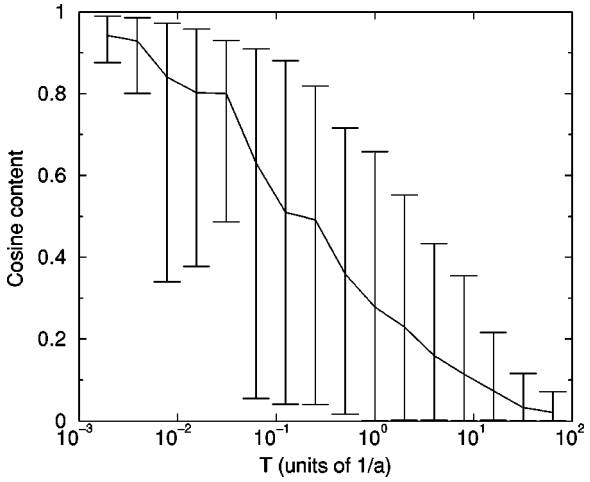

FIG. 11. Cosine content [expression (12)] of the first principal component as a function of time. The curve is the average over 100 simulations, the error bars indicate the intervals containing 90% of the points.

other. The time spent in each well is relatively long. The behavior of the proteins, in the 14 nanoseconds that they spend in one well, and the model system is nearly diffusive on short time scales and becomes more ordered on long time scales. When the longest correlation time in the model, $1/a$ is chosen as 2 ns for HPr and as 4 ns for lysozyme the quantitative agreement with the model is good. This holds for the total fluctuation, the overlap of the sampled spaces and the cosine content. The agreement can be partially explained by the three adjustable parameters, which are the exponent of the eigenvalue curve, the ratio of the two correlation times and the ratio of the amplitudes of the slow and fast fluctuations. However, the exact exponent of the eigenvalue curve is not critical; HPr and lysozyme have exponents of about $-2/3$ and $-4/3$, respectively, while in the model an exponent of -1 was used. The model is relatively simple in the sense that the force constants scale with the same exponent and the slow and fast diffusion constant do not depend on the spatial direction. An important observation is that the model also describes the spread around the averages correctly. This means that not only the average properties of the proteins and the model are similar, but also the ensembles of trajectories. Although the correlation times of the individual degrees of freedom can differ up to a factor of 2 from the algebraic model curve, this does not influence the convergence behavior significantly. For short time scales the model is compatible with the one proposed by Amadei *et al.* [17], which was not intended to model the long time behavior.

An advantage of the model system is that the convergence behavior can be studied accurately. This is impossible for proteins, not only because of the current speed of computers, but more importantly because proteins tend to jump to different conformations on a time scale that is not much longer than the longest correlation time within a conformation. One should realize that because of these jumps it is impossible to get a complete picture of the available phase space, with the current speed of the computers. Even when the protein stays in one conformation during a simulation, a jump could occur when the simulation is prolonged.

Thus, the main part of the fluctuations of the two proteins during the chosen intervals can be described as diffusion in a high-dimensional harmonic potential, of which the position of the minimum diffuses in a potential of the same shape, but on a much longer time scale. The behavior can be interpreted as thermal motion of slow, collective coordinates in a potential of mean force of the faster degrees of freedom of the protein and the solvent. The minimum of this potential fluctuates slowly around an average, probably due to slow rearrangement of the packing of the side chains.

Using the model system it can be estimated that the 14 ns intervals of the protein simulations are approximately eight and four times longer than the longest correlation time for HPr and lysozyme. The longest correlations times for the proteins obtained from the fits of the autocorrelations of the PC's are shorter than a nanosecond. However, fitting the autocorrelation of the PC's of the model system shows that $1/a$ is underestimated on average by a factor of 2 and 3 for simulation times of $8/a$ and $4/a$, respectively. The simulation time needs to be increased by an order of magnitude to obtain a reasonable estimate of the longest correlation time. Nevertheless, the chosen intervals seem to be long enough to estimate the mean square fluctuation.

The hope was that with the model system some indicator could be found that provides a good prediction of the convergence of the sampling around one conformation. The total fluctuation, a simple property, is not suitable, since it increases logarithmically on intermediate time scales. The cosine content of the first principal component seems more promising. The sampling as defined by the overlap [expression (11)] is approximately equal to one minus the cosine content. Unfortunately, the fluctuations in the cosine content are of the size of the average. This renders it useless as an indicator, since an accurate value can only be obtained by averaging over many pieces of a long trajectory. The cosine content is a useful negative indicator. When the first principal component is similar to a cosine with half a period, the sampling is far from converged. From the results for the protein simulations and the model system we can conclude that all quantities are too uncertain to predict the long term sampling from a short simulation. The only way to assess the convergence of sampling of a short simulation seems to be by performing a longer one.

ACKNOWLEDGMENTS

The author thanks A. E. Mark for his support and H. J. C. Berendsen for stimulating discussions.

APPENDIX

The expectation of the square of an integral of the stochastic process $x(t)$, as defined by Eqs. (13), (14), and (15), with an arbitrary function $f(t)$ can be calculated as follows:

$$\left\langle \left(\int_0^T f(t)x(t)dt \right)^2 \right\rangle_{x(0)=X_0} \tag{A1}$$

$$= \left\langle \left\{ \int_0^T f(t)e^{-at}\left[x(0) + \int_0^t e^{av}r(v)dv \right] dt \right\}^2 \right\rangle_{x(0)=X_0} \tag{A2}$$

$$= \left\langle \left(\int_0^T f(t)e^{-at}x(0)dt \right)^2 + 2\int_0^T f(t)e^{-at}x(0)dt \int_0^T f(t)e^{-at}\int_0^t e^{av}r(v)dv\, dt + \left(\int_0^T f(t)e^{-at}\int_0^t e^{av}r(v)dv\, dt \right)^2 \right\rangle_{x(0)=X_0} \tag{A3}$$

$$= \left(\int_0^T f(t)e^{-at}dt \right)^2 X_0^2 + \int_0^T \int_0^T f(t)f(u)e^{-a(t+u)} \int_0^t \int_0^u \langle e^{a(v+w)}r(v)r(w)\rangle dw\, dv\, du\, dt \tag{A4}$$

$$= \left(\int_0^T f(t)e^{-at}dt \right)^2 X_0^2 + \int_0^T \int_0^T f(t)f(u)e^{-a(t+u)} \frac{2D}{a}\{\exp[2a\min(t,u)]-1\}du\, dt \tag{A5}$$

$$= \left(\int_0^T f(t)e^{-at}dt \right)^2 X_0^2 + \frac{2D}{a}\int_0^T \int_0^T f(t)f(u)(e^{-a|t-u|} - e^{-a(t+u)})du\, dt. \tag{A6}$$

The expectation of the square of the average of x can be obtained from expression (A6) by taking f equal to 1,

$$\left\langle \left(\int_0^T x(t)dt \right)^2 \right\rangle_{x(0)=X_0} = \frac{1}{a^2}(1-e^{-aT})^2 X_0^2 + \frac{D}{a^3}(-3+2aT+4e^{-aT}-e^{-2aT}). \tag{A7}$$

The expectation of the overlap of x with a cosine, ensemble averaged over the starting value (a Gaussian distribution with variance D/a) is

$$\left\langle \left(\frac{1}{T}\int_0^T \sqrt{2}\cos(k\pi t)x(t)dt \right)^2 \right\rangle = \frac{2DT\{k^2\pi^2 + a^2T^2 + 2aT[-1+e^{-aT}(-1)^k]\}}{(k^2\pi^2 + a^2T^2)^2}. \tag{A8}$$

The expectation of the integral over the square of x is

$$\left\langle \int_0^T x(t)^2 dt \right\rangle_{x(0)=X_0} \tag{A9}$$

$$= \left\langle \int_0^T \left\{ e^{-at}\left[x(0) + \int_0^t e^{av}r(v)dv \right] \right\}^2 dt \right\rangle_{x(0)=X_0} \tag{A10}$$

$$= \left\langle \int_0^T [e^{-at}x(0)]^2 + 2e^{-2at}x(0)\int_0^t e^{av}r(v)dv + e^{-2at}\left(\int_0^t e^{av}r(v)dv \right)^2 dt \right\rangle_{x(0)=X_0} \tag{A11}$$

$$= \int_0^T e^{-2at}dt X_0^2 + \int_0^T e^{-2at} \times \int_0^t \int_0^t \langle e^{a(v+w)}r(v)r(w)\rangle dw\, dv\, dt \tag{A12}$$

$$= \int_0^T e^{-2at}dt X_0^2 + \int_0^T e^{-2at}\int_0^t De^{2av}dv\, dt \tag{A13}$$

$$= \frac{1}{2a}(1-e^{-2aT})X_0^2 + \frac{D}{2a^2}(-1+2aT+e^{-2aT}). \tag{A14}$$

[1] R.M. Levy, A.R. Srinivasan, W.K. Olson, and J.A. McCammon, Biopolymers **23**, 1099 (1984).
[2] A. E. García, Phys. Rev. Lett. **68**, 2696 (1992).
[3] A.E. García and J.G. Harman, Protein Sci. **5**, 62 (1996).
[4] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen, Proteins: Struct., Funct., Genet. **17**, 412 (1993).
[5] B. Hess, Phys. Rev. E **62**, 8438 (2000).
[6] S. Hayward and N. Gō, Annu. Rev. Phys. Chem. **46**, 223 (1995).
[7] A. Kitao and N. Gō, J. Comput. Chem. **12**, 359 (1991).
[8] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. [Electronic Publication] **7**, 306 (2001).
[9] W. F. van Gunsteren *et al.*, *Biomolecular Simulation: GROMOS96 MANUAL AND USER GUIDE* (BIOMOS b.v., Zürich, Groningen, 1996).
[10] A.K. Feenstra, B. Hess, and H.J.C. Berendsen, J. Comput. Chem. **20**, 786 (1999).
[11] B. Hess, H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[12] H.J.C. Berendsen *et al.*, J. Chem. Phys. **81**, 3684 (1984).
[13] Z. Jia, J.W. Quail, E.B. Waygood, and L.T.J. Delbaere, J. Biol. Chem. **268**, 22 490 (1993).
[14] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans, in *Intermolecular Forces*, edited by B. Pullman (Reidel, Dordrecht, The Netherlands, 1981), pp. 331–342.
[15] L.H. Weaver and B.W. Matthews, J. Mol. Biol. **193**, 189 (1987).
[16] S. Hayward, A. Kitao, and H.J.C. Berendsen, Proteins: Struct., Funct., Genet. **27**, 425 (1997).
[17] A. Amadei *et al.*, Proteins: Struct., Funct., Genet. **35**, 283 (1999).