New Insights into the Meaning and Usefulness of Principal Component Analysis of Concatenated Trajectories

Gustavo Pierdominici-Sottile and Juliana Palma*

A comparison between different conformations of a given protein, relating both structure and dynamics, can be performed in terms of combined principal component analysis (combined-PCA). To that end, a trajectory is obtained by concatenating molecular dynamics trajectories of the individual conformations under comparison. Then, the principal components are calculated by diagonalizing the correlation matrix of the concatenated trajectory. Since the introduction of this approach in 1995 it has had a large number of applications. However, the interpretation of the eigenvectors and eigenvalues so obtained is based on intuitive foundations, because analytical expressions relating the concatenated correlation matrix with those of the individual trajectories under consideration have not been provided yet. In this article, we present such expressions for the cases of two, three, and an arbitrary number of concatenated trajectories. The formulas are simple and show what is to be expected and what is not to be expected from a combined-PCA. Their correctness and usefulness is demonstrated by discussing some representative examples. The results can be summarized in a simple sentence: the correlation matrix of a concatenated trajectory is given by the average of the individual correlation matrices plus the correlation matrix of the individual averages. From this it follows that the combined-PCA of trajectories belonging to different free energy basins provides information that could also be obtained by alternative and more straightforward means. © 2014 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23811

Introduction

Principal component analysis (PCA) has been widely used to characterize the dynamics of proteins since it allows to detect important directions in their multidimensional configurational space.[1] These directions are obtained from molecular dynamics (MD) simulations by diagonalizing the correlation matrix.[2]Usually, a few eigenvectors stand out for having eigenvalues far larger than the rest. Movements along these directions account for the largest structural variations of the peptidic chain, describing the so-called essential dynamics (ED) of the protein. Motions along the remaining eigenvectors just correspond to trivial, nearly Gaussian fluctuations. There have been many discussions on the reliability, usefulness, and meaning of the vectors identified by PCA.[3–7] Besides, several tools have been provided to assess their stability and convergence.[1] The main hypothesis of the approach is that the ED of a protein, determined with PCA, contains the motions relevant to its function.[2] This hypothesis has gained support from the build-up of MD studies that describe a close relationship between the first eigenvectors of the correlation matrix and the functional motions of several proteins.[8–14] The PCA method is closely related to quasiharmonic analysis, a method that provides an affordable approach to compute configurational entropies.[15–17] Finally, we should note that PCA only identifies linear correlations between atomic fluctuations. More sophisticated procedures have to be used to detect correlations beyond linearity.[18,19]

An extension of the PCA method consists of diagonalizing the correlation matrix obtained by concatenating two or more independent trajectories, each corresponding to an alternative conformation of the same protein.[20] They could be, for example, the trajectories for the holo and apo forms of a protein, the active and inactive forms of an enzyme, the open and closed structures of a channel protein or distinct oligomerization states of a given protein subunit. It is known that the main eigenvectors of these combined correlation matrices (CCM) no longer describe the largest deformations of the conformations involved. Instead, it has been asserted that they highlight differences in the structure and dynamics of the proteins under comparison. The occurrence of static modes among the eigenvectors of the CCM has frequently been reported (see for example Refs. [20–24]). They are identified as eigenvectors of the CCM for which the projections of the individual trajectories differ significantly.

To the best of our knowledge, an analytical expression relating the CCM to the structures and correlation matrices of the individual trajectories involved has not been provided yet. In this article, we present such formulas for the cases of two, three, and an arbitrary number of concatenated trajectories. We believe that they will be useful to enlighten the interpretation of the results of combined-ED analysis and to guide its discussion. Among other things, these expressions allow to predict the number of static modes to be expected and afford a precise and clear meaning for the eigenvalues and directions of these eigenvectors.

G. Pierdominici-Sottile, J. Palma
Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Sáenz Peña 352, Bernal, B1876BXD, Argentina
E-mail: juliana@unq.edu.ar

Contract grant sponsor: CONICET and Universidad Nacional de Quilmes

Methodology

In the following, we denote by N the number of structures sampled from each single trajectory, and we assume that this number is the same for all the trajectories included in the concatenated one. This is the usual procedure in combined-PCA studies. Thus, a concatenated trajectory involving n single trajectories has nN structures. The vector containing the atomic Cartesian coordinates of the kth structure will be denoted by $\mathbf{x}^{(k)}$ and its ith element by $x_i^{(k)}$. Therefore, index k runs from 1 to nN while index i runs from 1 to $3N_{at}$, where N_{at} is the number of atoms considered.

It should be noted that, when the individual trajectories remain within their own free energy wells during the whole simulation, the construction of a combined correlation matrix that uses the same N for all the trajectories leads to a wrong weighting of the different conformations. Therefore, the combined-PCA is deprived of any thermodynamic meaning. However, even though it is the normal practice, there is no need to use the same N for all the trajectories. Accordingly, for completeness, we derive at the appendix the formula for the correlation matrix of two concatenated trajectories for the case in which the numbers of snapshots sampled from each trajectory differ.

Formulas for two concatenated trajectories

Let us A and B denote two different stable conformations of the same protein. After running MD simulations for A and B, and sampling N structures from each trajectory, a new set is formed by arranging the N structures coming from the simulation of A in the first place, followed by the N structures coming from the simulation of B. In this way, the new set $\{\mathbf{x}\}$, corresponding to the concatenated trajectory, contains $2N$ structures:

$$\{\mathbf{x}\}=\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)},\mathbf{x}^{(N+1)},\ldots,\mathbf{x}^{(2N)}\}. \tag{1}$$

According with the usual PCA procedure, the $2N$ structures are aligned to the same reference to eliminate overall rotations and translations. After that, the correlation matrix for the concatenated trajectory, $\mathbf{C}^{AB}$, is calculated following the standard protocols. The ijth element of $\mathbf{C}^{AB}$ is given by

$$C_{ij}^{AB}=\frac{1}{2N}\sum_{k=1}^{2N}\left(x_i^{(k)}-\langle x_i\rangle^{AB}\right)\left(x_j^{(k)}-\langle x_j\rangle^{AB}\right). \tag{2}$$

Here, $\langle x_i\rangle^{AB}$ and $\langle x_j\rangle^{AB}$ are the average values for coordinates x_i and x_j, respectively, evaluated with the structures of the concatenated trajectory. Thus, for example,

$$\langle x_i\rangle^{AB}=\frac{1}{2N}\sum_{k=1}^{2N}x_i^{(k)}. \tag{3}$$

Equation (2) can be rearranged by splitting the summation into two contributions, one involving the structures coming from A and one with the structures coming from B,

$$\begin{aligned}2NC_{ij}^{AB}=&\sum_{k=1}^{N}\left(x_i^{(k)}-\langle x_i\rangle^{AB}\right)\left(x_j^{(k)}-\langle x_j\rangle^{AB}\right)\\&+\sum_{k=N+1}^{2N}\left(x_i^{(k)}-\langle x_i\rangle^{AB}\right)\left(x_j^{(k)}-\langle x_j\rangle^{AB}\right).\end{aligned} \tag{4}$$

Formally, both summations are equivalent as they just differ in the structures under consideration. Thus, we only analyze one of them. Let us Σ^A be the first summation on the right side of eq. (4). It can be rewritten as

$$\Sigma^A=\sum_{k=1}^{N}\left(x_i^{(k)}-\langle x_i\rangle^{A}+\Delta\langle x_i\rangle^{A,AB}\right)\left(x_j^{(k)}-\langle x_j\rangle^{A}+\Delta\langle x_j\rangle^{A,AB}\right), \tag{5}$$

where $\langle x_i\rangle^A$ and $\langle x_j\rangle^A$ are the averages for coordinates x_i and x_j, respectively, evaluated with the N structures taken from the A-simulation, while $\Delta\langle x_i\rangle^{A,AB}$ and $\Delta\langle x_j\rangle^{A,AB}$ express differences between the average values calculated with the N structures coming from the A-trajectory and the $2N$ structures of the concatenated trajectory,

$$\Delta\langle x_i\rangle^{A,AB}=\langle x_i\rangle^{A}-\langle x_i\rangle^{AB},\ \Delta\langle x_j\rangle^{A,AB}=\langle x_j\rangle^{A}-\langle x_j\rangle^{AB}. \tag{6}$$

By partially distributing the right side of eq. (5), we obtain,

$$\begin{aligned}\Sigma^A=&\sum_{k=1}^{N}\left(x_i^{(k)}-\langle x_i\rangle^{A}\right)\left(x_j^{(k)}-\langle x_j\rangle^{A}\right)\\&+\Delta\langle x_j\rangle^{A,AB}\sum_{k=1}^{N}\left(x_i^{(k)}-\langle x_i\rangle^{A}\right)+\Delta\langle x_i\rangle^{A,AB}\sum_{k=1}^{N}\left(x_j^{(k)}-\langle x_j\rangle^{A}\right)\\&+N\Delta\langle x_i\rangle^{A,AB}\Delta\langle x_j\rangle^{A,AB}.\end{aligned} \tag{7}$$

In this expression, the first summation equals to NC_{ij}^A, where C_{ij}^A is the ijth element of the correlation matrix corresponding to the A-simulation. Conversely, the second and third terms vanish because of the very definitions of $\langle x_i\rangle^A$ and $\langle x_j\rangle^A$. Therefore, the final expression for Σ^A is,

$$\Sigma^A=NC_{ij}^A+N\Delta\langle x_i\rangle^{A,AB}\Delta\langle x_j\rangle^{A,AB}, \tag{8}$$

with an equivalent equation bearing for Σ^B, the second summation on the right side of eq. (4). Replacing the expressions for Σ^A and Σ^B in eq. (4) and rearranging we obtain,

$$2C_{ij}^{AB}=C_{ij}^A+C_{ij}^B+\Delta\langle x_i\rangle^{A,AB}\Delta\langle x_j\rangle^{A,AB}+\Delta\langle x_i\rangle^{B,AB}\Delta\langle x_j\rangle^{B,AB}. \tag{9}$$

Equation (9) indicates that both, structural and dynamical features, contribute to the correlation matrix for two concatenated trajectories. The dynamical contribution is contained in the correlation matrices of the individual trajectories. The structural or static contribution comes from the differences between the average structures. We collect the static contributions in matrix $\mathbf{S}^{AB}$,

$$S_{ij}^{AB}=\frac{\Delta\langle x_i\rangle^{A,AB}\Delta\langle x_j\rangle^{A,AB}+\Delta\langle x_i\rangle^{B,AB}\Delta\langle x_j\rangle^{B,AB}}{2}. \tag{10}$$

Thus, the final expression for C_{ij}^{AB} is,

$$2C_{ij}^{AB}=C_{ij}^{A}+C_{ij}^{B}+2S_{ij}^{AB}. \quad (11)$$

A simpler and more explanatory expression for S_{ij}^{AB} can be obtained by noting that the average structure of the concatenated trajectory is a combination of the individual averages,

$$2\langle x_i\rangle^{AB}=\langle x_i\rangle^{A}+\langle x_i\rangle^{B}, 2\langle x_j\rangle^{AB}=\langle x_j\rangle^{A}+\langle x_j\rangle^{B}. \quad (12)$$

Introducing these relationships in eq. (10), we obtain,

$$S_{ij}^{AB}=\frac{\Delta\langle x_i\rangle^{A,B}\Delta\langle x_j\rangle^{A,B}}{4}, \quad (13)$$

where,

$$\Delta\langle x_i\rangle^{A,B}=\langle x_i\rangle^{A}-\langle x_i\rangle^{B}, \Delta\langle x_j\rangle^{A,B}=\langle x_j\rangle^{A}-\langle x_j\rangle^{B}. \quad (14)$$

Equation (13) shows that the elements of matrix $\mathbf{S}^{AB}$ will be negligible for those rows and columns corresponding to Cartesian coordinates that do not change significantly, on average, in going from *A* to *B*. On the contrary, those pairs of coordinates that change their averages in a considerable amount will appear as large positive or negative peaks.

Formulas for n-concatenated trajectories

In this section, we derive the expression for the correlation matrix obtained by concatenating *n* individual trajectories, $\mathbf{C}^n$, assuming that we know the formula for the correlation matrix corresponding to $(n-1)$ concatenated trajectories, $\mathbf{C}^{(n-1)}$. As we already have the expression for two concatenated trajectories, the formula so obtained allows us to go from two concatenated trajectories to three, from three to four and so on. The structures used in the calculation of $\mathbf{C}^n$ are arranged in the set,

$$\{\mathbf{x}\}=\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(nN-N)},\mathbf{x}^{(nN-N+1)},\ldots,\mathbf{x}^{(nN)}\}, \quad (15)$$

where the first $(n-1)N$ structures are used in the computation of C^{n-1}, while the last *N* structures belong to the trajectory being added. We call this last trajectory as the *Z*-trajectory and its correlation matrix as $\mathbf{C}^Z$. The *ij*th element of $\mathbf{C}^n$ is given by

$$C_{ij}^{n}=\frac{1}{nN}\sum_{k=1}^{nN}\left(x_i^{(k)}-\langle x_i\rangle^{n}\right)\left(x_j^{(k)}-\langle x_j\rangle^{n}\right), \quad (16)$$

where $\langle x_i\rangle^n$ and $\langle x_j\rangle^n$ are the average values for coordinates x_i and x_j, respectively, calculated with the *nN* structures of set 15. The summation of eq. (16) can be split into two summations, one involving the first $(n-1)N$ structures of set 15 and another one with the remaining *N* structures coming from the *Z*-trajectory,

$$\begin{aligned}nNC_{ij}^{n}=&\sum_{k=1}^{nN-N}\left(x_i^{(k)}-\langle x_i\rangle^{n}\right)\left(x_j^{(k)}-\langle x_j\rangle^{n}\right)\\&+\sum_{k=nN-N+1}^{nN}\left(x_i^{(k)}-\langle x_i\rangle^{n}\right)\left(x_j^{(k)}-\langle x_j\rangle^{n}\right).\end{aligned} \quad (17)$$

Let us call Σ^{n-1} to the first summation on the right side of eq. (17) and Σ^Z to the second one. We manipulate Σ^{n-1} by adding $\langle x_i\rangle^{n-1}-\langle x_i\rangle^{n-1}$ and $\langle x_j\rangle^{n-1}-\langle x_j\rangle^{n-1}$ within the corresponding parenthesis, where $\langle x_i\rangle^{(n-1)}$ and $\langle x_j\rangle^{(n-1)}$ are the average values for coordinates x_i and x_j, respectively, calculated with the first $(n-1)N$ structures of set 15. After this, Σ^{n-1} reads,

$$\begin{aligned}\Sigma^{n-1}=&\sum_{k=1}^{nN-N}\left(x_i^{(k)}-\langle x_i\rangle^{n-1}\right)\left(x_j^{(k)}-\langle x_j\rangle^{n-1}\right)\\&+\Delta\langle x_j\rangle^{n-1,n}\sum_{k=1}^{nN-N}\left(x_i^{(k)}-\langle x_i\rangle^{n-1}\right)+\Delta\langle x_i\rangle^{n-1,n}\sum_{k=1}^{nN-N}\left(x_j^{(k)}-\langle x_j\rangle^{n-1}\right)\\&+(n-1)N\Delta\langle x_i\rangle^{n-1,n}\Delta\langle x_j\rangle^{n-1,n},\end{aligned} \quad (18)$$

where

$$\Delta\langle x_i\rangle^{n-1,n}=\langle x_i\rangle^{n-1}-\langle x_i\rangle^{n}, \Delta\langle x_j\rangle^{n-1,n}=\langle x_j\rangle^{n-1}-\langle x_j\rangle^{n}. \quad (19)$$

The first term of eq. (18) equals to $(n-1)NC_{ij}^{n-1}$ while the second and third term vanish. Therefore, we obtain,

$$\Sigma^{n-1}=(n-1)NC_{ij}^{n-1}+(n-1)N\Delta\langle x_i\rangle^{n-1,n}\Delta\langle x_j\rangle^{n-1,n}. \quad (20)$$

A similar reasoning shows that Σ^Z is given by,

$$\Sigma^{Z}=NC_{ij}^{Z}+N\Delta\langle x_i\rangle^{Z,n}\Delta\langle x_j\rangle^{Z,n}, \quad (21)$$

where,

$$\Delta\langle x_i\rangle^{Z,n}=\langle x_i\rangle^{Z}-\langle x_i\rangle^{n}, \Delta\langle x_j\rangle^{Z,n}=\langle x_j\rangle^{Z}-\langle x_j\rangle^{n}, \quad (22)$$

while $\langle x_i\rangle^Z$ and $\langle x_j\rangle^Z$ are the average values of coordinates x_i and x_j, respectively, computed with the last *N* structures of set 15. Replacing the expressions obtained for Σ^{n-1} and Σ^Z in eq. (17) and rearranging we obtain,

$$nC_{ij}^{n}=(n-1)C_{ij}^{n-1}+C_{ij}^{Z}+(n-1)\Delta\langle x_i\rangle^{n-1,n}\Delta\langle x_j\rangle^{n-1,n}+\Delta\langle x_i\rangle^{Z,n}\Delta\langle x_j\rangle^{Z,n}. \quad (23)$$

This equation looks similar to eq. (9) except for the fact that the contributions coming from the $(n-1)$ concatenated trajectories have a weight that is $(n-1)$ times larger than that of the *Z*-trajectory. For $n=2$, eq. (23) becomes equivalent to eq. (9).

Formulas for 3-concatenated trajectories

Here, we use eq. (23) to obtain the correlation matrix for three concatenated trajectories, $\mathbf{C}^3$, from the correlation matrix of two concatenated trajectories, $\mathbf{C}^2$, and the matrix corresponding to the trajectory being added, $\mathbf{C}^Z$. It should be noted that names $\mathbf{C}^2$ and C_{ij}^2 used in this section are equivalent to $\mathbf{C}^{AB}$ and C_{ij}^{AB} of the previous section, assuming that the two trajectories considered in the computation of $\mathbf{C}^2$ are called *A* and *B*. According with eq. (23) we have,

$$3C_{ij}^{3}=2C_{ij}^{2}+C_{ij}^{Z}+2\Delta\langle x_i\rangle^{2,3}\Delta\langle x_j\rangle^{2,3}+\Delta\langle x_i\rangle^{Z,3}\Delta\langle x_j\rangle^{Z,3}. \quad (24)$$

After replacing C_{ij}^2 for the expression given in eq. (9) we obtain,

$$3C_{ij}^3 = C_{ij}^A + C_{ij}^B + C_{ij}^Z + 2\Delta\langle x_i\rangle^{2,3}\Delta\langle x_j\rangle^{2,3} + \Delta\langle x_i\rangle^{Z,3}\Delta\langle x_j\rangle^{Z,3} + \Delta\langle x_i\rangle^{A,2}\Delta\langle x_j\rangle^{A,2} + \Delta\langle x_i\rangle^{B,2}\Delta\langle x_j\rangle^{B,2}. \quad (25)$$

Even correct, this expression is not appropriate because it does not treat the three trajectories on the same foot. To overcome this drawback, we note that,

$$\Delta\langle x_i\rangle^{2,3} = \Delta\langle x_i\rangle^{A,3} - \Delta\langle x_i\rangle^{A,2} = \Delta\langle x_i\rangle^{B,3} - \Delta\langle x_i\rangle^{B,2},$$
$$\langle x_i\rangle^3 = \left(\langle x_i\rangle^A + \langle x_i\rangle^B + \langle x_i\rangle^Z\right)/3, \quad (26)$$
$$\langle x_i\rangle^2 = \left(\langle x_i\rangle^A + \langle x_i\rangle^B\right)/2,$$

with similar equations holding for coordinate x_j. Introducing these relationships into eq. (25) and rearranging, a simple expression for C_{ij}^3 that treats the three trajectories on the same foot can be obtained. It reads,

$$3C_{ij}^3 = C_{ij}^A + C_{ij}^B + C_{ij}^Z + \Delta\langle x_i\rangle^{A,3}\Delta\langle x_j\rangle^{A,3} + \Delta\langle x_i\rangle^{B,3}\Delta\langle x_j\rangle^{B,3} + \Delta\langle x_i\rangle^{Z,3}\Delta\langle x_j\rangle^{Z,3}. \quad (27)$$

This equation is qualitatively similar to eq. (9). It shows that the dynamical contributions to $\mathbf{C}^3$ are embodied in the correlation matrices of the individual trajectories while the static contributions come from differences between the average structure of the concatenated trajectory and those of the individual trajectories. As before, we collect the static contribution in a single matrix that we denote as $\mathbf{S}^3$,

$$S_{ij}^3 = \frac{\Delta\langle x_i\rangle^{A,3}\Delta\langle x_j\rangle^{A,3} + \Delta\langle x_i\rangle^{B,3}\Delta\langle x_j\rangle^{B,3} + \Delta\langle x_i\rangle^{Z,3}\Delta\langle x_j\rangle^{Z,3}}{3}. \quad (28)$$

Numerical evaluation

To assess the correctness of the expressions presented at the previous sections, we performed PCA and combined-PCA for three different structures of human serum albumin (HSA). This protein consists of a single chain with 585 residues and contains 17 pairs of disulfide bridges.[25] The structures used in this numerical evaluation were: the apo form (HSA-apo), the form complexed with lauric acid (HSA-lau), and the form complexed with myristic acid (HSA-myr). In the following paragraphs, we provide the details of the setting of the MD trajectories and the implementation of the PCA and combined-PCA.

The crystal structures used in the MD simulations were taken from the Protein Data Bank. Their PDB-ID codes are 1AO6 (HSA-apo), 1E7F (HSA-lau), and 1E7G (HSA-myr). In these structures, HSA-lau and HSA-myr have eight binding sites occupied by the corresponding substrate molecules. The MD simulations as well as the analysis presented below were performed using the AMBER 12 suite of programs,[26] with the Amber99SB force field,[27] following equivalent protocols for the three systems. Briefly, the PDB files were fed into the Leap module. There, they were solvated in a TIP3P parallelepiped water box whose walls were 14.0 Å away from the solute. All crystallographic water molecules were conserved. Periodic boundary conditions were used, with a cutoff of 10.0 Å. The SHAKE algorithm was applied to all hydrogen atoms thus allowing an integration time step of 2.0 fs. The Particle Mesh Ewald method was used to calculate the long-range Coulomb forces.[28] The initial structures were first minimized at constant volume and then heated from 0 K to a target temperature of 310 K during 100.0 ps, using the weak-coupling algorithm with τ_{tp}= 2.8 ps. During this heating, the volume was kept constant. After this, we switched to constant temperature and pressure conditions, using a value of 2.0 ps for both τ_{tp} and τ_p. Finally, we allowed for 1 ns of equilibration before a production run of 30 ns. Snapshots were sampled every 20 ps. Thus, each production phase contains in 1500 frames.

We performed standard PCA for the three individual trajectories, as well as combined-PCA for the three possible pairs (HSA-apo/HSA-lau, HSA-apo/HSA-myr, and HSA-lau/HSA-myr) and for the three concatenated trajectories. Water molecules, ions, and substrate molecules were removed from the snapshots obtained from the MD simulations, thus rendering equivalent structures for the three systems. Then, following the standard procedures, all the structures were least-square fitted onto the same reference structure to eliminate overall translations and rotations. After that, the covariance matrices for the C_α atoms were built and diagonalized to yield the first 20 eigenvectors. Matrices $\mathbf{S}^2$ and $\mathbf{S}^3$ appearing in the combined-PCA were built and diagonalized using our home-made FORTRAN programs.

Results

General remarks on the equations

Typical MD trajectories used in PCA (and also in combined-PCA) are not long enough to allow the protein to change from one stable conformation to the other. Thus, the fluctuations observed during any individual trajectory are expected to be significantly smaller than the differences between the average structures of alternative conformations. In this situation, the main contribution to the concatenated correlation matrix comes from matrix $\mathbf{S}$, while the correlation matrices of the individual trajectories only add relatively minor details. In other words, the static contribution becomes the dominant one. It is therefore relevant to analyze the properties of the $\mathbf{S}$ matrices in more depth.

We concentrate on the expressions obtained for $\mathbf{S}^2$ and $\mathbf{S}^3$ [eqs. (10) and (28)]. A careful inspection shows that they could be written in a more general form as,

$$S_{ij}^n = \frac{1}{n}\sum_\alpha \left(\langle x_i\rangle^\alpha - \langle x_i\rangle^n\right)\left(\langle x_j\rangle^\alpha - \langle x_j\rangle^n\right), \quad (29)$$

where $n = 2$ or 3, while α runs over {*A*, *B*} or {*A*, *B*, *Z*}, according we are considering two or three concatenated trajectories. Written in this form, it is clear that matrices $\mathbf{S}^2$ and $\mathbf{S}^3$ also have the form of a correlation matrix: the correlation matrix of the individual average structures. Thus, $\mathbf{S}^2$ is formed with only two structures: the averages for trajectories *A* and *B*. Therefore, there is a

single direction in the $3N_{at}$-dimensional space that can explain all the variations contained in $\mathbf{S}^2$. This is the direction of the line that connects the average structure of *A* with that of *B*. But this is also the direction of the first eigenvector of $\mathbf{S}^2$. Moreover, the eigenvalue of this vector has a clear physical meaning that can be appreciated by considering eq. (13). The equation shows that the trace of matrix $\mathbf{S}^2$ gives the total squared deviation between the two structures under consideration, divided by four. After diagonalization, all these variations are accumulated into the eigenvalue of the first eigenvector of $\mathbf{S}^2$. Therefore,

$$4\lambda_1^{S^2} = N_{at}.\mathrm{RMSD}^2, \tag{30}$$

where $\lambda_1^{S^2}$ is the first eigenvalue of $\mathbf{S}^2$ while RMSD is the root mean square deviation between the average structures of *A* and *B*. Finally, as the first eigenvector accounts for all the variations contained in $\mathbf{S}^2$, the eigenvalues for the rest of the eigenvectors are necessarily null. Alternatively, one could say that the two points considered in $\mathbf{S}^2$ span a 1-dimensional subspace within the $3N_{at}$-dimensional space of the protein. Then, the diagonalization of matrix $\mathbf{S}^2$ just identifies the direction of the vector that spans this subspace.

Similarly, $\mathbf{S}^3$ is a correlation matrix formed with three structures: the averages for trajectories *A*, *B*, and *Z*. These structures span a plane in the $3N_{at}$-dimensional space of the protein (except if they are aligned, but that would be extremely rare). Accordingly, the diagonalization of matrix $\mathbf{S}^3$ only has two eigenvectors with non-negligible eigenvalues. These vectors span the plane that contains the three average structures. Their orientations correspond to the directions of the largest deviations between the individual averages and the average of the concatenated trajectory.

A completely different situation is found when all the trajectories involved in the combined-PCA correspond to the same well of the free energy surface. That would be the case, for example, of a combined-PCA involving alternative structures obtained from nuclear magnetic resonance (NMR) experiments of the same protein. If such trajectories are run for long enough, they perform an equivalent sampling of the configurational space. Accordingly, their average structures are all pretty similar and the elements of the static matrix become negligible. For that case, eq. (29) indicates that the combined correlation matrix is just the average of the individual correlation matrices. Thus, in this case, the combined-PCA does not provide new information since equivalent results could be obtained by just extending the run time of any of the individual trajectories involved.

Before closing this section, we note that the content of eqs. (9) and (27) can be summarized in a simple sentence: "the correlation matrix of the concatenated trajectory is equal to the average of the individual correlation matrices plus the correlation matrix of the individual average structures."

Convergence of the individual PCA

We checked the convergence of the individual PCA by assessing the convergence of the essential spaces (ES) determined by them. To stablish the size of the essential spaces, we projected the individual trajectories onto the eigenvectors of the corresponding individual correlations matrices. Vectors of the essential space were detected as those for which the projection does not fit into a Gaussian function.[2] It was found that the ES of each system was contained within the first 10 eigenvectors. They represent 90.5, 90.7, and 89.4% of the protein motion for HSA-apo, HSA-lau, and HSA-myr, respectively. These percentages were evaluated from the ratio between of the fluctuations contained within the ES and the fluctuations contained within the set of 20 eigenvectors computed in each case. Then, the convergence of the essential spaces was checked by calculating the root mean square inner product (RMSIP) between the ESs evaluated from independent subparts (*a* and *b*) of a given trajectory,

$$\mathrm{RMSIP} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\mathbf{v}_i^a.\mathbf{v}_j^b\right)^2}. \tag{31}$$

Here, *M* denotes the dimension of the ES (equal to 10 in this case) while $\mathbf{v}_i^a$ and $\mathbf{v}_j^b$ are the *i*th and *j*th eigenvectors obtained from subparts *a* and *b* of the given trajectory, respectively. In Figure 1, we show the time evolution of the RMSIP for the systems under study. As can be seen, in the three cases, the RMSIPs become time-independent after ≈2 ns, indicating that the individual PCAs are well converged.

Before considering the eigenvalues and eigenvectors of the combined-PCA for the three trajectories used in this numerical evaluation, it is necessary to compare the fluctuations observed within each trajectory with the differences between their average structures. As a measure of the average fluctuations within a given trajectory, we compute,

$$\langle \mathrm{RMSD}_{C_\alpha} \rangle = \frac{1}{N}\sum_{j=1}^{N}\left(\mathrm{RMSD}_{C_\alpha}\right)_j, \tag{32}$$

where *N* is the number of snapshots involved and $\left(\mathrm{RMSD}_{C_\alpha}\right)_j$ is the C_α-root mean square deviation between snapshot *j* and

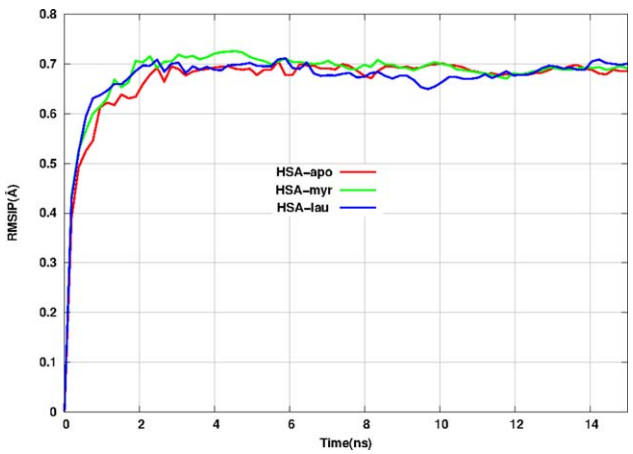

Figure 1. Time evolution of the RMSIPs for the three trajectories under analysis. The time indicated on the *x*-axis corresponds to the lengh of the independent simulations used to compute the RMSIP.

Table 1. Assessment of the contribution of matrix $\mathbf{S}^2$ to matrix $\mathbf{C}^2$ for the three pairs of trajectories under analysis.

	$\langle RMSD_{C_\alpha} \rangle$ [a]		$RMSD_{C_\alpha}$ [b]
Apo	1.67(0.35) Å	Apo-Myr	4.80 Å
Myr	1.44(0.30) Å	Apo-Lau	5.49 Å
Lau	1.77(0.28) Å	Lau-Myr	2.76 Å

[a] $\langle RMSD_{C_\alpha} \rangle$ measures the average fluctuations of the C_α atoms within an individual trajectory [see eq. (30)]. [b] C_α-Root mean square deviation between average structures of alternative trajectories. Numbers within parenthesis indicate the statistical uncertainties.

Table 2. Results obtained from the correlation matrix corresponding to the three concatenated trajectories.

	Apo-Myr	Apo-Lau	Lau-Myr
$\lambda_1^{S^2}$ [a]	3337.29 Å^2	4361.59 Å^2	1118.13 Å^2
$\lambda_1^{C^2}$ [b]	3435.47 Å^2	4499.40 Å^2	1231.73 Å^2
$V_1^{S^2} \cdot V_1^{C^2}$ [c]	0.999	0.999	0.998

[a] Eigenvalues of the first eigenvector of matrix S^2. [b] Eigenvalues of the first eigenvector of matrix C^2. [c] Scalar product between the first eigenvector of S^2 and that of C^2.

the average structure of the trajectory considered. Table 1 compares the $\langle RMSD_{C_\alpha} \rangle$ computed for each individual trajectory with the $RMSD_{C_\alpha}$ between the average structures of alternative trajectories.

It is observed that the differences between the average structures are larger than the fluctuations observed within each individual simulation. Therefore, the dominant contribution to the combined correlation matrix is the static one, embodied in matrix $\mathbf{S}^2$. An illustrative example of this situation is presented in Figure 2, where we compare matrices $\mathbf{C}^2$, $\mathbf{S}^2$, and the individual correlation matrices, for the trajectories of HSA-apo and HSA-myr. The similarities between matrices $\mathbf{C}^2$ and $\mathbf{S}^2$, and the differences between them and the correlation matrices of the individual trajectories, are readily appreciated. We note that matrices $\mathbf{C}^2$ and $\mathbf{S}^2$ have elements that are one order of magnitude larger than those of the individual correlation matrices. To make clearer the differences between these matrices, we used a logarithmic scale to plot the matrix elements. Similar pictures are obtained when the other pairs are considered.

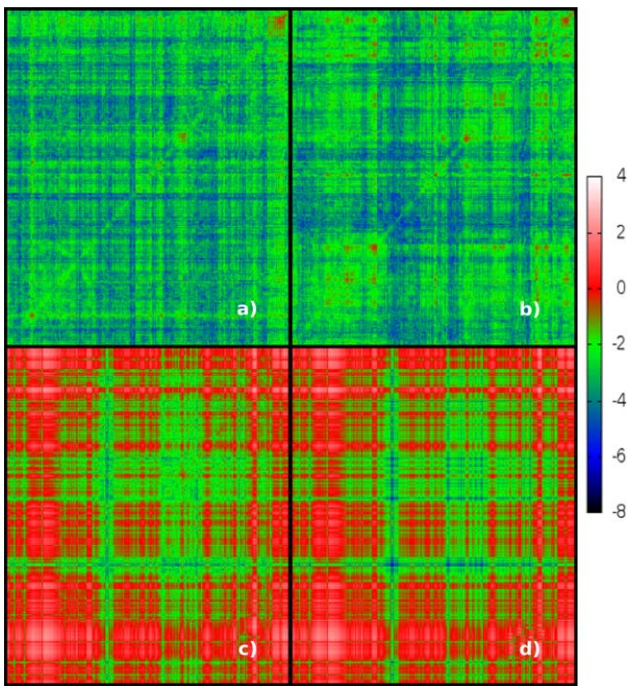

Figure 2. Correlations matrices involved in the combined-PCA of HSA-apo and HSA-myr. a) HSA-apo; b) HSA-myr; c) $\mathbf{C}^2$ matrix; and d) $\mathbf{S}^2$ matrix. A logarithmic scale was used to plot the contours of the absolute values of the matrix elements. This choice facilitates the visual inspection of the matrices (see text).

PCA of 2-concatenated trajectories

Table 2 compares the largest eigenvalue of each combined correlation matrix, $\mathbf{C}^2$, with the unique nonzero eigenvalue of matrix $\mathbf{S}^2$. The scalar product between these two vectors is also presented at the same table. As expected from the previous discussion, the scalar products are all close to one and the first eigenvalues of $\mathbf{C}^2$ and $\mathbf{S}^2$ are pretty similar to each other. The smallest scalar product and the largest relative difference between the eigenvalues were obtained in the combined-PCA of the two holo forms. This is because the $RMSD_{C_\alpha}$ between these two average structures is the smallest one. Therefore, the contribution of $\mathbf{S}^2$ is not so dominant in this case. Nevertheless, the differences are tiny.

For each combined-PCA, we computed the projections of the individual trajectories onto the first eigenvector of matrix $\mathbf{S}^2$. Then, we calculated the probability distributions for these projections. The results are plotted in Figure 3. As expected, the projections of the individual trajectories are located on either side of the *y*-axis. Moreover, for each pair, the averages of the individual distributions have the same absolute value, and the distances between them are just the square root of the corresponding eigenvalue. All these expectations, that derive from the equations presented previously, are closely fulfilled by the numerical examples presented in Figure 3. We end this section by noting that the eigenvalues of $\mathbf{S}^2$ given in Table 2 and the $RMSD_{C_\alpha}$ of Table 1 fulfil the relationship expressed in eq. (30), within the numerical accuracy of the values obtained.

PCA of 3-concatenated trajectories

In this case, the difference matrix has two eigenvectors with nonzero eigenvalues: $\lambda_1^{S^3} = 3235.03$ Å^2 and $\lambda_2^{S^3} = 683.00$ Å^2. Conversely, for matrix $\mathbf{C}^3$, we have $\lambda_1^{C^3} = 3369.09$ Å^2 and $\lambda_2^{C^3} = 781.98$ Å^2. Thus, the eigenvalues of corresponding eigenvectors of matrices $\mathbf{C}^3$ and $\mathbf{S}^3$ are similar to each other. The relative difference is somewhat larger for λ_2. The scalar products between corresponding eigenvectors are $\mathbf{v}_1^{S^3} \cdot \mathbf{v}_1^{C^3} = 0.999$ and $\mathbf{v}_2^{S^3} \cdot \mathbf{v}_2^{C^3} = 0.996$. Altogether these values confirm that the largest contribution to matrix $\mathbf{C}^3$ comes from matrix $\mathbf{S}^3$, while the individual correlation matrices only add minor details. The projections of the individual trajectories onto the first two eigenvectors of matrix $\mathbf{S}^3$ are presented in Figure 4. The average projection of each trajectory is indicated with a black dot

while their coordinates, in the plane expanded by vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, are given at the figure caption. We note that eigenvector $\mathbf{v}_1$ is nearly aligned with the line that connects the average of HSA-apo with the middle point between the averages of HSA-myr and HSA-lau. This occurs because the distance between the averages of HSA-myr and HSA-lau is significantly smaller than the distances between any of these structures and the average of HSA-apo. In the limiting case in which one of the distances is negligibly with respect to the other two, such alignment would be perfect. We finally note that, as expected, the distances between the averages in Figure 4 are the same as the corresponding distances in Figure 3.

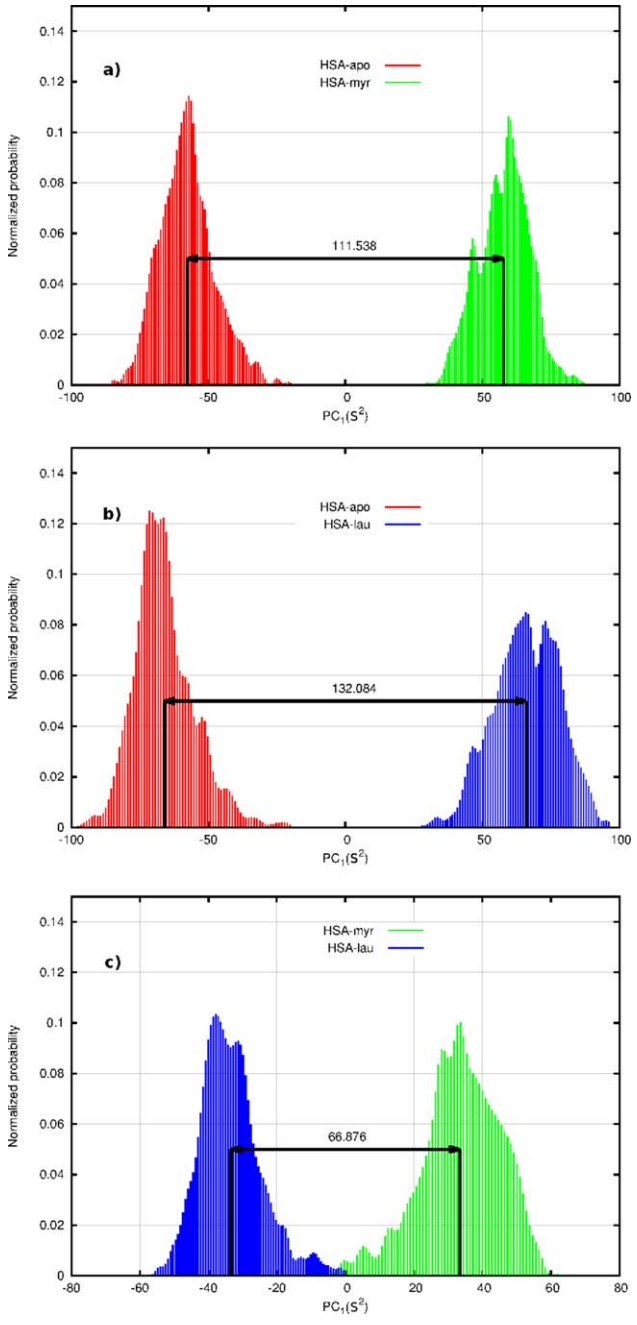

Figure 3. Probability distributions for the projections of the individual trajectories onto the first eigenvector of matrix $\mathbf{S}^2$. a) Combined-PCA for HSA-apo and HSA-myr; b) combined-PCA for HSA-apo and HSA-lau; and c) Combined-PCA for HSA-myr and HSA-lau. Lengths are measured in Å. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

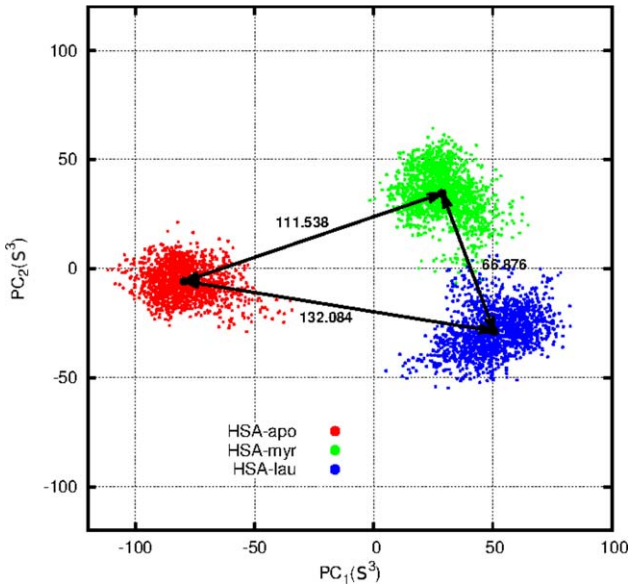

Figure 4. Projections of the individual trajectories onto the first two eigenvectors of matrix $\mathbf{S}^3$. The coordinates of the average projections are (−79.454, −5.779) for HSA-apo, (28.833, 34.503) for HSA-myr and (50.621, −28.724) for HSA-lau. Lengths are measured in Å. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Discussion

In this section, we summarize the main outcomes of this work. They derive from the formulas presented at the Methodology section and have been confirmed by the examples discussed at the Results section.

- The correlation matrix of a concatenated trajectory, $\mathbf{C}^n$, is given by the average of the individual correlation matrices $\sum_i \mathbf{C}^i/n$, plus the correlation matrix of the individual average structures, $\mathbf{S}^n$.
- When the RMSD_{C_α} between the individual average structures is significantly larger than the $\langle \mathrm{RMSD}_{C_\alpha} \rangle$ of the individual trajectories, the dominant contribution to $\mathbf{C}^n$ comes from $\mathbf{S}^n$.
- The number of eigenvectors with non-negligible eigenvalues of matrix $\mathbf{S}^n$ is $n-1$. When the dominant contribution to $\mathbf{C}^n$ comes from $\mathbf{S}^n$ these vectors appear at the first positions among the eigenvectors of $\mathbf{C}^n$, constituting the so-called "static modes."
- When the dominant contribution to $\mathbf{C}^n$ comes from $\mathbf{S}^n$, the projections of the individual trajectories onto the static modes are unavoidably large. Moreover, if they are plot in the subspace expanded by the $n-1$ eigenvectors of $\mathbf{S}^n$, the individual averages appear on either side of the axes. This occurs because, according to the standard PCA procedure, all the trajectories are referred to the global average structure. This structure necessarily lies in between the individual averages.

- In the case of two concatenated trajectories, much of the information contained in the static mode can be obtained by alternative means (i.e., by computing the average for each trajectory and the difference vector between them).
- A few eigenvectors of matrix $\mathbf{C}^n$ (the static modes) inform us about differences in the average structures of the trajectories considered. The rest of the eigenvectors afford dynamical information. However, their precise meaning is more difficult to envision. The expressions presented here indicate that they are the eigenvectors of a matrix obtained by first averaging the individual correlation matrices, and then projecting-out the directions of the static modes.

Conclusions

We have presented several analytical expressions relating the correlation matrix of a concatenated trajectory with the correlation matrices and the average structures of the trajectories involved. First, we derived the expression for two concatenated trajectories. Then, we developed the more general case of n concatenated trajectories, assuming that one already knows the formulas for $n-1$ trajectories. Finally, we use this more general equation to obtain a compact and illustrative expression for the case of three concatenated trajectories. The analysis of the cases considered shows that the results can be summarized in a single sentence: the correlation matrix of a concatenated trajectory is given by the average of the individual correlation matrices plus the correlation matrix of the individual averages. From this, it follows that the combined-PCA of trajectories belonging to different free energy basins provides information that could also be obtained by alternative and more straightforward means.

We performed combined-PCAs for two and three concatenated trajectories, using the structures of HSA in the the apo form as well as complexed with myristic acid and lauric acid. These numerical examples demonstrated the accuracy of the expressions presented in the article. We believe that these expressions will help to enlighten the interpretation of the results of combined-ED analysis and to guide their discussion.

Keywords: principal component analysis · essential dynamics · concatenated trajectories · conformational changes · protein dynamics

How to cite this article: G., Pierdominici-Sottile, J., Palma *J. Comput. Chem.* **2014**, DOI: 10.1002/jcc.23811

[1] I. Daidone, A. Amadei, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 762.
[2] A. Amadei, A. B. M. Linssen, H. J. C. Berendsen, *Proteins Struct. Funct. Bioinf.* **1993**, *17*, 412.
[3] M. A. Balsera, W. Wriggers, Y. Oono, K. Schulten, *J. Phys. Chem.* **1996**, *100*, 2567.
[4] A. Amadei, M. A. Ceruso, A. Di Nola, *Proteins Struct Funct Bioinf* **1999**, *36*, 419.
[5] J. B. Clarage, T. Romo, B. K. Andrews, B. M. Pettitt, G. N. Phillips, *Proc. Natl. Acad. Sci.* **1995**, *92*, 3288.
[6] B. Hess, *Phys. Rev. E* **2002**, *65*, 031910.
[7] O. F. Lange, H. Grubmuller, *J. Phys. Chem. B* **2006**, *110*, 22842.
[8] D. M. F. van Aalten, J. B. C. Findlay, A. Amadei, H. J. C. Berendsen, *Protein Eng.* **1995**, *8*, 1129.
[9] B. L. de Groot, D. M. F. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, H. J. C. Berendsen, *Proteins Struct. Funct. Bioinf.* **1997**, *29*, 240.
[10] A. Kitao, N. Go, *Curr. Opin. Struct. Biol.* **1999**, *9*, 164.
[11] H. J. C. Berendsen, S. Hayward, *Curr. Opin. Struct. Biol.* **2000**, *10*, 165.
[12] B. L. de Groot, X. Daura, A. E. Mark, H. Grubmuller, *J. Mol. Biol.* **2001**, *309*, 299.
[13] H. Lei, C. Wu, H. Liu, Y. Duan, *Proc. Natl. Acad. Sci.* **2007**, *104*, 4925.
[14] A. L. Tournier, J. C. Smith, *Phys. Rev. Lett.* **2003**, *91*, 208106.
[15] M. Karplus, J. N. Kushick, *Macromolecules* **1981**, *14*, 325.
[16] M. M. Teeter, D. A. Case, *J. Phys. Chem.* **1990**, *94*, 8091.
[17] T. Ichiye, M. Karplus, *Proteins Struct. Funct. Bioinf.* **1991**, *11*, 205.
[18] O. F. Lange, H. Grubmuller, *Proteins Struct. Funct. Bioinf.* **2006**, *62*, 1053.
[19] H. Kamberaj, A. van der Vaart, *Biophys. J.* **2009**, *97*, 1747.
[20] D. M. F. van Aalten, A. Amadei, A. B. M. Linssen, V. G. H. Eijsink, G. Vriend, H. J. C. Berendsen, *Proteins Struct. Funct. Bioinf.* **1995**, *22*, 45.
[21] G. H. Peters, D. M. F. van Aalten, A. Svendsen, R. Bywater, *Protein Eng.* **1997**, *10*, 149.
[22] M. Filizola, S. X. Wang, H. Weinstein, *J. Comput.-Aided Mol. Des.* **2006**, *20*, 405.
[23] J. Shan, G. Khelashvili, S. Mondal, E. L. Mehler, H. Weinstein, *PLoS Comput. Biol.* **2012**, *8*, e1002473.
[24] N. Bung, M. Pradhan, H. Srinivasan, G. Bulusu, *PLoS Comput. Biol.* **2014**, *10*, e1003484.
[25] A. Dugaiczyk, S. W. Law, O. E. Dennison, *Proc. Natl. Acad. Sci.* **1982**, *79*, 71.
[26] D.A. Case, T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Götz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P. A. Kollman **2012**, AMBER 12, University of California, San Francisco.
[27] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. L. Simmerling, *Proteins Struct. Funct. Bioinf.* **2006**, *65*, 712.
[28] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98*, 10089.

APPENDIX

Here, we derive an expression for the elements of the correlation matrix of two concatenated trajectories, for the more general situation in which the number of structures taken from the individual trajectories is not the same. Let $\{\mathbf{x}\}$ be a set of N_t snapshots obtained from MD simulations of trajectories *A* and *B*. We assume that the first N_A structures of the set were taken from trajectory *A* while the remaining $N_B = N_t - N_A$ come from trajectory *B*. The *ij*th element of the combined correlation matrix fulfils,

$$N_t \; C_{ij}^{AB} = \sum_{k=1}^{N_t} \left(x_i^{(k)} - \langle x_i \rangle^{AB} \right) \left(x_j^{(k)} - \langle x_j \rangle^{AB} \right). \tag{A1}$$

Here, $\langle x_i \rangle^{AB}$ and $\langle x_j \rangle^{AB}$ are the averages of coordinates x_i and x_j computed with the set $\{\mathbf{x}\}$. These averages are related to the averages computed from trajectories *A* and *B* according to,

$$N_t \langle x_i \rangle^{AB} = N_A \langle x_i \rangle^{A} + N_B \langle x_i \rangle^{B}, \tag{A2}$$

$$N_t \langle x_j \rangle^{AB} = N_A \langle x_j \rangle^{A} + N_B \langle x_j \rangle^{B}. \tag{A3}$$

The summation in eq. (A1) is now split into two summations, one containing the structures taken from trajectory *A* and one with the structures taken from trajectory *B*,

$$N_t C_{ij}^{AB} = \Sigma^A + \Sigma^B, \tag{A4}$$

with

$$\Sigma^A = \sum_{k=1}^{N_A} \left(x_i^{(k)} - \langle x_i \rangle^{AB} \right) \left(x_j^{(k)} - \langle x_j \rangle^{AB} \right), \tag{A5}$$

$$\Sigma^B = \sum_{k=N_A+1}^{N_t} \left(x_i^{(k)} - \langle x_i \rangle^{AB} \right) \left(x_j^{(k)} - \langle x_j \rangle^{AB} \right). \tag{A6}$$

Following a procedure similar to those used in the main text of this work, the expressions for Σ^A and Σ^B become,

$$\Sigma^A = N_A C_{ij}^A + N_A \Delta\langle x_i \rangle^{A,AB} \Delta\langle x_j \rangle^{A,AB}, \tag{A7}$$

$$\Sigma^B = N_B C_{ij}^B + N_B \Delta\langle x_i \rangle^{B,AB} \Delta\langle x_j \rangle^{B,AB} \tag{A8}$$

where $\Delta\langle x_i \rangle^{A,AB}, \Delta\langle x_j \rangle^{A,AB}$, so forth have the meaning expressed in eq. (6). Replacing these expressions in eq. (A4) we obtain,

$$C_{ij}^{AB} = w_A C_{ij}^A + w_B C_{ij}^B + S_{ij}^{AB}, \tag{A9}$$

where $w_A = N_A/N_t, w_B = N_B/N_t$, while the static contribution to the combined correlation matrix is given by

$$S_{ij}^{AB} = w_A \Delta\langle x_i \rangle^{A,AB} \Delta\langle x_j \rangle^{A,AB} + w_B \Delta\langle x_i \rangle^{B,AB} \Delta\langle x_j \rangle^{B,AB}. \tag{42}$$

Finally, using the relationships given in eqs. (A2) and (A3), the expression for S_{ij}^{AB} simplifies to

$$S_{ij}^{AB} = w_A w_B \Delta\langle x_i \rangle^{A,B} \Delta\langle x_j \rangle^{A,B}. \tag{A10}$$

which becomes equal to eq. (13) when the number of snapshots taken from each trajectory is the same. Moreover, as the factor $w_A w_B$ is maximum when $w_A = w_B = 0.5$ this last equation demonstrates that the usual procedure of using the same number of snapshots for both trajectories maximizes the static contribution to the combined correlation matrix.

Received: 21 August 2014
Revised: 12 November 2014
Accepted: 23 November 2014
Published online on 00 Month 2014