Similarities between principal components of protein dynamics and random diffusion

Berk Hess
Department of Biophysical Chemistry, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

(Received 11 February 2000)

Principal component analysis, also called essential dynamics, is a powerful tool for finding global, correlated motions in atomic simulations of macromolecules. It has become an established technique for analyzing molecular dynamics simulations of proteins. The first few principal components of simulations of large proteins often resemble cosines. We derive the principal components for high-dimensional random diffusion, which are almost perfect cosines. This resemblance between protein simulations and noise implies that for many proteins the time scales of current simulations are too short to obtain convergence of collective motions.

PACS number(s): 87.14.Ee, 87.15.Aa, 87.15.He

INTRODUCTION

This article focuses on the use of covariance analysis as a tool to gain information about the conformational freedom of proteins. Knowledge about the conformational freedom of a protein can give insight into the stability and functional motions of the protein. Conformational freedom is related to the potential energy: in equilibrium the distribution of conformations can, in principle, be calculated from the potential. In practice this is not possible, since proteins are too complex, not only because of their high dimensionality, but also because the energy landscape usually consists of many minima. Normal mode analysis can be used to analyze any of these potential energy minima, but it does not take entropy due to the occupation of several minima into account, which plays an important role at room temperature. For reviews on normal mode and principal component analysis applied to proteins, see [1,2]. The first application of principal component analysis to macromolecules was performed to estimate the configurational entropy [3]. More recently the application of principal component analysis to proteins has been termed ‘‘molecule optimal dynamic coordinates’’ [4,5] and ‘‘essential dynamics’’ [6]. It has now become a standard technique for analyzing molecular dynamics trajectories. However, the effects of insufficient sampling on the results are not well understood.

Covariance analysis, also called principal component analysis, is a mathematical technique for analyzing high-dimensional data sets. Essentially, it defines a new coordinate system for the data set, with the special property that the covariance is zero for any two coordinates. In this sense these new coordinates can be called uncorrelated. These coordinates are to be ordered according to the variance of the data in that coordinate. This can allow for a reduction of the dimensionality of the space by neglecting the coordinates with small variance, thus concentrating on the coordinates with larger spread or fluctuations.

The procedure for N-dimensional data $\mathbf{x}(t)$ is as follows. First the covariance matrix has to be constructed. The covariance C_{ij} of coordinate i and coordinate j is defined as

$$C_{ij}=\langle (x_i-\langle x_i\rangle)(x_j-\langle x_j\rangle)\rangle,$$

where $\langle\ \rangle$ is the average over all data points. The data can consist of a finite number of N-dimensional points, in which case the average is a summation. $\mathbf{x}(t)$ can also be an N-dimensional function; then the average is an integral. The symmetric $N\times N$ matrix C can be diagonalized with an orthonormal transformation matrix R:

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

The columns of R are the eigenvectors or principal modes. The eigenvalues λ are equal to the variance in the direction of the corresponding eigenvector; they can be chosen such that $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_N\geqslant 0$. The original data can be projected on the eigenvectors to give the principal components p_i, $i=1,\ldots,N$:

$$\mathbf{p}=R^T(\mathbf{x}-\langle\mathbf{x}\rangle).$$

When the data are the result of a dynamic process, the principal components are a function of time and $p_1(t)$ will be the principal component with the largest mean square fluctuation. Note that if the system obeys Newton’s laws every coordinate has to be weighted with the square root of the mass to obtain physically relevant dynamic principal components.

For a system moving in a (quasi)harmonic potential, the principal modes are similar to the normal modes, which are the eigenvectors of the Hessian in the energy minimum. The advantage of normal mode analysis is that it depends only on the shape of the potential energy surface; however, its use is restricted to quasiharmonic systems. Covariance analysis can be applied to any dynamic system, but the results will also depend on the sampling. The problem is how to separate intrinsic properties of the system from sampling artifacts. As was shown by Linssen [7], the principal components of multidimensional random diffusion are cosine shaped, with amplitudes inversely proportional to the eigenvalue ranking number. This behavior is very similar to that observed from simulations of proteins. In order to interpret protein data correctly, it is important that the behavior of random diffusion is basically understood. In the next section we will perform a theoretical analysis of random diffusion. The results of random diffusion depend only on the sampling, because there is no potential.

ANALYSIS OF RANDOM DIFFUSION

N-dimensional diffusion is described by a system of N independent stochastic differential equations:

$$\frac{dx_i(t)}{dt}=r_i(t), \quad i=1,\ldots,N. \tag{1}$$

$r_i(t)$ is a stationary stochastic noise term, which is δ-correlated:

$$E[r_i(t)]=0, \tag{2}$$

$$E[r_i(t)r_j(t+\tau)]=2D\delta_{ij}\delta(\tau) \tag{3}$$

where $E[\,]$ denotes the expectation, δ_{ij} is the Kronecker delta and $\delta(\tau)$ is the Dirac delta function. The solution of this system is a set of N nonstationary stochastic functions $x_i(t)$, which represent Wiener processes [8]. The displacements $x_i(t+\tau)-x_i(t)$ are Gaussian distributed with

$$E[x_i(t+\tau)-x_i(t)]=0, \tag{4}$$

$$E[\{x_i(t+\tau)-x_i(t)][x_j(t+\tau)-x_j(t)\}]=2D\tau\delta_{ij}. \tag{5}$$

We will use this continuum description for the analysis of random diffusion. Realizations can be generated only for a finite number of time points. This is not a problem, since the results of the covariance analysis will be virtually identical when we have more than N time points.

We are interested in the eigenvalues and eigenvectors of the covariance matrix C and in the shapes of the principal components:

$$C_{ij}=\frac{1}{T}\int_0^T[x_i(t)-\langle x_i\rangle][x_j(t)-\langle x_j\rangle]dt, \tag{6}$$

$$p_k(t)=\mathbf{e}_k\cdot[\mathbf{x}(t)-\langle\mathbf{x}\rangle], \tag{7}$$

where $\mathbf{e}_k$ is eigenvector k and $\langle x_i\rangle$ is the average of x_i:

$$\langle x_i\rangle=\frac{1}{T}\int_0^T x_i(t)dt. \tag{8}$$

The eigenvectors are ordered according to descending eigenvalue. The first eigenvector $\mathbf{e}_1$ is the vector that produces the linear combination of x_i's with the largest mean square fluctuation. Thus $\mathbf{e}_1$ is the vector with $\|\mathbf{e}_1\|=1$ which maximizes

$$\lambda_1=\max_{\|\mathbf{e}_1\|=1}\frac{1}{T}\int_0^T[\mathbf{e}_1\cdot(\mathbf{x}(t)-\langle\mathbf{x}\rangle)]^2dt. \tag{9}$$

We can write $\mathbf{x}(t)-\langle\mathbf{x}\rangle$ as a sum of a series of orthonormal functions f_k:

$$\lambda_1=\max_{\|\mathbf{e}_1\|=1}\frac{1}{T}\int_0^T\left(\mathbf{e}_1\cdot\sum_{k=0}^{\infty}\mathbf{c}^k f_k(t)\right)^2dt, \tag{10}$$

$$\int_0^T f_k(t)f_l(t)dt=\delta_{kl}, \tag{11}$$

$$c_i^k=\int_0^T f_k(t)[x_i(t)-\langle x_i\rangle]dt. \tag{12}$$

We can choose the f_k such that $f_k(t)$ is proportional to the projection on eigenvector k: $f_k(t)\sim\mathbf{e}_k\cdot[\mathbf{x}(t)-\langle\mathbf{x}\rangle]$, thus f_k will contribute only to λ_k. $f_1(t)$ is the function that maximizes the variance:

$$\lambda_1=\max_{\|\mathbf{e}_1\|=1}\max_{f_1}\frac{1}{T}\int_0^T[\mathbf{e}_1\cdot\mathbf{c}^1f_1(t)]^2dt$$

$$=\max_{\|\mathbf{e}_1\|=1}\max_{f_1}\frac{1}{T}(\mathbf{e}_1\cdot\mathbf{c}^1)^2. \tag{13}$$

It can easily be derived that $\mathbf{e}_1$ is proportional to $\mathbf{c}^1$:

$$\mathbf{e}_1=\frac{\mathbf{c}^1}{\sqrt{\mathbf{c}^1\cdot\mathbf{c}^1}}. \tag{14}$$

When N is 1, $f_1(t)$ is proportional to $x_1(t)-\langle x_1\rangle$. As N gets larger, the set of x_i's will form a better representation of the full ensemble. For large N we can approximate λ_1 with an ensemble average:

$$\lambda_1=\max_{f_1}\frac{1}{T}\mathbf{c}^1\cdot\mathbf{c}^1\approx\max_{f_1}\frac{1}{T}\,\mathrm{E}[\mathbf{c}^1\cdot\mathbf{c}^1]$$

$$=\max_{f_1}\frac{1}{T}\,\mathrm{E}[Nc_i^1c_i^1]=\lambda_1^*. \tag{15}$$

The integral over f_k is zero:

$$\int_0^T f_k(t)dt\sim\int_0^T\mathbf{e}_k\cdot[\mathbf{x}(t)-\langle\mathbf{x}\rangle]dt=0. \tag{16}$$

For large N we can approximate f_1 by the function that maximizes λ_1^*; we will call this function f_1^*. Since the integral over f_1 is zero and we want to approximate f_1, we demand that the integral over f_1^* is zero as well. Using this we can calculate the expectation of $c_i^1c_i^1$:

$$E[(c_i^1)^2]=E\left[\left(\int_0^T f_1^*(t)x_i(t)dt\right)^2\right]$$

$$=2D\int_0^T\left(\int_0^t f_1^*(u)du\right)^2dt. \tag{17}$$

The full derivation is given in the Appendix. When we define $g(t)$ as

$$g(t)=\int_0^t f_1^*(u)du, \tag{18}$$

we can rephrase the optimization problem in terms of g and add the constraints (11) and (16) using Lagrange multipliers μ_1 and μ_2. We have to find the function g that maximizes

$$\int_0^T [g(t)]^2 + \mu_1 g'(t) + \mu_2([g'(t)]^2 - 1)dt$$

$$= \int_0^T L(t,g,g')dt \tag{19}$$

with the boundary conditions

$$g(0) = \int_0^0 f_1^*(t)dt = 0, \tag{20}$$

$$g(T) = \int_0^T f_1^*(t)dt = 0. \tag{21}$$

According to the Euler-Lagrange formalism, the optimal g is the solution of

$$\frac{\partial L}{\partial g} - \frac{d}{dt}\frac{\partial L}{\partial g'} = 2g(t) - 2\mu_2 g''(t) = 0. \tag{22}$$

The solutions are

$$g_k(t) = C \sin\left(\frac{\pi k t}{T}\right) \quad \text{with} \quad k = 0,1,2,3, \ldots . \tag{23}$$

g_0 only satisfies the boundary conditions when $C = 0$, so we discard this function. By differentiating g_k and using Eq. (11) we obtain a set of orthonormal functions h_k:

$$h_k(t) = \frac{dg_k(t)}{dt} = \sqrt{\frac{2}{T}}\cos\left(\frac{\pi k t}{T}\right) \quad \text{with} \quad k = 1,2,3, \ldots . \tag{24}$$

We have to find the h_k that maximizes λ_1^*:

$$f_1^* = h_k \Rightarrow \lambda_1^* = \frac{1}{T} E[N c_i^1 c_i^1] = \frac{2NDT}{\pi^2 k^2}. \tag{25}$$

Thus the best guess for the projection on the first eigenvector f_1^* is h_1. We can apply the same procedure for f_k^*, with the extra restriction that Eq. (11) should hold for all $1 \leqslant l < k$. The result is

$$f_k^* = h_k, \tag{26}$$

$$\lambda_k^* = \frac{1}{T} E[N c_i^1 c_i^1] = \frac{2NDT}{\pi^2 k^2}. \tag{27}$$

The derivation of λ_k^* is based on expectations and statistical fluctuations are not taken into account. Since the difference between λ_1^* and λ_2^* is a factor of 4, we expect that the projection on the first eigenvector is very close to a half cosine. But for large k the statistical fluctuations might cause mixing of the cosines.

The projections f_k^* suggest that the covariance matrix can be approximately diagonalized with a matrix of cosine coefficients. To test this approximation we have to write each x_i as a sum of cosines:

$$x_i(t) - \langle x_i \rangle = \sum_{k=1}^{\infty} c_i^k f_k^*(t) = \sqrt{\frac{2}{T}} \sum_{k=1}^{\infty} c_i^k \cos\left(\frac{\pi k t}{T}\right), \tag{28}$$

where c_i^k is defined as

$$c_i^k = \int_0^T x_i(t) f_k^*(t) dt = \sqrt{\frac{2}{T}} \int_0^T x_i(t) \cos\left(\frac{\pi k t}{T}\right) dt. \tag{29}$$

The cosine coefficients are zero on average and uncorrelated:

$$E[c_i^k] = 0, \quad \text{where} \quad k = 1,2,3, \ldots, \tag{30}$$

$$E[c_i^k c_j^l] = \frac{2DT^2}{\pi^2 kl}\delta_{ij}\delta_{kl}, \quad \text{where} \quad k,l = 1,2,3, \ldots . \tag{31}$$

Full derivations are given in the Appendix. The covariance matrix can be expressed in terms of cosine coefficients:

$$C_{ij} = \frac{1}{T}\int_0^T \sum_{k=1}^{\infty} c_i^k f_k^*(t) \sum_{l=1}^{\infty} c_j^l f_l^*(t) dt$$

$$= \frac{1}{T}\sum_{k=1}^{\infty} c_i^k c_j^k = \frac{1}{T}\sum_{k=1}^{N} c_i^k c_j^k + O\left(\frac{1}{N}\right)$$

$$= C_{ij}^* + O\left(\frac{1}{N}\right). \tag{32}$$

We can write C^* as a matrix product which involves a diagonal matrix and a matrix with independent stochastic elements with variance 1:

$$C_{ij}^* = \sum_{k=1}^{N}\left(\frac{\sqrt{2DT}}{\pi k}\right)\left(\frac{\pi k}{\sqrt{2DT}} c_i^k\right)\left(\frac{\pi k}{\sqrt{2DT}} c_j^k\right)\left(\frac{\sqrt{2DT}}{\pi k}\right)$$

$$= (Y^{1/2} X X^T Y^{1/2})_{ij}, \tag{33}$$

where Y is a diagonal matrix with $Y_{kk} = 2DT\pi^{-2}k^{-2}$. Recently Bai and Silverstein proved that for large N the eigenvalues of $(1/N)C^*$ and Y have the same separation [9]. Since Y is a diagonal matrix with separated eigenvalues $Y_{kk} = \lambda_k^*/N$, the eigenvalues of C^* and Y are identical for large N. Because C^* is a good approximation of C, the eigenvalues of C are λ_k^*.

We want to prove that the projections on the first few eigenvectors of C resemble cosines. When this is the case, we can approximate the eigenvector matrix of C with a matrix of the first N normalized cosine coefficients for each coordinate:

$$R_{ik}^* = \frac{1}{\sqrt{\mathbf{c}_k \cdot \mathbf{c}_k}} c_i^k \approx \frac{1}{\sqrt{E[\mathbf{c}_k \cdot \mathbf{c}_k]}} c_i^k = \frac{\pi k}{\sqrt{2NDT}} c_i^k$$

$$= R'_{ik}, \quad \text{where} \quad i,k = 1, \ldots, N. \tag{34}$$

For reasonably large N, R' will be a good approximation of R^*. The use of R' instead of R^* simplifies the analysis. On average the rows and columns of R' are uncorrelated and have norm 1:

$$E\left[\sum_{k=1}^{N} R'_{ik}R'_{jk}\right]=\delta_{ij}\,, \tag{35}$$

$$E\left[\sum_{i=1}^{N} R'_{ik}R'_{il}\right]=\delta_{kl}\,. \tag{36}$$

However, the columns are only approximately orthogonal:

$$E\left[\left(\sum_{i=1}^{N} R'_{ik}R'_{il}\right)^2\right]=E\left[\left(\frac{\pi^2 kl}{2NDT^2}\right)^2\sum_{i=1}^{N}\sum_{j=1}^{N} c_i^k c_j^k c_i^l c_j^l\right] \tag{37}$$

$$=\left(\frac{\pi^2 kl}{2NDT^2}\right)^2\sum_{i=1}^{N} E[(c_i^k)^2]E[(c_i^l)^2] \tag{38}$$

$$=\frac{1}{N}\quad \text{for}\quad i\neq j. \tag{39}$$

We can use the matrix R' to transform the covariance matrix:

$$B_{ij}=(R'^{T}CR')_{ij}=\sum_{l=1}^{N} R'_{li}\sum_{m=1}^{N} C_{lm}R'_{mj} \tag{40}$$

$$=\sum_{l=1}^{N}\frac{\pi i}{\sqrt{2NDT}}c_l^i\sum_{m=1}^{N}\frac{1}{T}\sum_{k=1}^{\infty} c_l^k c_m^k\frac{\pi j}{\sqrt{2NDT}}c_m^j \tag{41}$$

$$=\frac{\pi^2 ij}{2NDT^3}\sum_{l=1}^{N} c_l^i\sum_{m=1}^{N} c_m^j\sum_{k=1}^{\infty} c_l^k c_m^k\,. \tag{42}$$

Since the columns of R' are not completely orthogonal, matrix B will not inherit all the properties of C. We hope that the matrix B is approximately diagonal. To check this we have to calculate the expectation of each matrix element:

$$E[B_{ii}]=\frac{\pi^2 i^2}{2NDT^3}E\left[\left(\sum_{l=1}^{N} c_l^i\sum_{m=1}^{N} c_m^i\sum_{k=1}^{\infty} c_l^k c_m^k\right)\right] \tag{43}$$

$$=\frac{\pi^2 i^2}{2NDT^3}\left(E\left[\sum_{l=1}^{N} c_l^{i2}\sum_{m=1}^{N} c_m^{i\,2}(1-\delta_{lm})\right]+E\left[\sum_{l=1}^{N} c_l^{i2}\sum_{k=1}^{\infty} c_l^{k2}(1-\delta_{ik})\right]+E\left[\sum_{l=1}^{N} c_l^{i4}\right]\right) \tag{44}$$

$$=\frac{2DTi^2}{N\pi^2}\left[\frac{N^2-N}{i^4}+\frac{N}{i^2}\left(\frac{\pi^2}{6}-\frac{1}{i^2}\right)+\frac{3N}{i^4}\right] \tag{45}$$

$$=2DT\left(\frac{1}{6}+\frac{N+1}{\pi^2 i^2}\right), \tag{46}$$

$$E[B_{ij}]=0\quad \text{for } i\neq j, \tag{47}$$

where we have used the fourth moment of a Gaussian distribution for $E[c_l^{i4}]$. The trace of B is twice as large as the trace of C. Comparing B_{ii} with λ_i^* shows that this difference is caused by the term 1/6, which arises from correlations between the columns of R'. The 1/6 is negligible for $i\ll\sqrt{N}$. The expectation of the off-diagonal elements is zero, because every term contains at least one c with an odd power. For the same reason the cross correlation between elements of B is zero:

$$E[B_{ij}B_{kl}]=0\quad \text{for } i\neq j \text{ or } k\neq l. \tag{48}$$

The off-diagonal elements are zero on average, but they might be very large. The variance of the off-diagonal elements of B is

$$E[(B_{ij})^2]=4D^2T^2\left[\frac{1}{90}+\frac{1}{3\pi^2}\left(\frac{1}{i^2}+\frac{1}{j^2}\right)+\frac{N+3}{\pi^4}\left(\frac{1}{i^4}+\frac{1}{j^4}\right)+\frac{2N+4}{i^2 j^2}+O\left(\frac{1}{N}\right)\right]. \tag{49}$$

The derivation is given in the Appendix. When $i\ll\sqrt{N}$, the off-diagonal elements are order $\sqrt{N}$, which is $\sqrt{N}$ smaller than the diagonal elements. Although there are $N-1$ off-diagonal elements, they are completely uncorrelated. The sum of N random elements is $\sqrt{N}$ larger than each element and also random. A cumulative effect can be achieved for one inner product of an eigenvector with a row of B, but for an eigenvalue of order N the inner products with all N rows should be N larger than the eigenvector elements and of the correct sign. Since this is impossible and for $i\ll\sqrt{N}$ the eigenvalues of C are $B_{ii}+O(1)$, the projections will resemble cosines. When $i\geqslant\sqrt{N}$, the off-diagonal elements are of the same order as the diagonal elements and the resemblance will be lost.

VALIDATION

To see how much real principal components resemble cosines, we have to compare the theory with results from Langevin dynamics simulations. An $x_i(t)$ can be generated only for a finite number of time points. This can be done using the algorithm $x_i((n+1)\tau)=x_i(n\tau)+\Delta x$. The value of $x_i(0)$ is irrelevant. Δx should be drawn from a Gaussian distribution which obeys Eqs. (4) and (5). When τ is small, the distribution does not have to be Gaussian, since the convolution of many distributions will converge to a Gaussian distribution. The simulations and covariance analysis were performed with the GROMACS package [10]. For the simulations we always used 1000 steps and a convolution of four uniform distributions, which corresponds to 4000 steps with a uniform distribution. Figure 1 shows the eigenvalues obtained from a simulation and the B_{ii} from Eq. (40) with normalized eigenvector length for $N=120$, $D=0.5$, and $T=1$. The first three eigenvalues are predicted quite accurately. After ten eigenvalues the diagonal elements of B are too small compared to the off-diagonal elements to be a reasonable approximation of the eigenvalues. The predictions are close to the estimated values [Eq. (46)]; the term 1/6 in Eq. (46) starts to dominate for $i>10$. When this term is discarded and the eigenvalues are estimated by λ_i^* [Eq. (27)], a good correspondence is obtained over the whole range. Using λ_i^* it can be calculated that on average the first three eigenvalues contain 80% of the mean square fluctuation when $N=10$ and 84% for large N. The first five princi-

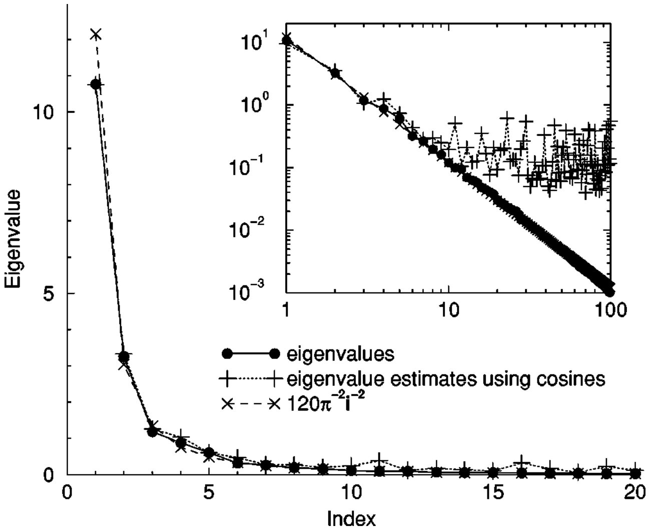

FIG. 1. The eigenvalues of the covariance matrix of a 120-dimensional Brownian dynamics simulation. The eigenvalue estimates are the B_{ii} from Eq. (40), using the normalized vectors of Eq. (34).

pal components are shown in Fig. 2; they resemble cosines.

The cosine content of an eigenvector can be determined by taking inner products with columns of R^*. Because these inner products fluctuate heavily, we have calculated average inner products over 100 Langevin dynamics simulations with N set to 30, 120, and 480. Figure 3 shows average inner products of eigenvector i with column i of R^*. The first eigenvector is almost a perfect cosine, with an inner product larger than 0.996 for all three system sizes. The inner products are one-half for $i=\sqrt{N}$, which is exactly the behavior we expected.

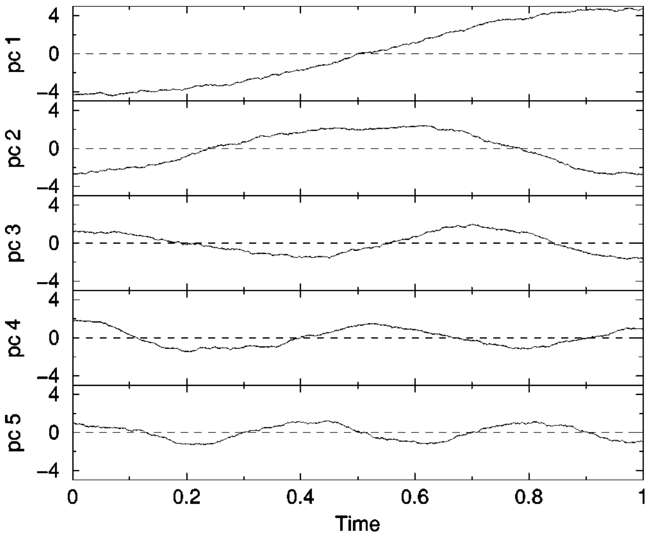

FIG. 2. The first five principal components of a 120-dimensional Brownian dynamics simulation.

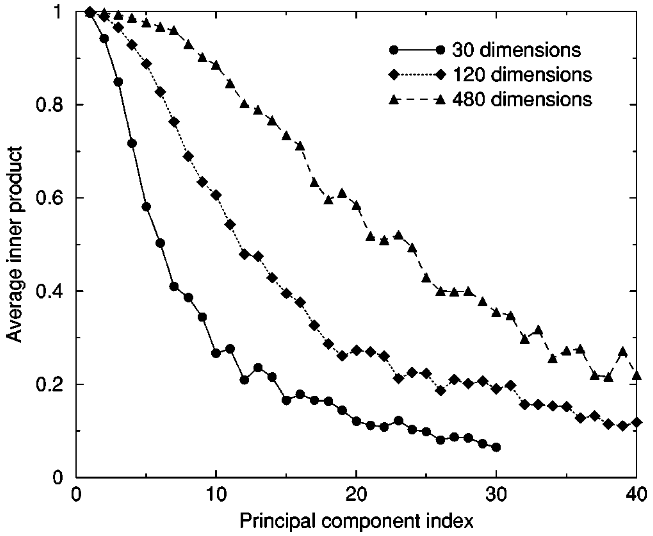

FIG. 3. The average inner products of the eigenvectors of the covariance matrix and the estimated eigenvectors R^* [Eq. (34)]. The inner products were averaged over 100 simulations for each system size. Since the columns of R' are cosine coefficients, these inner products show how much each principal component resembles a cosine with the number of periods equal to one-half the principal component index. The inner product is one-half at the square root of the system size. The error bars show the standard deviation.

The cosinelike principal components lead to strange dynamic effects. One example is the mean square displacement, which is proportional to time for a single pure diffusive coordinate. The mean square displacement along the cosine-shaped principal components is proportional to the time squared. This can be shown with a simple derivation. The principal component i can be approximated as

$$p_i(t)\approx\frac{\sqrt{4NDT}}{i\pi}\cos\left(\frac{\pi i t}{T}\right). \tag{50}$$

From this we can calculate the mean square displacement $\mathcal{M}$ along eigenvector i:

$$\mathcal{M}_i(t)\approx\frac{1}{T-t}\int_0^{T-t}\frac{4NDT}{i^2\pi^2}\left[\cos\left(\frac{i\pi y}{T}\right)-\cos\left(\frac{i\pi(y+t)}{T}\right)\right]^2 dy \tag{51}$$

$$=\frac{8NDT}{i^2\pi^2}\sin^2\left(\frac{i\pi t}{2T}\right)\left[1+\frac{T}{i\pi(T-t)}\sin\left(\frac{i\pi t}{T}\right)\right] \tag{52}$$

$$\approx\frac{2ND}{T}t^2 \quad \text{for} \quad t<\frac{T}{2i}. \tag{53}$$

The mean square displacements of the principal components for the 120-dimensional Brownian system (Fig. 2) are shown in Fig. 4. The whole system behaves diffusively, but the first few eigenvectors exhibit ballistic motion.

PROTEIN SIMULATIONS

Molecular simulations of macromolecules are good examples of high-dimensional systems where principal component analysis can be useful. It will reveal global motions when they are present in the system, without having to visually inspect the whole trajectory. In order to analyze internal

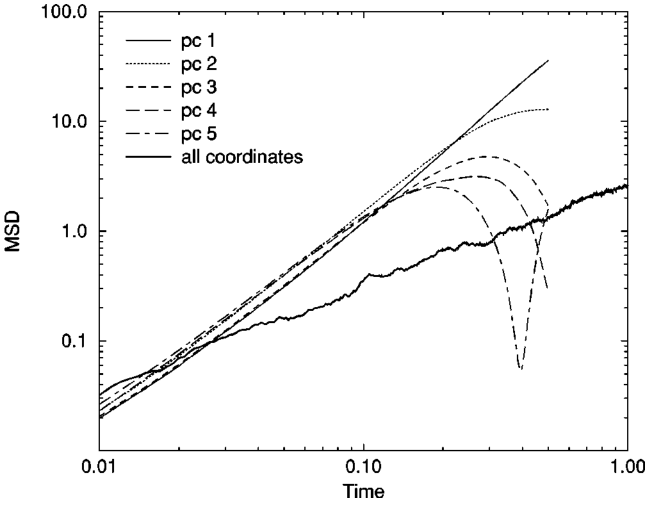

FIG. 4. The mean square displacement (MSD) of the first five principal components of a 120-dimensional Brownian simulation. The principal components are ballistic (MSD $\sim t^2$) over a large range of times. The MSD of the whole system is proportional to time.

motions of a molecule, the overall rotation has to be removed. The most simple way to accomplish this is least squares fitting each structure on a reference structure. Care should be taken when applying this procedure. Because different reference structures will lead to different orientations of the structures with respect to each other, the reference structure should be representative for the whole ensemble of structures. When large movements take place the fit might not be well determined and the fitted structures might jump between different orientations. This problem is most apparent in elongated linear chains, such as polymers, where the rotational orientation around the chain axis is not well defined.

We will present principal component analysis results for three different protein systems. In all these systems the dynamics of the first few principal components is mainly diffusive, but the sampled part of the free-energy landscape has different characteristics for the different simulations. The three simulations and the covariance analyses were performed with the GROMACS package [10].

OMPf porin in a bilayer

Outer membrane protein f (OMPf) is a trimeric protein that consists of three β barrels of 340 residues each. The protein was simulated with molecular dynamics (MD) at pH 3 in a bilayer of 153 di-miristoyl-phosphatidyl-choline (DMPC) lipids, surrounded by 11 419 simple-point-charge (SPC) water molecules and 66 chloride ions. This system of 51 813 atoms was simulated for 3 ns with a time step of 2 fs; the coordinates were written to file every 10 ps. A twin-range cutoff was used: 1.0 nm for the Lennard-Jones and short range Coulomb interactions and 1.8 nm for the long range Coulomb interactions, updated every 10 steps. The temperature and pressure were coupled to 310 K and 1 bar, respectively. The system is similar to the system in [11]. A description of the force-field parameters can be found in [12].

We split the trajectory into three parts of 1 ns each. Covariance analyses were performed for each part on the C_α atoms of monomer 1 (340 atoms, 1020 dimensions). All structures were least squares fitted on the C_α atoms of the monomer 1 at time 0. Since there are 101 structures in each part, there are only $100-6=94$ nonzero eigenvalues. The eigenvalues of the covariance matrix for time 1–2 ns are shown in Fig. 5. The eigenvalues decay with approximately a power of -1.1. The first five principal components are shown in Fig. 6. The first four resemble cosines with the number of periods equal to half the eigenvalue index. Thus the sampling in these directions is far from complete. The principal components resemble those of the Brownian system, which suggests that the protein might be diffusing on a part of the free-energy landscape that is almost flat in a few dimensions. If this is the case, then these eigenvectors do not describe relevant motions, but only give an indication in which directions the protein is more free to move.

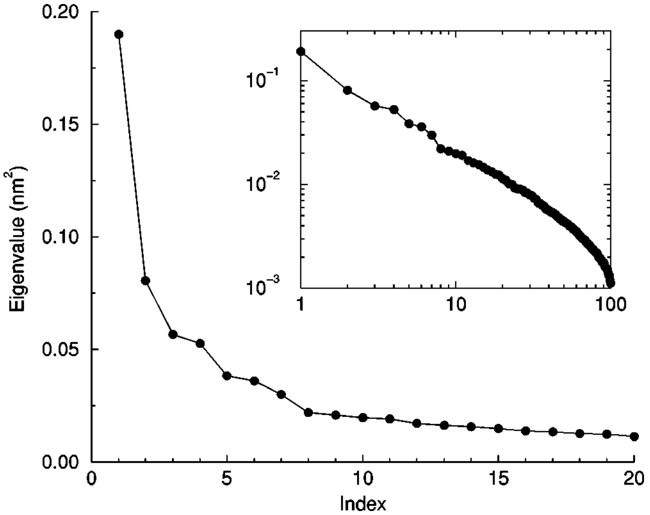

FIG. 5. Eigenvalues of the covariance matrix of subunit 1 of an OMPf trimer. The analysis was done over time 1–2 ns of the MD simulation.

To check the relevance of the eigenvectors, we can look at the overlap between subspaces spanned by eigenvectors

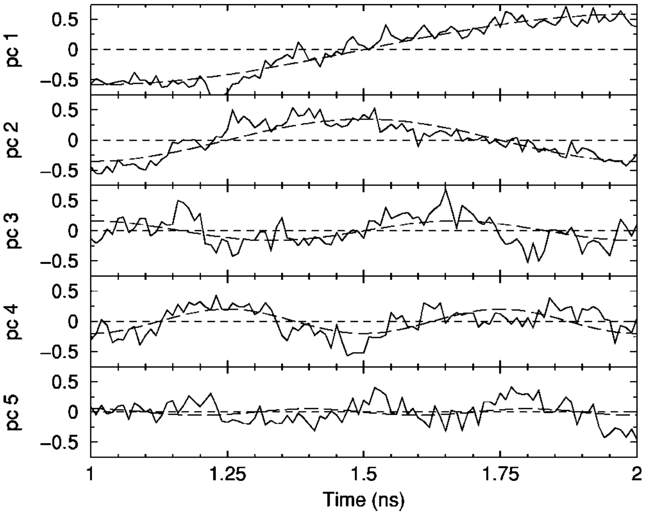

FIG. 6. The first five principal components of subunit 1 of the OMPf trimer. The analysis was done over time 1–2 ns of the MD simulation. Each principal components is fitted to a cosine with number of periods equal to one-half the index. In nm.

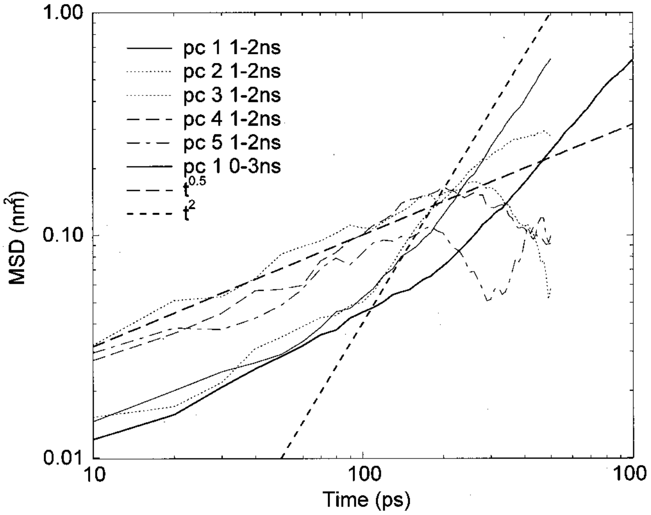

FIG. 7. The mean square displacement of the first five principal components of subunit 1 of the OMPf trimer. The analysis was done over time 1–2 ns of the MD simulation. The first principal component for the analysis over 0–3 ns is shown as well. These principal components are subdiffusive on short time scales. The first two principal components show ballistic behavior, due to their cosine shape (see Fig. 6).

for different parts of the simulation. The overlap between two sets of n orthonormal vectors $\mathbf{v}_1,\ldots,\mathbf{v}_n$ and $\mathbf{w}_1,\ldots,\mathbf{w}_n$ can be quantified as follows:

$$\text{overlap}(\mathbf{v},\mathbf{w})=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{v}_i\cdot\mathbf{w}_j)^2. \tag{54}$$

When sets $\mathbf{v}$ and $\mathbf{w}$ span the same subspace, the overlap is 1. The overlap between the subspace spanned by the first 10 eigenvectors is 0.13 between the first and second nanosecond, 0.16 between the second and third nanosecond, and 0.12 between the first and third nanosecond. This means that 1 ns is not enough to get a good impression of the conformational freedom of this protein. We have also calculated the mean square displacement along eigenvectors; these are shown in Fig. 7. On time scales below 100 ps the behavior is subdiffusive, which is reasonable for a set of connected atoms on a relatively flat part of the free-energy landscape. On longer time scales eigenvectors 1 and 2 go ballistic. This is not a property of the system, but rather an artifact of the short simulation time in combination with the covariance analysis, which filters ballistic motions out of a diffusive system. This can be illustrated by doing the same analysis for the whole 3 ns. Again the principal components look like cosines, but now the wavelength is three times as long. Thus the cosine behavior is linked to the analysis interval and not to inherent properties of the protein. The mean square displacement looks similar as well, but the transition to ballistic motion is shifted by approximately a factor of 3 (see Fig. 7).

β hairpin in water

The second system is a 16-residue β hairpin solvated in a box of 1414 SPC water molecules and three sodium ions. This system was simulated for 10 ns; structures were written to trajectory every 0.1 ps. A complete description of the simulation setup can be found in [13]. We performed the covariance analysis on the 24 001 structures from 5.3 to 7.7 ns. We chose this period because the peptide seems to reside in one free-energy minimum for these 2.4 ns. Because the termini are very flexible, only the backbone atoms of residues 2 to 15 were included in the analysis. The first five eigenvalues can be found in Table I; they contain 65% of the fluctuations and the first ten eigenvalues contain 81% of the fluctuations. The eigenvectors are well defined for this part of the simulation; the subspace overlap between the first and second 1.2 ns is 0.89 for the first five eigenvectors. The first five principal components are shown in Fig. 8. These look very noisy and resemble diffusion in a harmonic potential; their distributions are almost Gaussian. The force constants k for the harmonic well can be estimated from the eigenvalues and the temperature (see Table I). Kinetic information can be obtained from the autocorrelation functions of the principal components, shown in Fig. 9. There is a rapid drop in the first half picosecond followed by an approximately exponential decay. After 20 ps the curves level off, which is an indication that the peptide is not diffusing in a single harmonic well, but in several connected wells. The fast effect can also be found in the velocity autocorrelation functions of the eig-

TABLE I. Properties of the first five eigenvectors of the β-hairpin simulation. i is the eigenvector index, λ is the eigenvalue, $k=k_BT/\lambda$ is an estimate of the harmonic force constant in the direction of eigenvector i, a and τ are the result of an exponential fit $a\exp(-t/\tau)$ from 1 to 20 ps to the autocorrelation function of principal component i, and $\zeta=k\tau$ is a rough estimate of the friction coefficient.

i	λ (nm^2)	k $\left(\frac{\text{kJ}}{\text{mol nm}^2}\right)$	a	τ (ps)	ζ $\left(\frac{\text{amu}}{\text{ps}}\right)$
1	17.1×10^{-3}	146	0.81	19	2.9×10^{3}
2	10.8×10^{-3}	236	0.78	17	4.1×10^{3}
3	9.9×10^{-3}	252	0.73	14	3.4×10^{3}
4	5.8×10^{-3}	429	0.66	22	9.4×10^{3}
5	4.2×10^{-3}	600	0.38	20	12.2×10^{3}

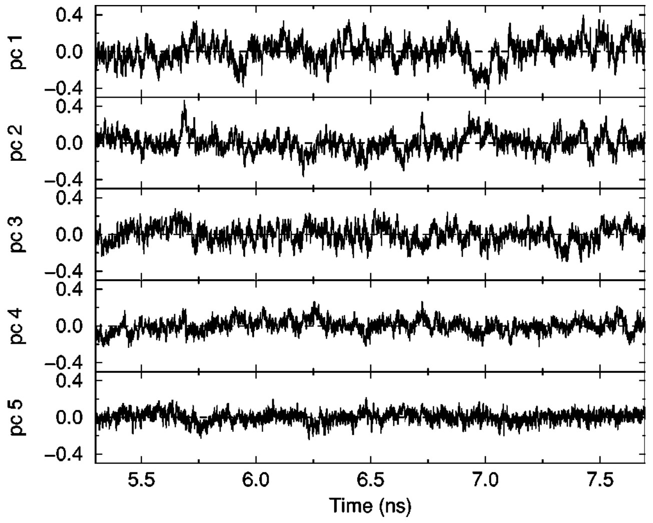

FIG. 8. The first five principal components of the β hairpin. In nm.

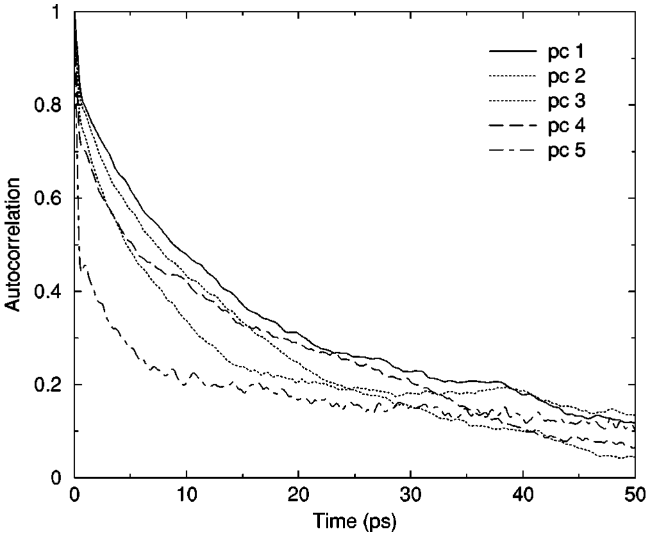

FIG. 9. The autocorrelation of the first five principal components of the β hairpin.

envectors functions; these have a negative peak at 0.3 to 0.4 ps. This time scale corresponds to the momentum effect, which should occur on a time scale of $\sqrt{m/k}$, where m is approximately 12 amu. The long time scale motions are overdamped, since the velocity autocorrelation is almost zero after a few picoseconds. In a diffusive system the correlations decay exponentially with $\tau=\zeta/k$, where ζ is the friction coefficient. We can estimate the friction coefficient for the first three eigenvectors by fitting the autocorrelation function with an exponential from 1 to 20 ps (see Table I).

HPr in water

The last system is the 85-residue protein HPr (histidine-containing protein) [18] simulated in a box of 5315 SPC water molecules. The total simulation time was 5 ns; structures were written to trajectory every picosecond. We left the first nanosecond out of the analysis to remove equilibration effects. The first five and ten eigenvalues contain 63% and 74% of the fluctuations, respectively. The first five principal components show plateaus with distinct jumps at 2.5 and 4.2 ns (Fig. 10). The global shape of the first and second eigenvectors resemble a half and a full cosine. The jumping behavior can also be observed in the root mean square deviation (RMSD) matrix (Fig. 11), which has several light triangles along the diagonal. The structures can be clustered using RMSD as a distance measure, for instance, a Jarvis-Patrick algorithm. In this algorithm a list of the nine closest neighbors is constructed for each structure. Two structures are in the same cluster when they are in each other's list and have at least three list members in common. The algorithm finds three large clusters (Fig. 11). The first five eigenvectors mainly describe the difference between the clusters. Because of the jumping behavior the overlap between the first and second halves of the 4 ns is only 0.32 and 0.46 for the subspace spanned by the first five and ten eigenvectors, respectively. When the simulation is prolonged, jumps to new clusters might occur, which can cause large changes in the eigenvectors.

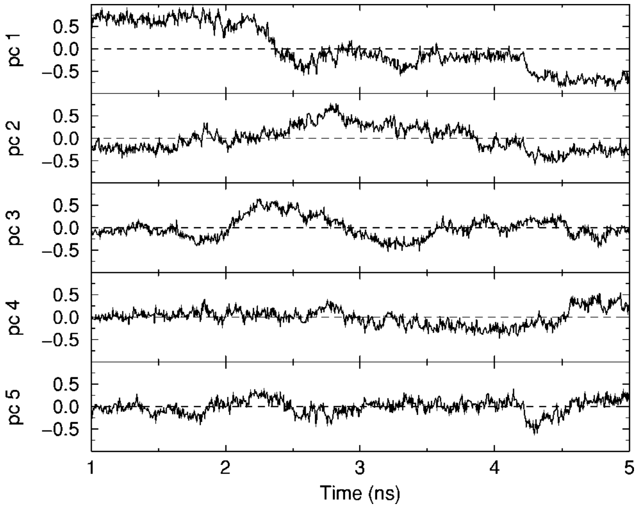

FIG. 10. The first five principal components of HPr. In nm.

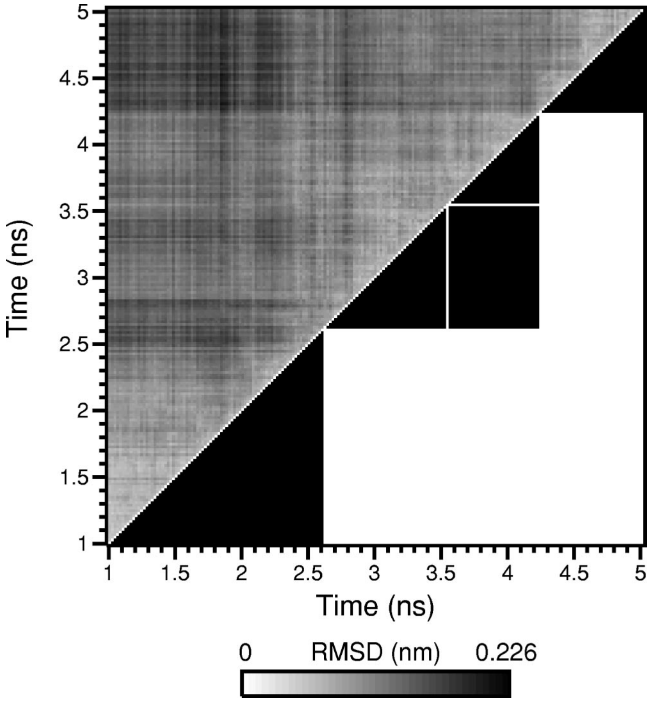

FIG. 11. Matrix showing the RMS deviation of the C_α's of HPr for each pair of structures in the upper left half. The lower right half shows the result of Jarvis-Patrick clustering. A black dot means that two structures are in the same cluster. There are three large clusters.

DISCUSSION

Principal component analysis is considered to be a powerful tool for finding large scale motions in proteins. The advantages of the analysis do not lie in the analysis itself, but in the fact that most of the fluctuations can be captured with the first few principal modes. This means that many analyses can be performed in only a few dimensions, which makes visual inspection of the results easier. This can reveal features of the free-energy landscape. Kinetic properties can also be analyzed, since all the time information is still present in the principal components.

However, one cannot conclude from the eigenvalues alone that only a small subspace is important. We have shown that for high-dimensional random diffusion, the first three eigenvalues contain 84% of the fluctuations. This does not mean that the first three principal modes reveal special properties of the system. In random diffusion all directions are equivalent; a second simulation will produce completely different principal modes. How can the relevance of the principal modes be determined? One should always divide the simulation into two or more parts and compare the principal modes for each part. A simple measure is the subspace overlap of the first few principal modes [Eq. (54)]. Amadei *et al.* extensively analyzed the overlap for molecular dynamics MD simulations of protein L and cytochrome C [14]. They found that the overlap for the first 10 principal components is 0.6 for two parts of 50 ps, irrespective of the time interval between the two parts. This indicates that the protein samples only one free-energy minimum, as is the case for the β-hairpin simulation. The overlap can be lower when the protein samples multiple minima, as was shown for HPr. In that case the principal modes describe the differences between the minima, rather than the shape of the energy minima. Although the subspace of the first five or ten principal modes might not be well defined, there will always be a splitting in a subspace of diffusive modes, which contain most of the fluctuations and a subspace of (near-)harmonic modes. This splitting will not change significantly between different simulations.

We have shown that the first few principal components of high-dimensional random diffusion are a half cosine, a full cosine, one and one-half cosine, etc. Many protein simulations with similar principal components can be found in the literature, for instance [5,15,16]. When the trajectory is projected on the plane of the first two principal components, the half and full cosine produce a characteristic semicircle, as can been seen for bovine pancreatic trypsin inhibitor (BPTI) simulations in [17]. Often these simulations are relatively short (a few nanoseconds) and of relatively large proteins (more than 200 residues), like the OMPf simulation presented in this article. It is very improbable that properties of the protein will produce cosine-shaped atomic displacements, consisting of exactly a half period, a full period, etc. It is more likely that this behavior is the result of random diffusion, since the simulation time is not long enough to repeatedly reach barriers on the free-energy landscape. The first principal component, a half cosine, can be very misleading, as it seems if the protein moves from one ''well-defined'' conformation to another. The cosine-shaped principal components lead to ballistic behavior on longer time scales. García *et al.* [15] write that this behavior belongs to the ''Lévy flight'' class. However, this behavior is caused by the incomplete sampling and scales with the simulation time, as shown for OMPf, and it will disappear when a free-energy minimum has been sampled to a reasonable extent, as shown for the β hairpin. When one sees cosinelike principal components, one should interpret the results carefully and always keep in mind that most of the fluctuations could be caused by random diffusion. The analysis is still useful, because it can separate the ''random diffusion'' degrees of freedom from the more restricted degrees of freedom. In such cases insight into the principal modes of random diffusion in a harmonic potential will make better estimates of the conformational freedom possible. In future work we hope to present simulations of proteins of reasonable size, which are long enough to analyze the convergence.

ACKNOWLEDGMENTS

The author thanks D. P. Tieleman and K. A. Feenstra for providing the simulations of OMPf and HPr, A. Morozov for checking $E[(B_{ij})^2]$, and H. Grubmüller, H. J. C. Berendsen, and J. G. E. M. Fraaije for stimulating discussions.

APPENDIX

The variance of the integral over $f(t)x_i(t)$ can be calculated, given that

$$\int_0^T f(t)dt = 0. \tag{A1}$$

We can apply partial integration

$$\begin{aligned}
\mathrm{E}\left[\left(\int_0^T f(t)x_i(t)dt\right)^2\right] &= \mathrm{E}\left[\left(-\int_0^T\int_0^t f(v)dv\,\frac{dx_i(t)}{dt}dt + \int_0^t f(v)dv\,x_i(t)\bigg|_{t=0}^{t=T}\right)^2\right]\\
&= \mathrm{E}\left[\left(-\int_0^T\int_0^t f(v)dv\,\frac{dx_i(t)}{dt}dt\right)^2\right]\\
&= \mathrm{E}\left[\int_0^T\int_0^t f(v)dv\,\frac{dx_i(t)}{dt}dt\int_0^T\int_0^u f(w)dw\,\frac{dx_i(u)}{du}du\right]\\
&= \int_0^T\int_0^T\int_0^t f(v)dv\int_0^u f(w)dw\,\mathrm{E}\left[\frac{dx_i(t)}{dt}\,\frac{dx_i(u)}{du}\right]du\,dt\\
&= \int_0^T\int_0^T\int_0^t f(v)dv\int_0^u f(w)dw\,2D\,\delta(t-u)du\,dt\\
&= 2D\int_0^T\left(\int_0^t f(u)du\right)^2 dt.
\end{aligned} \tag{A2}$$

The expectations of c_i^k and $c_i^k c_j^l$ can be calculated using integration by parts:

$$E[c_i^k]=E\left[\sqrt{\frac{2}{T}}\int_0^T x_i(t)\cos\left(\frac{\pi kt}{T}\right)dt\right]=E\left[-\frac{\sqrt{2}}{\sqrt{T}\pi k}\int_0^T\frac{dx_i(t)}{dt}\sin\left(\frac{\pi kt}{T}\right)dt\right]=-\frac{\sqrt{2}}{\sqrt{T}\pi k}\int_0^T E\left[\frac{dx_i(t)}{dt}\right]\sin\left(\frac{\sqrt{T}\pi kt}{T}\right)dt=0. \tag{A3}$$

For $k=1,2,\ldots$ and $l=1,2,\ldots,$

$$\begin{aligned}
E[c_i^k c_j^l]&=E\left[\frac{2}{T}\int_0^T x_i(s)\cos\left(\frac{\pi ks}{T}\right)ds\int_0^T x_j(t)\cos\left(\frac{\pi lt}{T}\right)dt\right]\\
&=E\left[\frac{2T}{\pi^2 kl}\int_0^T\frac{dx_i(s)}{dt}\sin\left(\frac{\pi ks}{T}\right)ds\int_0^T\frac{dx_j(t)}{dt}\sin\left(\frac{\pi lt}{T}\right)dt\right]\\
&=\frac{2T}{\pi^2 kl}\int_0^T\int_0^T E\left[\frac{dx_i(s)}{dt}\,\frac{dx_j(t)}{dt}\right]\sin\left(\frac{\pi ks}{T}\right)ds\,\sin\left(\frac{\pi lt}{T}\right)dt\\
&=\frac{2T}{\pi^2 kl}\int_0^T\int_0^T 2D\,\delta(s-t)\,\delta_{ij}\sin\left(\frac{\pi ks}{T}\right)ds\,\sin\left(\frac{\pi lt}{T}\right)dt\\
&=\frac{4DT}{\pi^2 kl}\,\delta_{ij}\int_0^T\sin\left(\frac{\pi kt}{T}\right)\sin\left(\frac{\pi lt}{T}\right)dt\\
&=\frac{2DT^2}{\pi^2 kl}\,\delta_{ij}\delta_{kl}\,.
\end{aligned}\tag{A4}$$

The size of the off-diagonal elements of B is

$$\begin{aligned}
E[(B_{ij})^2]&=E\left[\left(\frac{\pi^2 ij}{2NDT^3}\sum_{l=1}^{N}c_l^i\sum_{m=1}^{N}c_m^j\sum_{k=1}^{\infty}c_l^k c_m^k\right)^2\right]\\
&=\left(\frac{\pi^2 ij}{2NDT^3}\right)^2 E[(U+V_{ij}+V_{ji})^2]\\
&=\left(\frac{\pi^2 ij}{2NDT^3}\right)^2(E[U^2]+E[V_{ij}^2]+E[V_{ji}^2]+2\mathrm{E}[UV_{ij}]+2E[UV_{ji}]+2E[V_{ij}V_{ji}]),
\end{aligned}\tag{A5}$$

where

$$U=\sum_{l=1}^{N}c_l^i\sum_{m=1}^{N}c_m^j\sum_{k=1}^{\infty}c_l^k c_m^k(1-\delta_{ik})(1-\delta_{jk}), \tag{A6}$$

$$V_{ij}=\sum_{l=1}^{N}c_l^{i^2}\sum_{m=1}^{N}c_m^j c_m^i\,. \tag{A7}$$

$$E(U^2)=E\left[\sum_{l=1}^{N}c_l^{i^2}\sum_{m=1}^{N}c_m^{j^2}\sum_{k=1}^{\infty}c_l^{k^2}c_m^{k^2}(1-\delta_{ik})(1-\delta_{jk})\right]=\left(\frac{2DT^2}{\pi^2}\right)^2\frac{N^2+O(N)}{i^2 j^2}\left(\frac{\pi^4}{90}-\frac{1}{i^4}-\frac{1}{j^4}\right), \tag{A8}$$

$$E[V_{ij}^2]=E\left[\left(\sum_{l=1}^{N}c_l^{i^2}\right)^2\sum_{m=1}^{N}c_m^{j^2}c_m^{i^2}\right]=\left(\frac{2DT^2}{\pi^2}\right)^2\frac{N^3+2N^2(3-1)+O(N)}{i^6 j^2}, \tag{A9}$$

$$E[UV_{ij}]=E\left[\sum_{l=1}^{N}c_l^{i^2}\sum_{m=1}^{N}c_m^{j^2}c_m^{i^2}\left(\sum_{k=1}^{\infty}c_m^{k^2}-c_m^{i^2}-c_m^{j^2}\right)\right]=\left(\frac{2DT^2}{\pi^2}\right)^2\frac{N^2+O(N)}{i^4 j^2}\left(\frac{\pi^2}{6}-\frac{1}{i^2}-\frac{1}{j^2}\right), \tag{A10}$$

$$E[V_{ij}V_{ji}]=E\left[\sum_{l=1}^{N} c_l^{i^2}\sum_{m=1}^{N} c_m^{j^2}c_m^{i^2}\sum_{k=1}^{N} c_k^{j^2}\right]=\left(\frac{2DT^2}{\pi^2}\right)^2\frac{N^3+2N^2(3-1)+O(N)}{i^4j^4}, \tag{A11}$$

$$E[(B_{ij})^2]=4D^2T^2\left[\frac{1}{90}+\frac{1}{3\pi^2}\left(\frac{1}{i^2}+\frac{1}{j^2}\right)+\frac{N+3}{\pi^4}\left(\frac{1}{i^4}+\frac{1}{j^4}\right)+\frac{2N+4}{i^2j^2}+O\left(\frac{1}{N}\right)\right]. \tag{A12}$$

[1] S. Hayward and N. Gō, Annu. Rev. Phys. Chem. **46**, 223 (1995).

[2] A. Kitao and N. Gō, Curr. Opin. Struct. Biol. **9**, 164 (1999).

[3] M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).

[4] A. E. García, Phys. Rev. Lett. **68**, 2696 (1992).

[5] A. E. García and J. G. Harman, Protein Sci. **5**, 62 (1996).

[6] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Struct., Funct., Genet. **17**, 412 (1993).

[7] A. B. M. Linssen. Ph.D. thesis, University of Groningen, the Netherlands, 1998. Available at http://www.ub.rug.nl/eldoc/dis/science/a.b.m.linssen

[8] A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1965).

[9] Z. Bai and J. W. Silverstein, Ann. Prob. **27**, 1536 (1999).

[10] D. van der Spoel, B. Hess, K. A. Feenstra, E. Lindahl, and H. J. C. Berendsen, GROMACS User Manual version 2.0, 1999. Nijenborgh 4, 9747 AG Groningen, The Netherlands. Available at http://md.chem.rug.nl/~gmx

[11] D. Tieleman and H. A. Berendsen, Biophys. J. **74**, 2786 (1998).

[12] D. P. Tieleman, H. J. C. Berendsen, and M. S. P. Sansom, Biophys. J. **76**, 1757 (1992).

[13] D. Roccatano, A. Amadei, A. D. Nola, and H. J. C. Berendsen, Protein Sci. **8**, 1 (1999).

[14] A. Amadei, M. A. Ceruso, and A. D. Nola, Proteins: Struct., Funct., Genet. **36**, 419 (1999).

[15] A. E. García, R. Blumenfeld, and G. Hummer, Physica D **107**, 225 (1997).

[16] G. H. Peters, T. M. Frimurer, J. N. Andersen, and O. H. Olsen, BTbj **78**, 2191 (2000).

[17] P. Eastman and M. Pellegrini, J. Chem. Phys. **110**, 10 141 (1999).

[18] N. A. J. van Nuland *et al.*, J. Mol. Biol. **237**, 544 (1994).