JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

Consistent principal component modes from molecular dynamics simulations of proteins

Rodrigo Cossio-Pérez, Juliana Palma, and Gustavo Pierdominici-Sottile

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.6b00646 • Publication Date (Web): 16 Mar 2017

ACS Publications

Journal of Chemical Information and Modeling is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

CONSISTENT PRINCIPAL COMPONENT MODES FROM MOLECULAR DYNAMICS SIMULATIONS OF PROTEINS

Rodrigo Cossio-Pérez, Juliana Palma,* and Gustavo Pierdominici-Sottile

Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Sáenz Peña 352, B1876BXD Bernal, Argentina.

E-mail: juliana@unq.edu.ar

Abstract

Principal component analysis is a technique widely used for studying the movements of proteins using data collected from molecular dynamics simulations. In spite of its extensive use the technique has a serious drawback: equivalent simulations do not afford the same PC-modes. In this article we show that concatenating equivalent trajectories and calculating the PC-modes from the concatenated one significantly enhances the reproducibility of the results. Moreover, the consistency of the modes can be systematically improved by adding more individual trajectories to the concatenated one.

*To whom correspondence should be addressed

Introduction

The understanding of how proteins work needs a joint description of their dynamical and structural characteristics. Molecular dynamics (MD) simulations constitute a powerful approach to investigate their dynamical features. Within this framework, the use of principal component analysis (PCA) has emerged as one of the most widely employed techniques to analyze protein movements.[1–4] The methodology was introduced in MD studies of proteins by Karplus *et. al.*,[5,6] but it has also been extensively used in other branches of science and technology.[7,8]

In MD studies, PCA is mainly applied to change from a description based on local atomic coordinates to one provided by collective coordinates, called the PC-modes. They describe the simultaneous motion of different parts of the protein. The PC-modes are the eigenvectors of the covariance matrix, which is calculated with the configurations sampled from a MD trajectory.[1] A key feature of PCA is that only a bunch of PC-modes, those associated with the highest eigenvalues, account for approximately 70% to 90% of the total fluctuations of the protein.[9–11] This allows for a huge reduction in the number of degrees of freedom required to indicate the deformations of the system: just a few PC-modes provide a description equivalent to hundreds or thousands of atomic coordinates. The subspace formed by the PC-modes associated to the largest eigenvalues is called the essential space (ES). It is usually assumed that the motions related to the biological function of a protein are contained within its ES.[12–15] The remaining PC-modes account for irrelevant, small-amplitude fluctuations. It is said that they span the "near-constraint subspace" which is normally of no interest.[16]

In spite of the many examples in which PCA has proved to be useful, the PC-modes calculated by the standard procedure have an undesirable characteristic that casts doubts on their actual significance and utility: equivalent simulations do not afford the same PC-modes.[17–19] This lack of reproducibility can be assessed by calculating the inner product between PC-modes obtained by equivalent but independent MD simulations. Ideally the

absolute value of these inner products should be 1.0 for modes having the same index and zero otherwise. A less restrictive condition consists of requiring that the ESs obtained by equivalent simulations describe the same subspace. The parameter that evaluates the overlap between such subspaces is called the root mean squared inner product (RMSIP), and has been widely used to assess the convergence of the main PC-modes of proteins. The RMSIP is equal to 1.0 if the two ESs span the same subspace while it is zero if they are orthogonal. It is also a common practice to calculate the RMSIP with the ESs obtained from different halves of the same trajectory, using increasingly longer simulation times. This is done to evaluate the convergence of the main PC-modes with respect to the length of the simulation. This test has been applied to many different systems and in all cases the same qualitative result was obtained.[20–26] Initially the RMSIP grows fast, but then it levels out reaching a plateau value which is sensibly smaller than 1.0. This behavior indicates that extending the simulation time, within the ranges typically used in current MD simulations, is not effective to improve the convergence of the ES.

Many studies have focused on evaluating the main features of protein PC-modes.[27–31] They comprise from small soluble proteins to large membrane proteins, and from short simulations of just a few ns to relatively long simulations of more than 50 ns. In all cases it was found that the main PC-modes obtained by equivalent trajectories were different, and that the RMSIP computed from them was smaller than one. In this article we show that the consistency of the PC-modes can be improved by employing a correlation matrix obtained by concatenating independent but equivalent trajectories. The performance of the procedure is demonstrated by applying it to the principal component analysis of bovine pancreatic trypsin inhibitor (BPTI) and lyzozyme.

The use of PCA of concatenated trajectories was introduced by Berendsen and co-workers in 1995.[32] Since then, it has been widely used as one of the diverse tools employed to characterize the most relevant motions of proteins and other systems of biological interest. However, the results of those studies were interpreted on intuitive foundations since analytical formulas

showing the precise meaning of the eigenvectors and eigenvalues obtained by concatenated-PCA had not been provided. Recently we presented such analytical expressions and discussed that two opposite limits could be found.[24] One extreme case occurs when the trajectories belong to two or more free energy minima, and the root mean square deviations between these minima are significantly larger than the typical fluctuations around them (i.e. the trajectories cannot go from one minimum to the other). This case was thoroughly analyzed in Ref. 24. The other extreme occurs when the concatenated trajectories are initiated and remain within the same free energy well. The method proposed in this article is aimed to univocally determine the main PC-modes in this case. We note that only in this situation the ES of a protein contains a set of collective coordinates useful to describe the most important fluctuations. When multiple minima are implied the main PC-modes contain the so-called "static contribution".[24] It can partially or completely mask the "dynamic contribution" of protein fluctuations.

Concatenated trajectories have also been used for other purposes. In particular, they were employed in different methodologies developed to characterize protein folding processes. In these methods, the concatenation has to be performed following specific prescriptions aimed to accelerate the conformational sampling. Important examples of these applications can be found in Ref. 33 - 34 and the references cited wherein. The analysis of the PC-modes obtained with this kind of trajectories lies outside the scope of this article. Instead we concentrate on the characterization of the dynamics of stable conformations of proteins, whose main PC-modes have a clear and straightforward interpretation as collective coordinates useful to describe their most important fluctuations.

Materials and Methods

Statement of the problem

In applications of PCA to the study of protein dynamics, the PC-modes are obtained by diagonalizing a correlation matrix whose elements are given by,

$$C_{ij} = \frac{1}{N}\sum_{k=1}^{N}(x_i^k - \overline{x}_i)(x_j^k - \overline{x}_j). \quad (1)$$

Here x_i^k, x_j^k, are a pair of elements of vector $\mathbf{x}^k$, which describes the configuration of the system at time step k, while $\overline{x}_i$ and $\overline{x}_j$ are their average values calculated from the N structures sampled in the MD simulation. Normally, $\mathbf{x}$ is a vector containing the Cartesian coordinates of the C_α atoms of the protein, but other choices can also be used.[35–37] During the setting of any MD simulation several parameters are chosen at random. Besides, the sampling interval and the time at which the first structure is recorded are arbitrarily decided. Therefore the atomic coordinates of the selected structures are random numbers too, as are random the C_{ij} coefficients calculated from them.

For an infinite long simulation, from which an infinite number of samples could be taken, the C_{ij} coefficients would assume perfectly defined values, free of statistical errors. However, infinite long simulations are not possible and the experience indicates that currently feasible simulations are not long enough to produce correlation matrices with sufficiently small statistical uncertainties. Thus, if $\mathbf{C}^\infty$ is the correlation matrix corresponding to an infinite long sampling and $\mathbf{C}$ is a correlation matrix computed from a finite number of samples, the relation between them can be written as,

$$\mathbf{C} = \mathbf{C}^\infty + \mathbf{E}, \quad (2)$$

where the elements of matrix $\mathbf{E}$ contain the statistical errors of the correlation coefficients computed from the sample. Since the elements of $\mathbf{E}$ are different than zero, the eigenvalues and eigenvectors of matrix $\mathbf{C}$ differ from those of the actual correlation matrix $\mathbf{C}^\infty$. These discrepancies can be estimated if the elements of $\mathbf{E}$ are small enough and fulfill some defined characteristics (i.e. they are normally distributed).[38] Such estimations have successfully been

done in applications of PCA to other branches of science, technology and economy.[39,40]

Recognizing that infinite sampling is not possible, one faces the problem of determining which sampling procedure minimizes the statistical error of the results. The use of increasingly longer simulations is normally employed but, as already described in the introduction, such strategy does not lead to the desired result. Alternatively, one can use several equivalent MD simulations that just differ from each other in the initial velocities of the atoms. We recently demonstrated that the correlation matrix $\mathbf{C}^{(n)}$, obtained by concatenating n trajectories with the same number of samples can be decomposed as,[24]

$$\mathbf{C}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_i + \mathbf{S}^{(n)}. \tag{3}$$

Here $\mathbf{C}_i$ is the correlation matrix corresponding to the i-th trajectory while $\mathbf{S}^{(n)}$ is the correlation matrix computed from the n average structures. If the individual MD simulations were able to sample all regions of the accessible configurational space, trajectories that just differ in their initial atomic velocities would produce almost the same average structures. In this case matrix $\mathbf{S}^{(n)}$ should be significantly smaller than the $\mathbf{C}_i$ matrices, because the deviation of the individual average structures with respect to the global average would be much smaller than the fluctuations observed in any single trajectory. Under such conditions, the correlation matrix of the concatenated trajectory becomes quite close to the average of the individual correlation matrices, $\mathbf{C}^{(n)}_{av}$.

$$\mathbf{C}^{(n)} \approx \frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_i = \mathbf{C}^{(n)}_{av}. \tag{4}$$

According to the Classical Central Limit Theorem the last equality implies that for large values of n, the statistical uncertainty in the elements of $\mathbf{C}^{(n)}_{av}$ will be smaller than those of the $\mathbf{C}_i$'s by a factor of about $1/\sqrt{n}$. Thus, one is induced to think that concatenating trajectories should provide a route for obtaining reproducible PC-modes.

However, it could also happen that the single trajectories are relatively short or get

trapped and sample only a portion of the available configurational space. In such case, the individual averages will be disperse with respect to the global average structure and matrix $\mathbf{S}^{(n)}$ will not be negligible. Even in this apparently adverse condition, the correlation matrix of the concatenated trajectory can still be equal to the average of some conveniently-defined correlation matrices $\mathbf{C}_i$. To see how to achieve this, one has to recognize that matrix $\mathbf{C}^{(n)}$ remains unchanged with respect to any permutation of the structures used to compute it. Usually, structures 1 to N correspond to the first simulation, structures $N+1$ to $2N$ to the second one, and so on. However one can shuffle the nN structures employed to calculate $\mathbf{C}^{(n)}$ and then divide them into n sets of N arbitrarily-selected structures. Any of these new sets will have structures originated from different MD simulations. Therefore, for sufficiently large n and N, the average structures of these sets will be pretty similar to each other, and the new $\mathbf{S}^{(n)}$ matrix will be negligible with respect to the new $\mathbf{C}_{av}^{(n)}$. Thus, the shuffling procedure does not affect $\mathbf{C}^{(n)}$ but it changes $\mathbf{C}_{av}^{(n)}$ and $\mathbf{S}^{(n)}$, in such a way that their changes mutually compensate. When the structures sampled from a single MD simulation are biased, the correlation matrix computed from them underestimates the actual correlations. This occurs because the deviations of the sampled structures with respect to their own average are smaller than their deviations with respect to the true average. For the same reason, also matrix $\mathbf{C}_{av}^{(n)}$ underestimates the actual correlations. The calculation of matrix $\mathbf{C}^{(n)}$ corrects this error because the correlations that get lost in $\mathbf{C}_{av}^{(n)}$ appear in $\mathbf{S}^{(n)}$. We therefore conclude that, even if the individual trajectories perform a biased sampling of the available configurational space, matrix $\mathbf{C}^{(n)}$ converges to an average of n conveniently-defined individual correlation matrices. Accordingly, its statistical uncertainly is reduced by a factor of $1/\sqrt{n}$.

In what follows, we will refer to matrix $\mathbf{C}^{(n)}$ as the correlation matrix of the concatenated trajectory. However, from the previous discussion, it should be clear that there is not a real need to "concatenate" the trajectories, since any order of the whole set of structures produces the same result. The key point here is to compute the correlation matrix using structures

sampled from multiple equivalent trajectories.

Molecular dynamics simulations

Principal component analysis of BPTI and lysozyme were used to test the consistency of the PC-modes. The initial coordinates of the proteins were obtained from the Protein Data Bank, ID=5PTI for BPTI[41] and 1REX for lysozyme.[42] The systems were solvated in a truncated octahedral cell of TIP3P explicit water molecules and minimized at constant volume. In a second stage they were heated at constant volume from 0 K to 310 K during 1 ns, using the weak coupling algorithm with τ_{P}=2.8 ps. After that, we switched to constant temperature and pressure conditions using a value of 2.0 ps for both, τ_{TP} and τ_{P}. Finally, an equilibration run of 10 ns was performed. For each system, the final structure of the equilibration stage was used as the initial configuration of the production runs. We run 180 equivalent trajectories of 5 ns and 80 trajectories of 50 ns for each system. These trajectories just differed in the initial velocities of the atoms, which were chosen from a maxwellian distribution at 310 K. Snapshots were taken every 25 ps in the 5-ns trajectories and every 250 ps in the 50-ns trajectories.

The simulations were performed with the AMBER 14 package using the ff99SB force field, applying periodic boundary conditions with a cutoff of 12.0 Å. The shake algorithm was employed to maintain bond distances to hydrogen, allowing for a time step of 2.0 fs. We tested that the projections of the individual trajectories onto their first two PC-modes did not resemble cosine functions.[43] Also, it was checked that the RMSIP calculated from different halves of the same trajectory was converged with respect to the simulation time.[20] Thus, any of the individual simulations employed in this work would pass the convergence assessments usually applied in PCA studies of protein.

Measurements of the convergence

The convergence of PCA is usually evaluated by comparing the PC-modes obtained from a reference trajectory against the ones derived from trial trajectories whose convergence one is trying to evaluate. Normally the reference trajectory is much longer than the trial trajectories or it has been obtained with procedures that improve the conformational sampling. This kind of assessment is based on the assumption that the reference trajectory provides a fair enough sampling, so that its PC-modes are already converged. Here we apply a criterion that does not require the *a-priori* knowledge of such reference trajectory. Instead, we consider that the PC-modes or the essential space of a protein are converged when two alternative but otherwise equivalent computations afford the same PC-modes or ESs. The similarity between the PC-modes and ESs of the alternative computations were measured using different parameters. They are described in the following paragraphs.

We analyzed the absolute value of the scalar product between corresponding PC-modes calculated from alternative simulations a and b, $|\mathrm{PC}_i^a \cdot \mathrm{PC}_i^b|$, with particular emphasis on the first mode, PC_1. We also employed the RMSIP, which measures the common portion of the ESs determined from the pair MD simulations,

$$RMSIP_M = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M}|\mathrm{PC}_i^a.\mathrm{PC}_j^b|^2}. \quad (5)$$

Here M denotes the dimension of the subspaces while PC_i^a and PC_j^b are the i-th and j-th eigenvectors obtained from simulations a and b, respectively. In our analysis, we set $M = 2$ since it is a common practice to analyze protein motions in the subspace spanned by the first two eigenvectors.[18,25,28,44–46] It should be noted that the use of $M = 2$ makes the evaluation as strict as possible. Finally, for each pair of trajectories, we also evaluated the covariance overlap, s, proposed by Hess.[43] The overlap is not a measure of the convergence of the essential space. Instead, it assesses the similarity of the spaces sampled from a given pair of

trajectories. The overlap is defined as,

$$s = 1 - d_N\left(\mathbf{C}_a, \mathbf{C}_b\right), \tag{6}$$

where d_N is the normalized distance between the correlation matrices $\mathbf{C}_a$ and $\mathbf{C}_b$. The normalized distance is calculated as

$$d_N\left(\mathbf{C}_a, \mathbf{C}_b\right) = \left[\frac{tr\left(\mathbf{C}_a + \mathbf{C}_b - 2\mathbf{C}_a^{1/2}\mathbf{C}_b^{1/2}\right)}{tr\left(\mathbf{C}_a\right) + tr\left(\mathbf{C}_b\right)}\right]^{1/2}, \tag{7}$$

where $\mathbf{C}_\alpha^{1/2}$, the square root of correlation matrix $\mathbf{C}_\alpha$, is calculated as,

$$\mathbf{C}_\alpha^{1/2} = \mathbf{R}_\alpha\, \mathbf{\Lambda}_\alpha^{1/2}\mathbf{R}_\alpha^T. \tag{8}$$

In the last equation $\mathbf{R}_\alpha$ is the matrix that diagonalizes $\mathbf{C}_\alpha$ and $\mathbf{\Lambda}_\alpha$ is the diagonal matrix that contains its eigenvalues. When two independent simulations afford the same sampling, the two correlation matrices are the same, the distance between them is zero and the overlap amounts to 1.0. On the contrary, if the subspaces sampled in the simulations are orthogonal, the normalized distance evaluates to 1.0 and the overlap is zero.

Results

Each individual simulation was used to compute a set of PC-modes. Therefore, for each system, we obtained 180 sets of PC-modes with the 5-ns trajectories and 80 sets with the 50-ns trajectories. For both, BPTI and lysozyme, PC-mode sets computed from trajectories of equal length were grouped into pairs. All possible pairs were generated. Thus, we formed 16110 pairs with the trajectories of 5 ns and 3160 pairs with the trajectories of 50 ns. For each pair we calculated the absolute value of the inner product $|\mathrm{PC}_i^a \cdot \mathrm{PC}_i^b|$, the RMSIP_2 and the overlap. This allowed us to estimate reliable probability distributions for the three

parameters.

To evaluate the hypothesis that concatenating trajectories improves the reproducibility of the PC-modes we collected individual simulations of equal length into batches of n simulations, so that each trajectory was allocated into a single batch. Then, we concatenated the trajectories of a given batch and computed $\mathbf{C}^{(n)}$, the correlation matrix of the concatenated trajectory. Finally, the reproducibility of the PC-modes so obtained was assessed by computing $|\mathrm{PC}_i^a \cdot \mathrm{PC}_i^b|$, RMSIP_2 and overlap, for all possible pairs of batches formed with the given n. We tried different values of n. In Table 1 we present the alternative values of n employed in this work, along with the number of batches and the number of pairs of batches that can be formed with the given n.

PC modes from single trajectories of BPTI

Figure 1 shows the probability distributions for $|\mathrm{PC}_1^a \cdot \mathrm{PC}_1^b|$, calculated from all the independent simulations of BPTI. The vertical black line indicates the value of the inner product that corresponds to 99% of cumulative probability, for normalized random vectors of the same size. The cumulative probability was evaluated as,

$$P_{\mathrm{cum}}(x^*) = \int_0^{x^*} \rho(x)dx, \tag{9}$$

where $\rho(x)$ is the probability density that a random vector of dimension M has a square projection x onto a subspace of dimension m. For the present case $M = 58 \times 3 = 174$, since the model of BPTI has 58 residues, while $m = 1$ since we are considering the projection onto a single PC-mode. According to the equations provided by Amadei et. al. in Ref. 20, $\rho(x)$ is given by,

$$\rho(x) = \frac{(M-1)!}{(m-1)!(M-m-1)!} x^{(m-1)}(1-x)^{(M-m-1)}, \tag{10}$$

that in our case simplifies to $\rho(x) = (M-1)(1-x)^{(M-2)}$.

It is seen that the scalar product between the first PC-modes of individual MD simulations

affords significantly larger values than those expected for random vectors. However very low values, which imply almost orthogonal PC_1, are likely to be obtained too. Figure 1 also shows that the reproducibility of PC_1 is better for the trajectories of 50 ns than for those of 5 ns. However, even for these longer simulations, pretty low values of the scalar product are usually found. The distributions presented in Fig. 1 remind the ones reported by Grossfield *et. al.*[31] who analyzed the reproducibility of PC_1 for different membrane systems, using relatively long trajectories. Thus, Fig. 1 does not provide new evidence about the characteristics of PC_1 obtained from individual MD simulations, but just reinforces the conclusion that its direction is random to a large extent.

Fig. 2 presents typical examples for the inner-product matrices formed with the first six PC-modes of two equivalent simulations. Ideally, the out-of-diagonal elements should be null. The pictures presented in Fig. 2 differ from this ideal, indicating that the first PC-modes of a given simulation are mostly distributed among the first PC-modes of the other simulation. This characteristic has already been described for other systems.[27,47] Because of this behavior, it is more difficult to obtain a one-to-one correspondence between the individual PC-modes than to converge the subspace spanned by some of them. This is in fact the conclusion afforded by the probability distributions of $RMSIP_2$, presented in Fig. 3, which show that pretty low values of RMSIP are rather unlikely. However, the agreement for $RMSIP_2$ is still far from satisfactory since the most likely values are just $\approx$0.55 (5-ns trajectories) and $\approx$0.65 (50-ns trajectories). The probability distributions for the covariance overlap are also presented in Fig. 3. They are somewhat narrower than those of $RMSIP_2$ and their maxima are shifted to the left. In spite of these differences the conclusions attained from both distributions, $RMSIP_2$ and overlap, are similar. We finally note that the distributions obtained with the longer trajectories are shifted to the right of those computed with the shorter ones. Thus, increasing the simulation time helps to improve the reproducibility of the essential space but, as noted above, the results are still far from satisfactory.

PC modes from the concatenated trajectories of BPTI

Figure 4 shows the evolution of the overlap and RMSIP_2 as the number of simulations in the concatenated trajectory, n, is increased. For the short trajectories one can readily appreciate that both, overlap and RMSIP_2, raise very rapidly between 1 and 30 trajectories. Then there is a change of behavior and the two parameters variate more slowly. In particular, the overlap seems to level out at about $n = 60$ where it reaches a value of 0.88. The increase of RMSIP_2, on the other hand, does not stop after $n = 60$ but just becomes slower. For $n = 90$ we found a RMSIP_2 of 0.98 indicating that the subspaces spanned by the first two PC-modes of such concatenated trajectories are nearly the same. RMSIP_2 and overlap of the long trajectories also increase with n. However, in this case, the initial improvement is not so marked as in the case of the short trajectories. This is mainly because the two parameters start from higher average values. In this case, we found a RMSIP_2 of 0.98 for $n = 40$. For this n the overlap evaluates to 0.89.

The fact that the RMSIP_2 is almost fully converged while the overlap is not indicates that the first two PC-modes are mostly contained in the subspace shared by the two concatenated trajectories, while the orthogonal subspace accounts for the non-important PC-modes. Besides, this reveals that the first PC-modes converge faster than those corresponding to higher indexes. This is really a good characteristic since PCA is normally used to reduce the dimensionality of the system under analysis. The concatenated-PCA technique can reliably determine the subspace that contains the most important collective motions of the protein, even though the sampling of the available configurational space is still not perfect.

In general, for a given n, RMSIP_2 calculated from the 50-ns trajectories gets higher averages than those derived from the 5-ns trajectories. For example, for $n = 5$, the average of RMSIP_2 is 0.68 for the short trajectories and 0.78 for the long ones. For $n = 20$, the values are 0.82 and 0.90, respectively. However, the benefits of concatenating trajectories become more evident if one considers not just the average values but the whole range of possible outcomes, for an equivalent simulation time. Thus for example, by concatenating

10 trajectories of 5 ns we obtained an average RMSIP_2 of 0.74. Exactly the same average was obtained for single trajectories of 50 ns. But, in the former case, the range of possible outcomes was 0.53-0.96 while in the second one it was 0.09-0.96. A closer inspection to the results presented in Fig. 4 shows that the increase in the average values of RMSIP_2 with n is mainly caused by the increase in the minimum values while the maxima are all similar. Thus, by doing a principal component analysis with concatenated trajectories one can avoid obtaining really ill-defined PC-modes.

It remains to be checked how the reproducibility of the individual PC-modes variates with the number of simulations included in the concatenated trajectory. In the following, we will just present the results obtained with the 5-ns trajectories. The previous discussion demonstrates that it is more difficult to obtain consistent PC-modes in this case than with the longer trajectories. Figure 5 shows typical inner product matrices for the first six PC-modes obtained from $\mathbf{C}^{(n)}$, for $n = 10$, 20, 30 and 90. For $n = 90$ we only have one matrix to show. For $n = 10$, 20 and 30 we have several options and decided to show matrices that present an average behavior. This is, neither the best nor the worst matrix for the given value of n, but an intermediate one. The improvement in the reproducibility of the individual PC-modes can be clearly seen. By increasing n, the elements in the diagonal or their closest neighbors reach significant values, while more distant elements become smaller.

For $n = 90$, the inner product between the two PC_1 is 0.979, while that of the PC_2 is 0.977. Similar almost perfect agreement is found for PC_5 and PC_6. However, something odd seems to happen between the PC_3 and PC_4 since the PC_3 of one batch is mostly contained in PC_4 of the other one and vice versa. This occurs because PC_3 and PC_4 are almost degenerate. The direction of degenerate eigenvectors is arbitrary since any linear combination of such vectors is also an eigenvector with the same eigenvalue. Thus, in cases like this, one cannot do better than determining the subspace spanned by the degenerate vectors. The near-constrained subspace is plenty of almost degenerate eigenvectors. However, this causes no troubles because these vectors are discarded in applications of PCA to molecular dynamics

of proteins. On the contrary, if the degeneracy occurs in vectors with high eigenvalues, all of them have to be included in the essential subspace. It is important to note that the mixing between PC_3 and PC_4 shown in Fig. 5 is fortuitous. We could have missed it if we had grouped the simulations in different batches of 90 elements.

PC modes for the concatenated trajectories of lysozyme

The distributions of $|PC_1^a \cdot PC_1^b|$, $RMSIP_2$ and overlap, computed from individual trajectories of lysozyme follow the same qualitative behavior as those of BPTI (presented in Figs. 1 and 3). They show that PC-modes calculated from a single simulation, either of 5 ns or 50 ns, are poorly defined. To be concise we will not present those distributions here. Instead we will focus on what happens when trajectories are concatenated since that is the main subject of this work. We show in Fig. 6 the evolution of $RMSIP_2$ and overlap as the number of concatenated trajectories, n, is increased. It is readily noted that the same trends observed for BPTI also apply to lysozyme. The average values of $RMSIP_2$ and overlap increase with n. This is mainly caused by the increase of the minimum possible values while the maximum values are all similar and high. For the same n, better results are obtained with trajectories of 50 ns than those of 5 ns. However, what is more relevant here, is the comparison of the results corresponding to the same simulation time. For example, the average $RMSIP_2$ between independent trajectories of 50 ns is 0.64, with a lower bound of 0.13 and an upper bound of 0.93. On the other hand, if one calculates $RMSIP_2$ between trajectories obtained by concatenating 10 simulations of 5 ns, the average value is 0.82 and the boundaries are 0.57 and 0.97. Thus, concatenating trajectories improves the reproducibility of the results and significantly reduces the chances of getting ill-defined PC-modes. As observed in the case of BPTI, for the highest n tried in this work, $RMSIP_2$ is very close to 1.0 while the overlap still noticeably deviates from that. This reinforces the observation that the first PC-modes converge faster than those corresponding to higher indexes. Therefore, well-defined essential spaces for proteins can be computed even though the sampling of the available

configurational space is not fully-converged.

Discussion

The results presented above demonstrate that reproducible PC-modes can be obtained by diagonalizing the correlation matrix of a concatenated trajectory, formed from n equivalent but independent simulations. Agreement of the most important PC-modes, within any desired accuracy, can be reached by systematically increasing n. A particularly convenient characteristic of the procedure is that relatively small values of n (i.e. between 10 and 20) significantly increase the average RMSIP by elevating the minimum possible value. In this way, one can avoid obtaining ill-defined essential spaces. The convergence of the individual PC-modes, on the other hand, requires larger values of n.

The reason why concatenating trajectories improves the reproducibility of the PC-modes was outlined in Section "Statement of the problem" and is based on the formulas presented in Ref. 24. A practical example of the performance of the procedure can be seen in Fig. 7. It shows the projection of typical 5-ns trajectories of BPTI onto the plane spanned by the first two eigenvectors of $\mathbf{C}^{(180)}$. A contour plot of the free energy calculated with the whole set of 5-ns trajectories of BPTI is also shown there. It is observed that single trajectories just occupy a fraction of the available area. They can repeatedly pass through a given region and never visit a nearby accessible zone. The use of different initial velocities takes the trajectories to different regions. Therefore, even though they individually move around a limited zone, the full accessible region is recovered when they are considered altogether.

Many years ago Caves et. al. described the same behavior for simulations of crambin[18]. They concluded that multiple equivalent trajectories are more efficient to provide a fair sampling of the available conformational space than a single long trajectory. Other studies of that time made similar observations and attained to the same conclusion[48–53]. However the

use of multiple trajectories is normally not considered a requirement for obtaining statistically significant results. Therefore it is not routinely used. Very recent investigations are re-installing the subject[54,55]. The present work aims to contribute in the same direction, but the focus is put on the behavior of PC-modes. To the best of our knowledge, this issue has never been systematically studied before.

Finally, it is interesting to discuss how the PC-modes obtained by concatenating short trajectories compare with those calculated by concatenating long trajectories. Unfortunately, there is not a single answer to that question since it depends on several factors such as the time-scale of the simulations being considered or the rigidity of the structure under analysis. Ref. 18 described that trajectories starting from the same structure but differing in the atomic velocities rapidly diverge from the initial point. Then they stabilize and move within a hyperspherical cortex defined by a nearly constant RMSD with respect to the original structure[18]. If that were strictly the case, converged PC-modes computed by concatenating short and long trajectories would be the same. However more recent studies showed that the description of Ref. 18 does not hold on the much longer time-scales affordable nowadays[56]. In general, if during the extra time the trajectories move further away from their initial point, converged PC-modes computed by concatenating short and long trajectories will differ. In the present study, we found that RMSIP_2 for converged PC-modes computed from 5-ns and 50-ns trajectories was 0.595 for BPTI and 0.795 for lyzozyme. This suggests that the description of Ref. 18 fits better the lyzozyme case than the BPTI case. It should also be noted that, since the PC-modes involved in these comparisons are already converged, the values obtained for the RMSIP_2 are insensitive to the total simulation time. To probe this we run extra trajectories of 5-ns, so that we had a set of 400 of such trajectories. Then we concatenated the 400 trajectories and calculated the PC-modes. These PC-modes were compared with those obtained by concatenating 40 trajectories of 50 ns so that, in the two cases, the total simulation time was 2.0 μs. In this case, the value of RMSIP_2 was 0.603 for BPTI and 0.804 for lyzozyme, which are nearly the same as reported above. It could also

happen that, in the extra time, some trajectories move between different wells. This would be rare for stable conformations of proteins if using the simulation times typically employed nowadays. However, the chances of observing such transitions increase as the number of concatenated trajectories gets large.[33] In that case, a more drastic effect on the converged PC-modes should be observed. As stated above, the method described in this article is not meant to address those cases. Nevertheless, it could be used to characterize the fluctuations observed within each well after the conformations sampled in the simulations have been clusterized so that they can be ascribed to each well. In any case, the most important conclusion to be drawn from this discussion is that PC-modes obtained by concatenating n trajectories of a simulation time T_s are representative of the configurational space attainable in time T_s. In general, they are not expected to be the same as PC-modes computed by concatenating longer trajectories.

Conclusions

We have shown that concatenating n independent but equivalent MD simulations, and computing the PC-modes from the correlation matrix of the concatenated trajectory $\mathbf{C}^{(n)}$, significantly improves the reproducibility of the main PC-modes. The procedure has two important and convenient properties. First, small values of n provide a significant enhancement against the results obtained from single MD simulations. In particular, the possibility of getting badly-defined essential spaces is greatly reduced. Second, if desired or needed, the quality of the results can be systematically improved by increasing n. The main limitation of the procedure has also been stated. The PC-modes so obtained are representative of a specific simulation time: the time of the individual trajectories. They are not expected to be the same as PC-modes converged from longer simulations. We believe that the procedure proposed here will be particularly useful in quantitative applications of PCA, such as

calculations of entropy or free energy, as well as for approximate methods that rely in the selection of appropriate coordinates to reduce the dimensionality of the system.

Acknowledgement

The authors thank CONICET (Project ID 11220130100260CO) for the financial support of this research. Financial and computational support from the Universidad Nacional de Quilmes (Project ID 1402/15) is also gratefully acknowledged.

Supporting information

A procedure to assess the consistency of the essential space determined by concatenating a given number of equivalent MD simulations is presented as Supporting information. The proposed method is completely general and easy to implement. It would allow to the users to recognize if the calculations already performed are enough or, on the contrary, some extra calculations are required. This information is available free of charge via the Internet at http://pubs.acs.org .

References

(1) Amadei, A.; Linssen, A.; Berendsen, H. Essential Dynamics of Proteins. *Proteins: Struct., Funct., Bioinf.* **1994**, *17*, 412–425.

(2) de Groot, B. L.; Amadei, A.; Scheek, R. M.; van Nuland, N. A. J.; Berendsen, H. J. C. An Extended Sampling of the Configurational Space of HPr from E. coli. *Proteins: Struct., Funct., Bioinf.* **1996**, *26*, 314–322.

(3) Nymeyer, H.; García, A. E. Simulation of the Folding Equilibrium of α-helical Peptides: a Comparison of the Generalized Born Approximation with Explicit Solvent. *Proc. Natl. Acad. Sci.* **2003**, *100*, 13934–13939.

(4) Zhang, W.; Wu, C.; Duan, Y. Convergence of Replica Exchange Molecular Dynamics. *J. Chem. Phys* **2005**, *123*, 154105.

(5) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14*, 325–332.

(6) Perahia, D.; Levy, R.; Karplus, M. Motions of an α-helical Polypeptide: Comparison of Molecular and Harmonic Dynamics. *Biopolymers* **1990**, *29*, 645–677.

(7) Shenai, P. M.; Xu, Z.; Zhao, Y. Applications of Principal Component Analysis (PCA) in Materials Science. *Principal Component Analysis-Engineering Applications. InTech* **2012**, 25–40.

(8) Yang, J. S.; Hu, X. J.; Li, X. X.; Li, H. Y.; Wang, Y. Application of Principal Component Analysis (PCA) for the Estimation of Source of Heavy Metal Contamination in Sediments of Xihe River, Shenyang City. Advanced Materials Research. 2012; pp 948–951.

(9) Hayward, S.; Kitao, A.; Hirata, F.; Gō, N. Effect of Solvent on Collective Motions in Globular Protein. *J. Mol. Biol.* **1993**, *234*, 1207–1217.

(10) Kitao, A.; Go, N. Investigating Protein Dynamics in Collective Coordinate Space. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169.

(11) Berendsen, H. J.; Hayward, S. Collective Protein Dynamics in Relation to Function. *Curr. Opin. Struct. Biol.* **2000**, *10*, 165–169.

(12) Spoel, D. V. D.; de Groot, B. L.; Hayward, S.; Berendsen, H. J.; Vogel, H. J. Bending of the Calmodulin Central Helix: a Theoretical Study. *Protein Sci.* **1996**, *5*, 2044–2053.

(13) van Aalten, D.; Jones, P.; De Sousa, M.; Findlay, J. Engineering Protein Mechanics: Inhibition of Concerted Motions of the Cellular Retinol Binding Protein by Site-directed Mutagenesis. *Protein Eng.* **1997**, *10*, 31–37.

(14) Vesper, M. D.; de Groot, B. L.; Livesay, D. R. Collective Dynamics Underlying Allosteric Transitions in Hemoglobin. *PLoS Comput. Biol.* **2013**, *9*, e1003232.

(15) Hub, J. S.; de Groot, B. L. Detection of Functional Modes in Protein Dynamics. *PLoS Comput. Biol.* **2009**, *5*, e1000480.

(16) Amadei, A.; Linssen, A. B.; de Groot, B. L.; Berendsen, H. J. Essential Degrees of Freedom of Proteins. **1995**, 85–93.

(17) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. Principal Component Analysis and Long Time Protein Dynamics. *J. Phys. Chem.* **1996**, *100*, 2567–2572.

(18) Caves, L. S.; Evanseck, J. D.; Karplus, M. Locally Accessible Conformations of Proteins: Multiple Molecular Dynamics Simulations of Crambin. *Protein Sci.* **1998**, *7*, 649–666.

(19) Hess, B. Similarities between Principal Components of Protein Dynamics and Random Diffusion. *Phys. Rev. E* **2000**, *62*, 8438.

(20) Amadei, A.; Ceruso, M. A.; Di Nola, A. On the Convergence of the Conformational Coordinates Basis Set Obtained by the Essential Dynamics Analysis of Proteins' Molecular Dynamics Simulations. *Proteins: Struct., Funct., Bioinf.* **1999**, *36*, 419–424.

(21) Lambrughi, M.; Papaleo, E.; Testa, L.; Brocca, S.; De Gioia, L.; Grandori, R. Intramolecular Interactions Stabilizing Compact Conformations of the Intrinsically Disordered Kinase-inhibitor Domain of Sic1: a Molecular Dynamics Investigation. *Front Physiol.* **2012**, *3*, 435.

(22) Martín-García, F.; Papaleo, E.; Gomez-Puertas, P.; Boomsma, W.; Lindorff-Larsen, K. Comparing Molecular Dynamics Force Fields in the Essential Subspace. *PloS One* **2015**, *10*, e0121114.

(23) Jónsdóttir, L. B.; Ellertsson, B. Ö.; Invernizzi, G.; Magnúsdóttir, M.; Thorbjarnardóttir, S. H.; Papaleo, E.; Kristjánsson, M. M. The Role of Salt Bridges on the Temperature Adaptation of Aqualysin I, a Thermostable Subtilisin-like Proteinase. *Biochim. Biophys. Acta* **2014**, *1844*, 2174–2181.

(24) Pierdominici-Sottile, G.; Palma, J. New Insights into the Meaning and Usefulness of Principal Component Analysis of Concatenated Trajectories. *J. Comput. Chem* **2015**, *36*, 424–432.

(25) Lou, H.; ; Cukier, R. I. Molecular Dynamics of Apo-Adenylate Kinase: A Distance Replica Exchange Method for the Free Energy of Conformational Fluctuations. *J. Phys. Chem. B* **2006**, *110*, 24121–24137.

(26) Daidone, I.; Amadei, A. Essential Dynamics: Foundation and Applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 762–770.

(27) de Groot, B.; van Aalten, D.; Amadei, A.; Berendsen, H. The Consistency of Large Concerted Motions in Proteins in Molecular Dynamics Simulations. *Biophys. J* **1996**, *71*, 1707.

(28) de Groot, B.; Hayward, S.; Van Aalten, D.; Amadei, A.; Berendsen, H. Domain Motions in Bacteriophage T4 Lysozyme: a Comparison between Molecular Dynamics and Crystallographic Data. *Proteins: Struct., Funct., Genet.* **1998**, *31*, 116–127.

(29) Amadei, A.; Ceruso, M. A.; Di Nola, A. On the Convergence of the Conformational Coordinates Basis Set Obtained by the Essential Dynamics Analysis of Proteins' Molecular Dynamics Simulations. *Proteins: Struct., Funct., Bioinf.* **1999**, *36*, 419–424.

(30) Skjaerven, L.; Martínez, A.; Reuter, N. Principal Component and Normal Mode Analysis of Proteins; a Quantitative Comparison Using the GroEL Subunit. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 232–243.

(31) Grossfield, A.; Feller, S. E.; Pitman, M. C. Convergence of Molecular Dynamics Simulations of Membrane Proteins. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 31–40.

(32) van Aalten, D. M. F.; Amadei, A.; Linssen, A. B. M.; Eijsink, V. G. H.; Vriend, G.; Berendsen, H. J. C. The Essential Dynamics of Thermolysin: Confirmation of the Hinge-bending Motion and Comparison of Simulations in Vacuum and Water. *Proteins: Struct., Funct., Bioinf.* **1995**, *22*, 45–54.

(33) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. Atomistic Protein Folding Simulations on the Submillisecond Time Scale Using Worldwide Distributed Computing. *Biopolymers* **2003**, *68*, 91–109.

(34) Ikebe, J.; Umezawa, K.; Kamiya, N.; Sugihara, T.; Yonezawa, Y.; Takano, Y.; Nakamura, H.; Higo, J. Theory for Trivial Trajectory Parallelization of Multicanonical Molecular Dynamics and Application to a Polypeptide in Water. *J. Comput. Chem.* **2011**, *32*, 1286–1297.

(35) Allen, L. R.; Krivov, S. V.; Paci, E. Analysis of the Free-Energy Surface of Proteins from Reversible Folding Simulations. *PLoS Comput. Biol.* **2009**, *5*, e1000428.

(36) Hori, N.; Chikenji, G.; Berry, R. S.; Takada, S. Folding Energy Landscape and Network Dynamics of Small Globular Proteins. *Proc. Natl. Acad. Sci.* **2009**, *106*, 73–78.

(37) Sittel, F.; Jain, A.; Stock, G. Principal Component Analysis of Molecular Dynamics: On the Use of Cartesian vs. Internal Coordinates. *J. Chem. Phys.* **2014**, *141*.

(38) Faber, N.; Meinders, M.; Geladi, P.; Sjöström, M.; Buydens, L.; Kateman, G. Random Error Bias in Principal Component Analysis. Part I. Derivation of Theoretical Predictions. *Anal. Chim. Acta.* **1995**, *304*, 257–271.

(39) Nadler, B. Finite Sample Approximation Results for Principal Component Analysis: A Matrix Perturbation Approach. *Ann. Stat* **2008**, 2791–2817.

(40) Allez, R.; Bouchaud, J.-P. Eigenvector Dynamics: General Theory and some Applications. *Phys. Rev. E* **2012**, *86*, 046202.

(41) Wlodawer, A.; Walter, J.; Huber, R.; Sjölin, L. Structure of Bovine Pancreatic Trypsin Inhibitor: Results of Joint Neutron and X-ray Refinement of Crystal form II. *J. Mol. Biol.* **1984**, *180*, 301–329.

(42) Muraki, M.; Harata, K.; Sugita, N.; Sato, K. Origin of Carbohydrate Recognition Specificity of Human Lysozyme Revealed by Affinity Labeling,. *Biochemistry* **1996**, *35*, 13562–13567.

(43) Hess, B. Convergence of Sampling in Protein Simulations. *Phys. Rev. E* **2002**, *65*, 031910.

(44) Boechi, L.; de Oliveira, C. A. F.; Da Fonseca, I.; Kizjakina, K.; Sobrado, P.; Tanner, J. J.; McCammon, J. A. Substrate-Dependent Dynamics of UDP-galactopyranose Mutase: Implications for Drug Design. *Protein Sci.* **2013**, *22*, 1490–1501.

(45) Lange, O. F.; Grubmüller, H. Full Correlation Analysis of Conformational Protein Dynamics. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1294–1312.

(46) Pierce, L. C.; Salomon-Ferrer, R.; de Oliveira, C. A. F.; McCammon, J. A.;

Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 2997–3002.

(47) Lange, O. F.; Grubmüller, H. Generalized Correlation for Biomolecular Dynamics. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 1053–1061.

(48) Elofsson, A.; Nilsson, L. How Consistent are Molecular Dynamics Simulations? *J. Mol. Biol.* **1993**, *233*, 766 – 780.

(49) Schulze, B. G.; Evanseck, J. D. Cooperative Role of Arg45 and His64 in the Spectroscopic A3 state of Carbonmonoxy Myoglobin: Molecular Dynamics Simulations, Multivariate Analysis, and Quantum Mechanical Computations. *J. Am. Chem. Soc.* **1999**, *121*, 6444–6454.

(50) Schiøtt, B.; Bruice, T. C. Reaction Mechanism of Soluble Epoxide Hydrolase: Insights from Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2002**, *124*, 14558–14570.

(51) Vitkup, D.; Ringe, D.; Karplus, M.; Petsko, G. A. Why Protein R-factors are so Large: A Self-consistent Analysis. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 345–354.

(52) Worth, G. A.; Nardi, F.; Wade, R. C. Use of Multiple Molecular Dynamics Trajectories to Study Biomolecules in Solution: the YTGP Peptide. *J. Phys. Chem. B* **1998**, *102*, 6260–6272.

(53) Gorfe, A. A.; Ferrara, P.; Caflisch, A.; Marti, D. N.; Bosshard, H. R.; Jelesarov, I. Calculation of Protein Ionization Equilibria with Conformational Sampling: pKa of a Model Leucine Zipper, GCN4 and Barnase. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 41–60.

(54) Perez, J. J.; Tomas, M. S.; Rubio-Martínez, J. Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory. *J. Chem. Inf. Model.* **2016**, *56*, 1950–1962.

(55) Coveney, P. V.; Wan, S. On the Calculation of Equilibrium Thermodynamic Properties from Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30236–30240.

(56) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341–346.

Figures

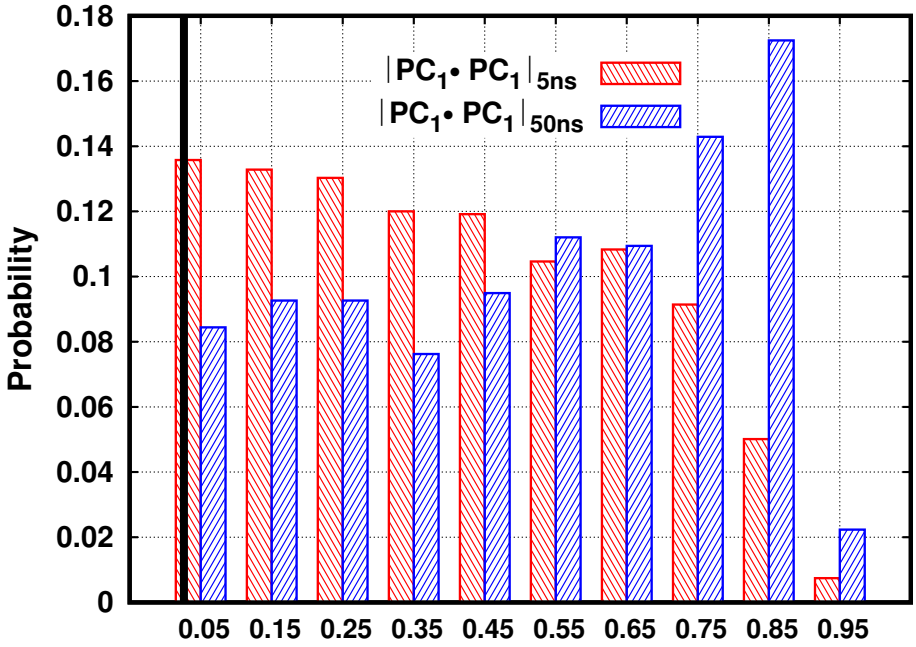

Figure 1: Normalized probability distributions for $|\mathrm{PC}_1^a \cdot \mathrm{PC}_1^b|$, computed from independent MD simulations of 5 ns (red) and 50 ns (blue) of BPTI. The vertical black line indicates the value of the inner product that contains 99% of cumulative probability, for normalized random vectors of the same dimension (see text).

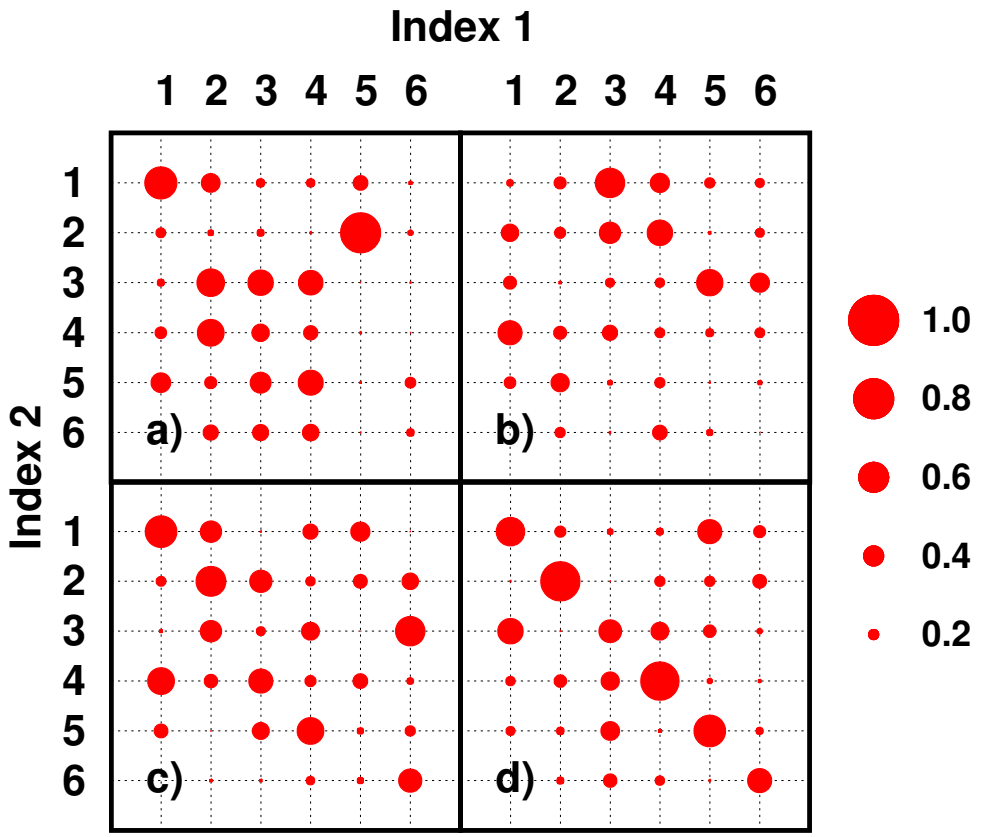

Figure 2: Typical examples of inner-product matrices for PC-modes computed from two independent MD simulations of BPTI. (a) and (b) correspond to 5-ns trajectories; (c) and (d) to 50-ns trajectories. The label of the axis refer to the index of the PC-modes while the radius of the circles measures the absolute value of the inner products.

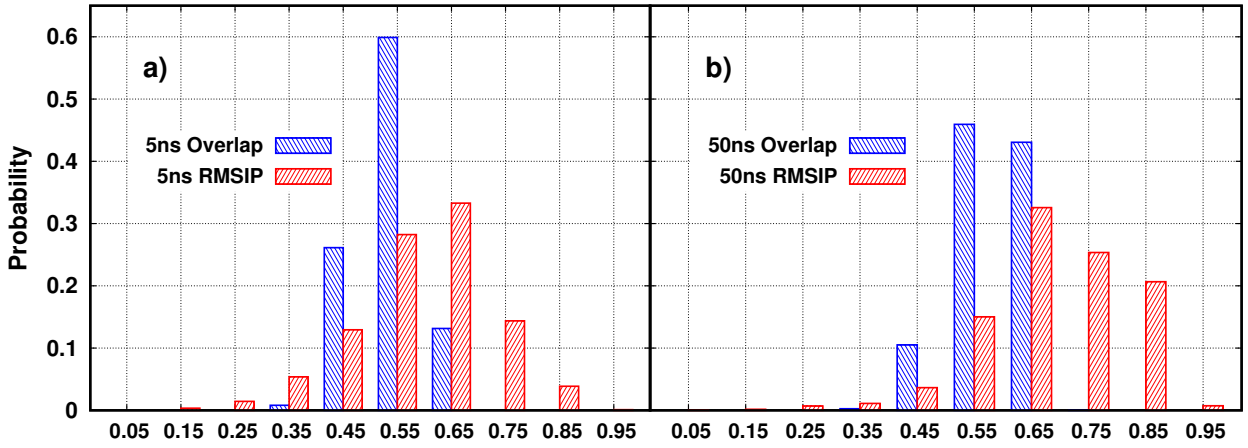

Figure 3: (a) Normalized probability distributions for $RMSIP_2$ and overlap computed from the 16110 pairs of PC-mode sets formed from the 180 individual MD simulations of 5 ns of BPTI. (b) Normalized probability distributions for $RMSIP_2$ and overlap computed from the 3160 pairs of PC-mode sets formed from the 80 individual MD simulations of 50 ns of BPTI.

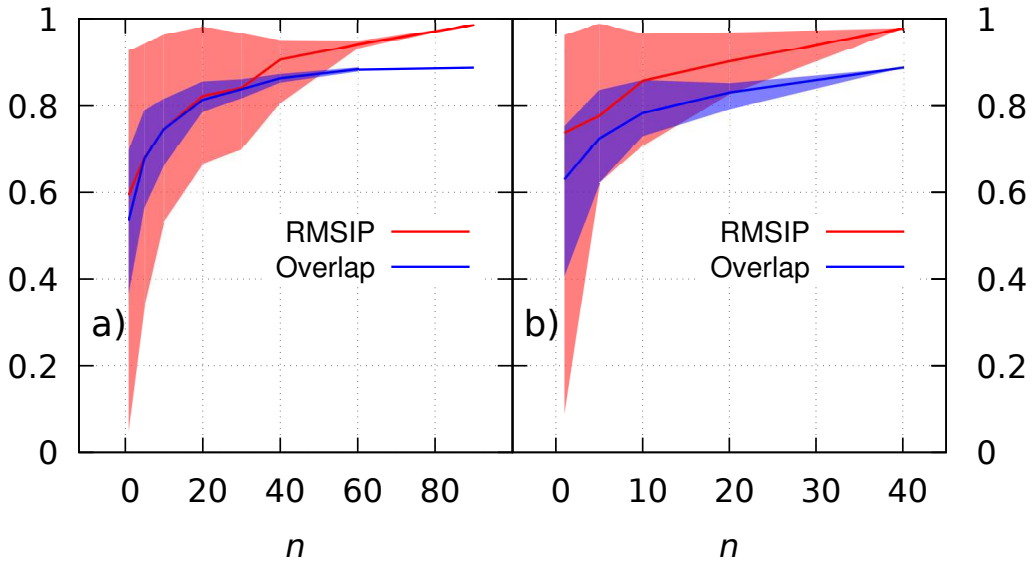

Figure 4: Evolution of $RMSIP_2$ and overlap with the number of trajectories, n, included in the concatenated correlation matrix. Data correspond to simulations of BPTI. The solid lines indicate the average values. The shadows go from the minimum to the maximum value observed in the sample. (a) 5-ns trajectories; (b) 50-ns trajectories.

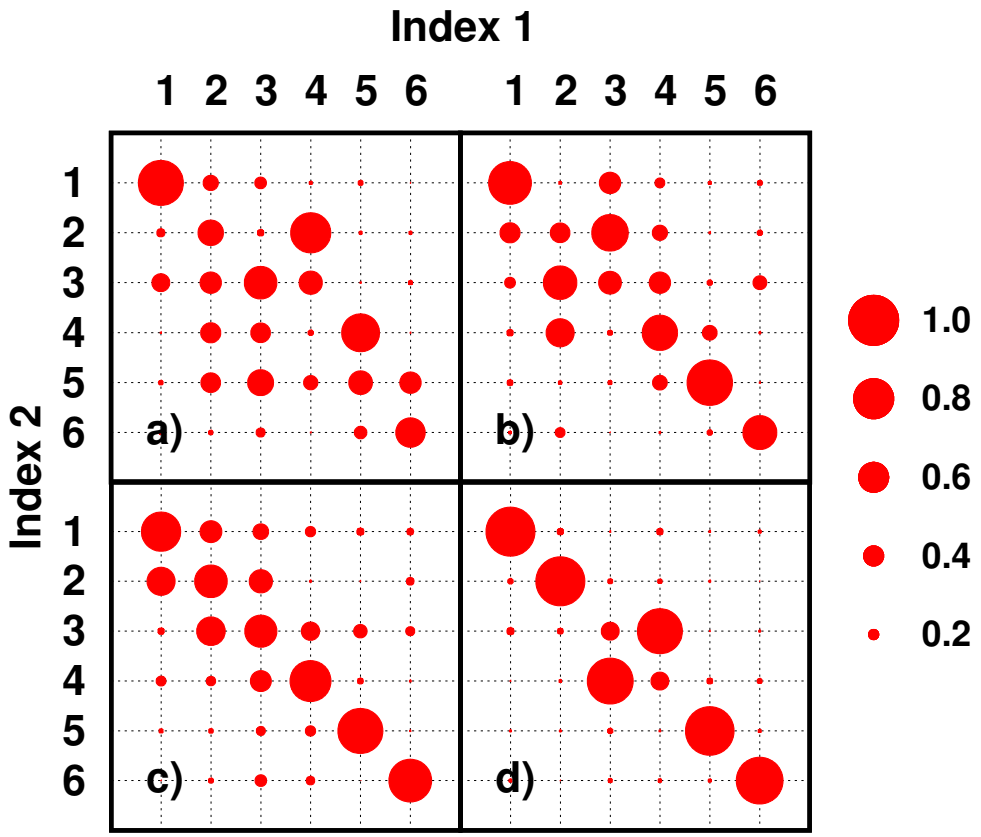

Figure 5: Inner product matrices for PC-modes computed from alternative batches of concatenated trajectories of BPTI. The label of the axis refer to the index of the PC-modes while the radius of the circles measures the absolute value of the inner products. a) $n = 10$; b) $n = 20$; c) $n = 30$; d) $n = 90$.

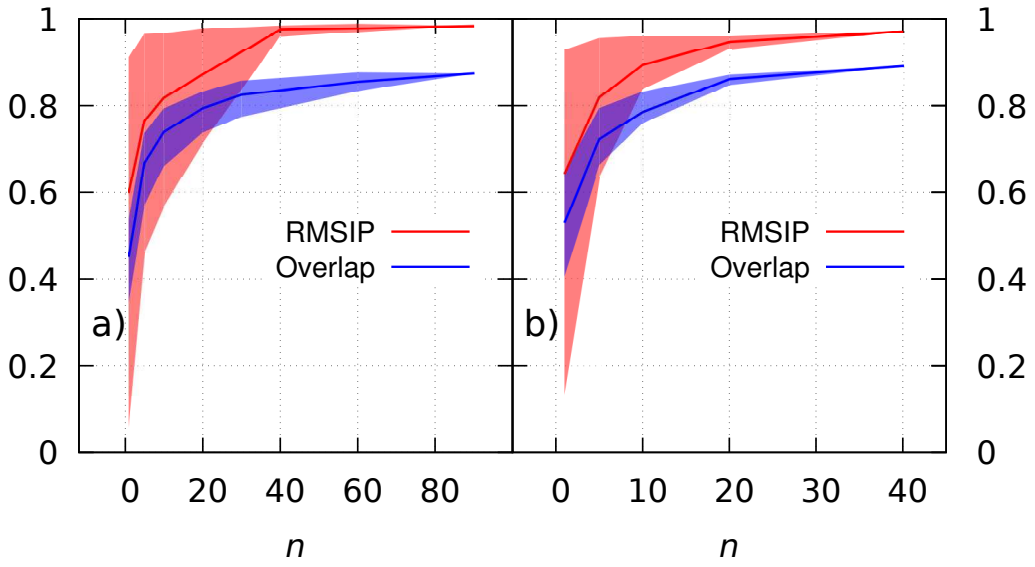

Figure 6: Evolution of $RMSIP_2$ and overlap with the number of trajectories, n, included in the concatenated correlation matrix. Data correspond to simulations of lysozyme. The solid lines indicate the average values. The shadows go from the minimum to the maximum value observed in the sample. (a) 5-ns trajectories; (b) 50-ns trajectories.

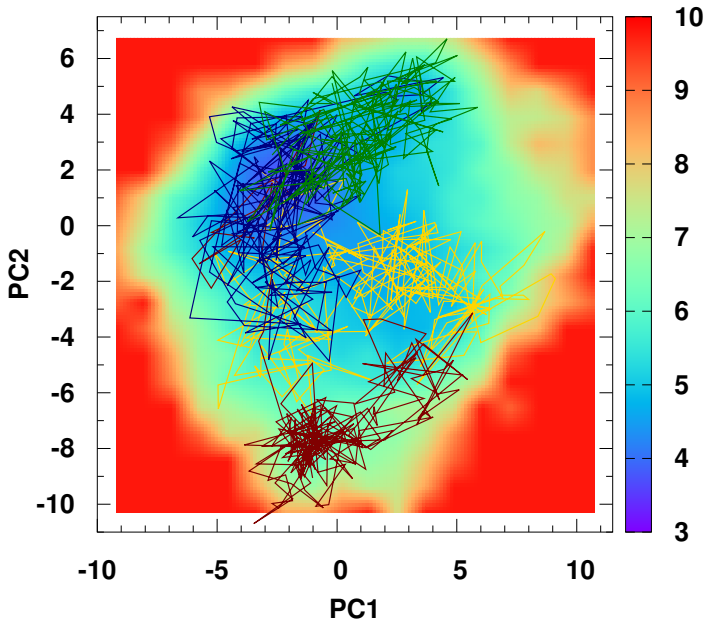

Figure 7: Typical individual trajectories projected onto the plane spanned by PC_1 and PC_2 of matrix $\mathbf{C}^{(180)}$ of BPTI. The colored contour plot shows the free energy (in arbitrary units) computed from the snapshots collected along the 180 independent trajectories.

Tables

Table 1: Number of batches (N_{batch}) and number of pairs of batches (N_{pairs}) that can be formed from the individual trajectories for each selected value of n.

180 traj (5 ns)			80 traj (50 ns)		
n	N_{batch}	N_{pairs}	n	N_{batch}	N_{pairs}
1	180	16110	1	80	3160
5	36	630	5	16	120
10	18	153	10	8	28
20	9	36	20	4	6
30	6	15	40	2	1
60	3	3			
90	2	1			

Graphical TOC Entry