JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

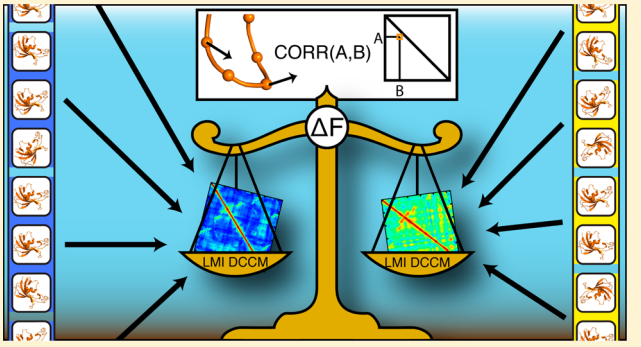
(Dis)similarity Index To Compare Correlated Motions in Molecular Simulations

Matteo Tiberti, Gaetano Invernizzi, and Elena Papaleo*,†

Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy

Supporting Information

ABSTRACT: Molecular dynamics (MD) simulations are widely used to complement or guide experimental studies in the characterization of protein dynamics, thanks to improvements in force-field accuracy, along with in the software and hardware to sample the conformational landscape of proteins. Among the different applications of MD simulations, the study of correlated motions is largely employed for different purposes. Several metrics have been developed to describe correlated motions in the MD ensemble, such as methods based on Pearson Correlation or Mutual Information. Cross-correlation analysis of MD trajectories is indeed appealing not only to identify residues characterized by coupled fluctuations in protein structures but also since it can be used to extrapolate motions along directions in which major conformational changes should occur, for example on longer time scales than the ones that are actually simulated. Nevertheless, most of the MD studies employ average correlation maps and mostly in a qualitative way, even when different systems or different replicates of the same system are compared. The broad application of correlation metrics in the analysis of MD simulations, especially for comparative purposes, requires a step forward toward more quantitative and accurate comparisons. We thus here employed a simple but effective index, which is based on a normalized Frobenius norm of the differences between protein correlation maps, to compare correlated motions. We applied this index for a quantitative comparison of correlated motions from MD simulations of seven proteins of different size and fold. We also employed the index to assess the robustness of correlation description when multi-replicate MD simulations of a same system are used, and we compared our index to metrics for comparison of structural ensembles such as Root Mean Square Inner Product and the Bhattacharyya Coefficient.



1. INTRODUCTION

Proteins are dynamic entities over different time scales, and the understanding of their conformational dynamics is fundamental for a complete picture of their biological function.[1−4] Small or large structural changes can occur in proteins, ranging from fast local fluctuations, as for example the ones upon ligand binding, or slow and global fluctuations as the ones involved in allosteric events.[2,5−9] In this context, molecular dynamics (MD) simulations are used to complement or guide experimental studies in the characterization of protein dynamics.[3,10−12] Different aspects can be investigated through protein MD simulations, such as residues characterized by coupled fluctuations.[13] Indeed, different metrics to describe correlated motions from MD ensembles have been proposed,[14,15] and since then they have been widely used in several applications.[16−29]

One of the first metrics proposed to analyze residues characterized by coupled fluctuations was based on the Pearson Correlation Coefficient, i.e., the dynamical cross-correlation matrix (DCCM).[14,30] In this case, cross-correlation coefficients vary from a value of −1 for anticorrelated motions to 1 when they are completely correlated (Figure 1A). The DCCM does not bear information about the magnitude of the motions, which could thus be either small local oscillations or large amplitude collective motions. It reflects correlations of displacements along a straight line, and thus if two atoms move exactly in phase and with the same period, but along perpendicular directions, they will be characterized by a cross-correlation of zero. As a consequence, only atoms moving along the same direction will give an accurate correlation value. Moreover, the covariance matrix upon which DCCM depends only contains linear correlations. This is not suitable for those protein motions that exhibit both linear and nonlinear components. To overcome these limitations, a generalized correlation approach based on Mutual Information (MI) and Linear Mutual Information (LMI, Figure 1B) was introduced.[15,31] MI allows for identifying more correlations than DCCM and to capture both linear and nonlinear correlations. Recently, other methods have been proposed in which either the time series of atom distances from a moving average[32,33] or

Received: February 17, 2015

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

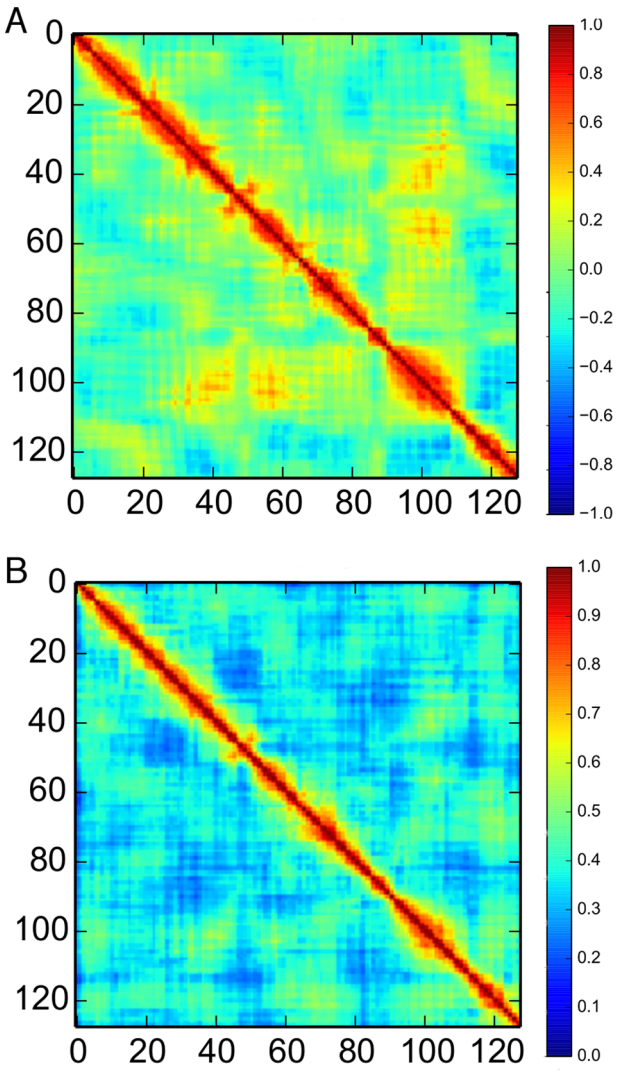


Figure 1. Correlation maps from protein MD simulations. Here we report an example of a correlation map calculated from 1 μs MD simulation of one of the target proteins used in this study (Dri DNA-binding domain) using the DCCM (A) or LMI (B) methods.

the recurrences of atom distances in the time series[34] are taken into account.

DCCM and MI-based methods have been extensively used in the study of protein dynamics[16−25] also thanks to the availability of computational tools to calculate them from MD trajectories that are available in almost all the MD packages. Both the approaches are, fox example, implemented in the *Wordom* toolkit for MD analysis.[35] Cross-correlation analysis is indeed appealing not only to identify coupled residues in the protein structure but also because it can be used to extrapolate motions along directions in which major conformational changes should occur, for example on longer time scales, without explicitly simulating them.[14] The seminal study by Hünenberger et al.[14] showed that it is crucial to verify that the atomic fluctuations of the system and, as a consequence, the cross-correlations between those fluctuations are sampled in a reproducible way. A dynamical cross-correlation analysis strongly depends on the time scale over which data on correlations have been collected. Despite this issue having been raised 20 years ago, most of the studies employing metrics for correlated motions in MD ensembles are using them in a qualitative way and often recurring to average values over the whole simulation time. These approaches can be subject to risk of having for instance a positive correlation in one map and a negative correlation in another map of the same system or in different portions of the same trajectory. These observations thus question the validity of comparing only average cross-correlation matrices from distinct simulations.

Moreover, the metrics for correlation motions, such as DCCM and LMI, are widely used not exclusively to identify cross-correlations between atomic fluctuations. Indeed, they are also integrated with graph-theory methods to detect paths of communications between distal sites in the protein,[36−50] or they are combined with methods for statistical coupling analysis.[51−54] The distance covariance has also been proposed to measure correlated motions in protein ensembles.[55] The MI-based methods can also be used to extract the causality of correlated motions in MD simulations and to reveal how correlated motions are used to transmit information through the protein.[56] Interestingly, difficulties related to the usage of average correlation metrics from a trajectory or arbitrary cutoffs have recently been pointed out.[57] The application of correlation metrics in MD analysis thus requires a more conscious usage, and we need to move a step forward toward a quantitative comparison of correlation maps.

We here contribute to this scenario, applying an index to compare correlation maps of six proteins of different size and fold. The index is based on the differences of the Frobenius norm of correlation maps normalized on the maximal difference that can be observed for the metrics under investigation. We here applied this index for a quantitative comparison of correlation motions and overall dynamics of MD simulations of two different DNA-binding domains of the ARID (AT-rich Interactive Domain) family of transcription factors, barnase, a cell-cycle inhibitor (p19INK), FK506 binding protein, ribonuclease H, and a bacterial xylanase. We illustrated its applications, and we defined a threshold of normalized Frobenius norm difference that allowed the identification of similar or dissimilar correlation patterns. Moreover, we used the index to compare the robustness of the correlation description across different MD replicates of the same protein. We also compared it with other metrics for ensemble comparison, such as Root Mean Square Inner Product (RMSIP)[58−60] and the Bhattacharyya coefficient.[9,61−66] The index here proposed is not limited to the comparison of DCCM- or MI-derived correlation maps, but it can be employed more broadly to a quantitative comparison of any metrics that can be represented as a list of quantitative pairwise relationship between protein residues.

2. MATERIALS AND METHODS

2.1. Molecular Dynamics (MD) Simulations. We performed explicit solvent MD simulations using the *GROMACS* software (www.gromacs.org) with the CHARMM22* force field[67] and the TIP3P solvent model.[68] The simulations have been carried out at 300 K and 1 bar in the isothermal−isobaric (NPT) ensemble. We employed periodic boundary conditions, and we set a distance equal or greater than 1.0 nm from the protein atoms and the box edges of a dodecahedral box. A number of water molecules equal to the protein net charge was replaced by counterions.

Each system was initially relaxed by 10000 steps of energy minimization by the steepest descent method, and then solvent equilibration was carried out for 50 ps at 300 K, while

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

restraining the protein atomic positions using a harmonical potential. Each system was then slowly equilibrated to the target temperature (300 K) and pressure (1 bar) through a thermalization and a series of pressurization simulations of 100 ps each. The LINCS algorithm was used to constrain heavy-atom bonds, allowing for a 2 fs time-step. Long-range electrostatic interactions were calculated using the Particle-Mesh Ewald (PME) summation scheme. van der Waals and short-range Coulomb interactions were truncated at 0.9 nm. The nonbonded pair list was updated every 10 steps, and conformations were stored every 4 ps.

We used as starting structures for our simulations, the NMR structures of the human ARID3A (PDB entry 2KK0),[69] the *Drosophila melanogaster* Dead ringer domains (Dri, PDB entry 1C20),[70] the X-ray structures of barnase (PDB entry 1A2P),[71] the human cyclin-dependent kinase inhibitor p19INK 4d (PDB entry 1BD8),[72] the FK506 binding protein (PDB entry 1BKF),[73] *Escherichia coli* ribonuclease H (PDB entry 2RN2),[74] and *Bacillus circulans* xylanase (PDB entry 1XNB, unpublished structure). These proteins have been selected since they are either proteins for which we already carried out investigation of correlated motions,[42] or they belong to a dataset of protein structures used for benchmarking of network analysis of protein structures.[75] We also carried out a 100 ns unfolding simulation at 500 K for each system. The five 100 ns replicates of both ARID3A and Dri, along with 1-μs simulation of Dri, have been previously published.[42] The main-chain Root Mean Square Deviation (RMSD) profiles over the simulation time for each trajectory used in this study are reported in Figure S1 as a reference. The 1-μs MD simulation of ARID3A, the four 100 ns MD replicates of the other five proteins, and their unfolding simulations are here discussed for the first time. Table 1 reports a summary of the proteins and number and length of the MD replicates that were analyzed in our study.

Table 1. Summary of MD Simulations Collected in This Study

system	PDB entry	no. of MD replicates	simulation length
ARID3A	2KK0	6	5 × 100 ns
			1 × 1 μs
Dri	1C20	2	1 × 1 μs
			1 × 100 ns (500 K)
barnase	1A2P	5	4 × 100 ns
			1 × 100 ns (500 K)
P19INK 4d	1BD8	5	4 × 100 ns
			1 × 100 ns (500 K)
FK506 binding protein	1BKF	5	4 × 100 ns
			1 × 100 ns (500 K)
ribonuclease H	2RN2	5	4 × 100 ns
			1 × 100 ns (500 K)
Bacillus xylanase	1XNB	5	4 × 100 ns
			1 × 100 ns (500 K)

2.2. Calculation of DCCM and LMI Correlation Metrics.

DCCM and LMI correlation matrices have been calculated as described by Hünenberger et al.[14] and Lange and Grubmüller,[15] respectively. We used the *Wordom* package[35] to calculate them. *Wordom* outputs have been then converted to the *xPyder*- (https://github.com/ELELAB/xpyder)[76] and *PyInteraph*- (https://github.com/ELELAB/pyinteraph)[77] compatible format by the *Python* tool *corr2dat.py*. The *Python* tool *deltmat.py* can be then used to compare different correlation matrices and calculate the normalized Frobenius norm of the differences between them, as well as to obtain two-dimensional plots of the correlation maps and their differences and other additional information on the comparison. The command line for the tools is provided running the script with the $-h$ option. Both the *Python* tools are available free of charge as Open Source at https://github.com/ELELAB/xpyder under the GPL v2 license. The tools require *Python* (2.7 or higher), *numpy* ($\geq$1.8.0), and *matplotlib* ($\geq$1.3.0) *Python* libraries.

2.3. Calculation of Root Mean Square Inner Product and Bhattacharyya Coefficient.

The Root Mean Square Inner Product (RMSIP) and the Bhattacharyya Coefficient (BC) are measures of the amount of overlap between two samples. They range from 0 to 1, where 1 indicates completely overlapping sets, while 0 indicates completely independent sets. In this case we have used them to compare the two halves of each trajectory. RMSIP was calculated as described in ref 58, on the first 20 eigenvectors from a Principal Component Analysis (PCA) of the Cα covariance matrix of the atomic positional fluctuations.[78]

BC was calculated following a procedure previously described,[62] using the atomic covariance matrix derived from our MD trajectories as an input. In particular, BC was calculated as

$$\mathrm{BC} = \exp\left(-\frac{1}{2q}\ln\left(\frac{|\Lambda|}{\sqrt{(|\mathbf{Q}^{\mathrm{T}}\mathbf{C}_{\mathbf{A}}\mathbf{Q}||\mathbf{Q}^{\mathrm{T}}\mathbf{C}_{\mathbf{B}}\mathbf{Q}|)}}\right)\right)$$

where $|\mathbf{X}|$ denotes the determinant of matrix $\mathbf{X}$, $\mathbf{Q}$ is the matrix of the principal components of $(\mathbf{C}_{\mathbf{A}} + \mathbf{C}_{\mathbf{B}})/2$, Λ is a diagonal matrix containing the corresponding eigenvalues, and q is the number of modes needed to capture 90% of the variance of $\mathbf{Q}$. BC was calculated using the implementation found in the *Bio3D* package for *R*.[65]

3. RESULTS AND DISCUSSION

3.1. A Normalized Frobenius Norm of the Differences between Correlated Maps to Compare MD Ensembles.

To quantify the differences among different correlation matrices, we calculated the Frobenius norm of their differences according to the following equation

$$\Delta F = \frac{1}{N}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(a_{ij} - b_{ij})^2}$$

where a_{ij} and b_{ij} are the elements of the correlation maps A and B, respectively, whereas the matrix order m is equal to the number of residues of the target protein, and N is the normalization factor (see below).

We introduced ΔF, a normalized Frobenius norm of the difference correlation matrix, to allow for a better quantitative comparison between different simulations of the same system (i.e., different MD replicates, changes in the simulation conditions, or simulations carried out with different force fields) or of simulations of different variants of the same protein (i.e., mutants, modified variants upon post-translational modifications, bound versus free states of a protein, ...). The normalization value (N) that we introduced here is the maximum Frobenius norm of the differences for a specific metrics of correlation motions and a protein of a certain size.

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

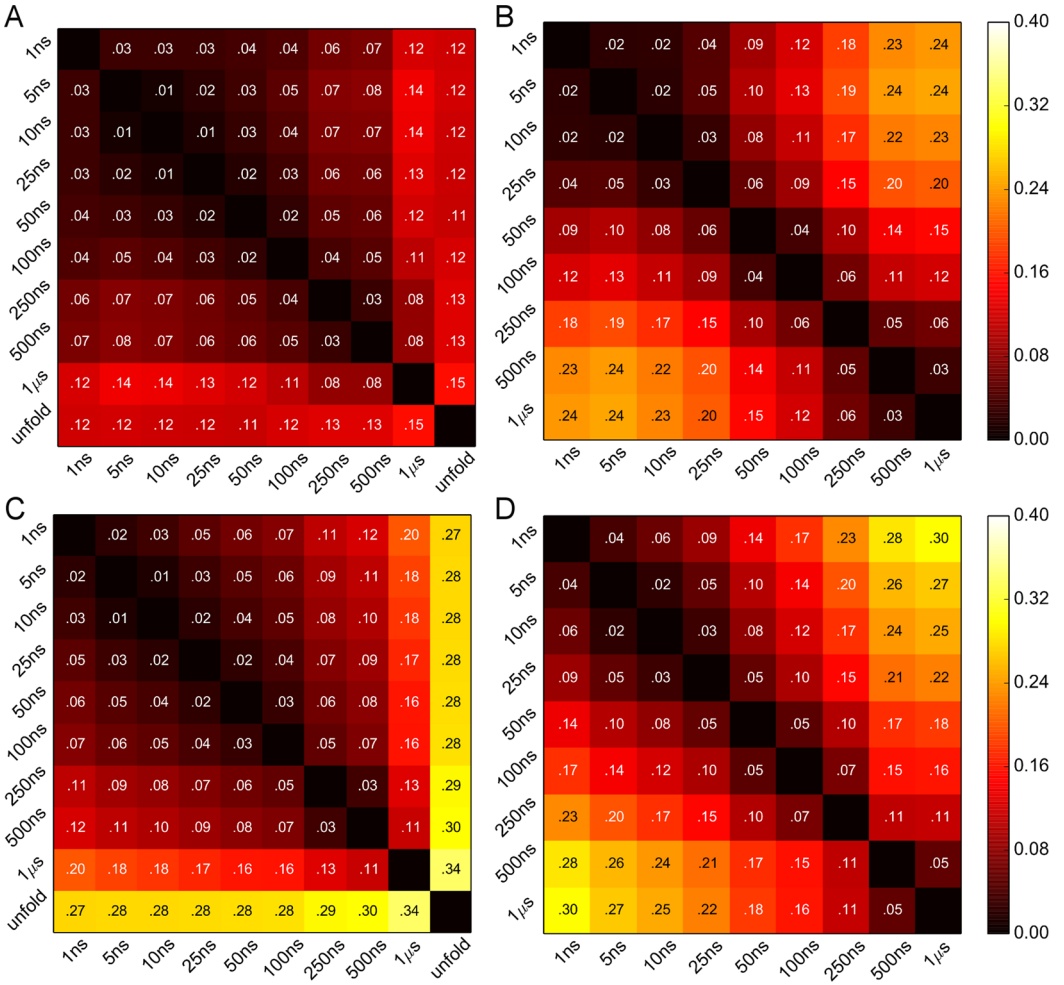


Figure 2. Frobenius norm of the differences between correlation maps calculated using different time windows for the averaging. We show a heat map of the pairwise comparisons of correlated maps derived by LMI (A,B) or DCCM (C,D) methods and using different time windows for the averaging. The ΔF values referred to each pairwise comparison are also reported in the heat map. In particular, we analyzed 1-μs MD simulations of Dri (A, C) and ARID3A (B, D). The length of each time windows used for the averaging is indicated by each label. "1 μs" and "unfold" correspond to the correlation maps calculated over the whole 1-μs MD simulation of the native proteins or over a 100 ns 500 K unfolding MD simulation, respectively.

The maximum Frobenius norm is thus calculated as the Frobenius norm of the differences between one matrix that includes as elements only the highest possible values of correlations and one matrix with all the lowest values of correlations. It thus depends on the minimum and maximum correlation values of the correlation metrics that we employ and on the matrix size (i.e., on the protein size). For LMI (or MI) correlation maps, the normalization faction is thus equal to the number of protein residues since these metrics range from 0 to 1, and the largest difference for each pair of residues in the correlation map is 1. In the DCCM case the normalization factor is two-fold the number of residues in the structure since the correlations are within a range between −1 and 1. ΔF is expected to be 0 for identical correlation maps and 1 when the two correlation maps are as different as possible.

3.2. Does Average Correlation Maps of the Same Proteins Calculated over Different Time Scales Provide the Same Description of Correlated Motions? In many applications, as stated in the introduction, an average correlation map from the whole MD trajectory is employed to describe residues characterized by coupled motions. In other cases, a correlation map averaging the correlations over a specific time window is used. It is not clear if the two approaches provide a sufficiently similar description of correlated motions and what procedure is better to adopt. We here tested the two different approaches on 1-μs MD simulations of two DNA-binding protein domains of ∼128 residues, i.e., the human ARID3A and the *D. melanogaster* Dri domain. The main-chain root-mean-square deviation (RMSD) profiles features different stabilities over the simulation time in the two proteins (Figure 1S). The MD simulation of Dri indeed features a more stable RMSD profile with minor fluctuations mainly due to the DNA-binding loops and the N- and C-terminal tails. The MD simulation of the NMR-derived structure of the ARID domain, on the contrary, is characterized by larger fluctuations mainly due to conformational heterogeneity of the C-terminal α-helix and loop dynamics. These two simulations thus provide us with two suitable candidates to compare correlated motions calculated over different time scales in cases in which we have reasonably stable or less stable MD trajectories. Moreover, we also included in the comparison a MD trajectory of Dri that we collected at 500 K for 100 ns[42] and in which the protein is partially unfolded. We used this unfolding trajectory as a control to estimate an upper limit of

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Frobenius norm difference to classify two simulations as featuring distinct correlated motions. At first, we calculated different correlation maps using both DCCM and LMI methods for each system, averaging over the whole trajectory (1 μs) or over shorter time windows (500, 250, 100, 50, 25, 10, 5, or 1 ns) (Figure 2). The comparison between the unfolding simulation of Dri and the simulations of the native protein suggests that a Frobenius norm difference (ΔF) of more than 0.12, which corresponds to ~15 pairs of residues characterized by different cross-correlation, is already an index of different correlated motions. This is close to the ΔF values achieved when the correlation map calculated from the whole 1-μs trajectory and any of the other average maps are compared. We also noticed that averages calculated on different time scales, i.e. on tens or on hundreds of ns, account for different correlated motions with ΔF values in the range of 0.06−0.08. On the contrary, when we used time windows that account for similar time scales (i.e., tens or hundreds of ns) for the averaging, the ΔF analysis highlights very similar patterns of correlated motions with ΔF values lower than 0.03. This trend is observed especially in the Dri simulation, whereas it is less evident, even if still detectable, in the less stable ARID3A trajectory. These results thus support the notion that a single average correlation map calculated over the whole trajectory is not a good estimate of the correlated motions. Indeed, it makes difficult to discriminate if the differences observed between the correlation map calculated from the whole 1-μs trajectory and the correlation maps calculated using shorter time windows are really genuine differences related to different motions sampled at different time scales. On the contrary, when we employed a sufficient number of correlation maps (more than 5−10 in our study) for the averaging, collected over shorter time windows, the description of correlated motions is more consistent. We should also notice that correlation analysis carried out with LMI (Figure 2A-B) and DCCM (Figure 2C-D) methods provided similar results.

3.3. Only Correlated Motions Occurring at Least on a Time Scale 10-Fold Shorter than the Actual Simulation Length Can Provide Significant Results. The results shown in Figure 2 suggest that only the averaging over a sufficient number of correlation maps, i.e. using smaller time windows with respect to the whole simulation length, provides a better description of the correlated motions. We thus estimated the convergence of the correlation maps derived by two different averaging procedures, i.e. calculating only one average correlation map from the whole trajectory or averaging correlation maps calculated over a 100 ns time window. With this goal in mind, we have thus monitored the evolution over the simulation time of ΔF between correlation maps achieved with the two procedures in the 1-μs MD simulation of Dri (Figure 3A). We observed that when the length of the time window is set equal to the simulation time, the average correlation map does not converge and features even in the last portion of the trajectory ΔF values close to the upper limit of 0.12. On the contrary, averaging over at least a 100 ns time window allows the correlation map to converge after 400 ns of simulation and reach ΔF values close to 0.

Overall, our data suggest that with 1-μs MD simulation in hand we can accurately estimate only correlated motions occurring on 10-fold shorter time scales or even shorter. Since we are using average matrices in any of the cases discussed above, it becomes thus crucial to verify that the average matrix really reflects the motions occurring on that specific time scale

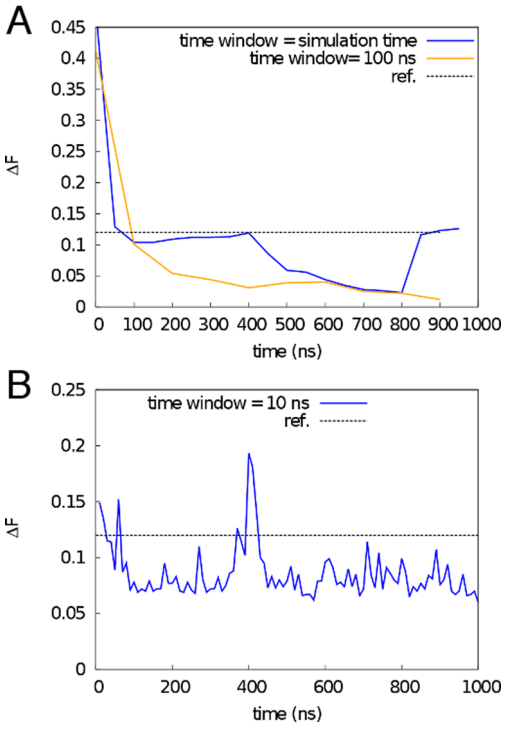


Figure 3. Convergence of correlation maps calculated using different time windows for the averaging. A) Convergence over the simulation time of correlation maps derived using a time window that is equal to the simulation length (i.e., using only one average correlation map from the whole trajectory, blue line) or averaging over 100 ns time window (yellow line). At each simulation time, we report the ΔF between the average correlation map relative to that specific simulation portion and the correlation map from the whole 1-μs MD trajectory and obtained using the same time windows for the averaging (i.e., one unique average map vs 100 ns time windows). B) The 1-μs MD trajectory has been divided in separate intervals of 10 ns. The ΔF between each individual correlation map calculated for each 10 ns interval and the average correlation map from the whole trajectory using a 10 ns time window for the averaging are reported over the simulation time. The 1-μs MD simulation of Dri is used for both the analyses in panels A and B, and the upper limit of ΔF estimated comparing correlation maps of native and partially unfolded Dri simulations (Figure 2) is shown as a reference as a black dotted line in both the panels.

over the simulation time. We thus monitored ΔF between the individual correlation maps calculated every 10 ns and the average correlation map calculated using the 10 ns time window in the Dri 1-μs simulation (Figure 3B). This example shows that the averaged correlation map does not necessarily reflect the correlation motions occurring over that specific window of time in the entire simulation, even on such short time scales and even with a MD simulation with an overall stable main chain RMSD profile. Indeed, there are several regions of the MD trajectory where the 10 ns correlation maps feature ΔF values larger (Figure 3B, blue line) than the ones observed comparing the native and partially unfolded Dri simulations (Figure 3B, dotted line). The same behavior holds when the 100 ns time window is used for the averaging and even amplified (*data not shown*). These data thus point out that it is crucial to carry out in-depth analysis, such as the ones presented here before employing only a unique average correlation map in the analysis of MD ensembles.

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

3.4. ΔF To Compare Different Replicates of the Same System.

We know that one of the major issues with canonical MD simulations is the quality of the sampling of the conformational space. Indeed, the target molecule encounters the risk to be entrapped in local minima for a long simulation time when we use canonical MD. Thus, it becomes crucial to assess the reproducibility of results achieved by MD simulations. One of the strategies more frequently employed is to collect more than one MD replicate of the same system either using different initial structures of the same molecule for the simulation (if more than one experimental deposition is available) or initializing the system with different atomic velocities taken from a Maxwellian distribution.[79] Another strategy is to carry out the simulation with more than one force field.[80]

We thus investigated how average correlation maps achieved by LMI and DCCM are described by independent replicates of the same system.

With this aim in mind, we first collected at least four 100 ns MD replicates of six different proteins (including ARID3A, barnase, a cell-cycle inhibitor (p19INK), FK506 binding protein, ribonuclease H, and a bacterial xylanase, Table 1), and we compared them both intra- and inter-replicates using ΔF. At first, we verified the ΔF value to use as an "upper limit" and assessed if the one proposed above is protein-dependent. Indeed, we are using the ΔF calculated between MD trajectories at 300 K and an unfolding trajectory at 500 K of the same protein as an estimate of an upper limit to classify two simulations as characterized by distinct correlated motions. The ΔF upper limit value of approximately 0.12 was referred to the lowest ΔF value obtained for the comparison between the unfolding trajectory and the 300 K trajectories of the ARID domain MD ensembles. We thus compared for each of the five additional target proteins the ΔF between the 500 K simulations and each of the four 100 ns replicates at 300 K, using both the 50 ns and 10 ns time window for the averaging and either LMI or DCCM to estimate the correlated motions. We also carried out intrareplicate comparisons i.e., between the two halves of each trajectory at 300 K. We observed in Figure 4 that there is always a clear distinction between the intrareplica and the 300 K vs 500 K comparisons independently on the protein fold and size and the method used for the averaging. We also noticed that LMI is more robust than DCCM with narrower peaks centered on very low ΔF values. This analysis suggests that we can use a ΔF value of approximately 0.12−0.15 to classify two correlation maps as clearly distinct. Indeed replicates which feature ΔF values higher than this value in intrareplica comparisons are the ones for which we observed the less stable main-chain RMSD profiles over the simulation time (Figure S1).

We then also calculated the similarity between the two halves of each MD replicate, expressed as RMSIP and BC, which are often used as metrics for ensemble comparison, and we compared them to the intrareplicate ΔF values with a time window of either 10 or 50 ns for the averaging (Figure 5). For this comparison the ΔF values have been normalized against the upper limit of 0.12 (ΔF^*) so that also this metric can range from 0 to 1, similarly to RMSIP and BC. One should remember that in the case of RMSIP and BC higher values mean more similar ensembles, whereas in the case of ΔF^* we observed the inverse relationship i.e., the closer the ΔF^* is to 0 the more similar the correlation maps are. Our data clearly show that both DCCM and LMI provide similar results and are highly correlated between each other, at least for the seven proteins here investigated. On the contrary, the correlation between this analysis and either RMSIP and especially BC is low, as well as the one between BC and RMSIP themselves. High RMSIP or BC values (higher overlap between the two halves of each trajectory) do not necessarily imply similar correlated motions, i.e., lower ΔF values. Indeed, RMSIP and BC values are generally higher than 0.7−0.8 for all the intrareplica comparisons that we collected (Figure S2). This would suggest a resampling of the same conformational spaces in the two halves of each trajectory, and, in turn, we would expect also a higher agreement between the correlated motions described by the two portions of each trajectory. On the contrary, this is not necessarily observed in our comparisons. Even trajectories in which the RMSIP value is equal or higher to 0.80 feature distinct correlated motions, even close to the ΔF upper limit value of 0.12. The criticality in the usage of RMSIP metrics in ensemble comparison has been also recently pointed out[81] in 10-μs simulations of fairly stable folded proteins with different force fields. RMSIP indeed provides an index of similarity of the conformational subspaces that are covered by two simulations or two portions of the same trajectory, whereas it does not provide information on the population of each state that has been sampled.

We here observed that an assessment of the sampling quality uniquely based on metrics for ensemble comparison that employ the covariance matrix of atomic fluctuations is not sufficient to ensure that the correlated motions will be

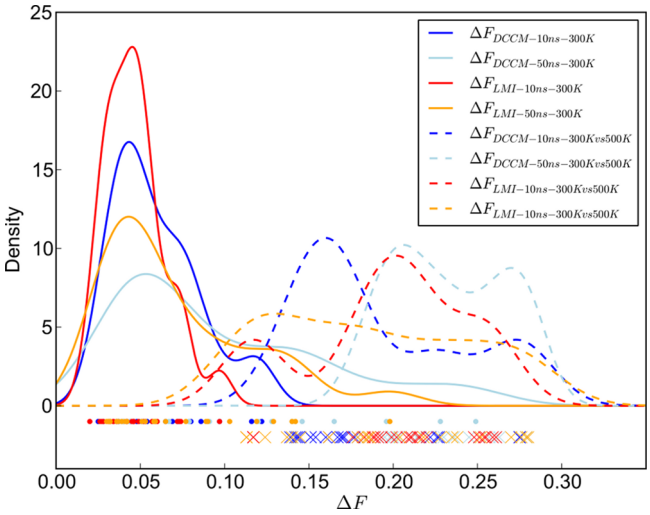


Figure 4. Comparison between the distributions of ΔF values in MD replicates which describe native dynamics or partially unfolded states. All the replicates of each protein target have been compared by ΔF. In particular, we estimated either the ΔF between two halves of each 300 K trajectory (solid lines) or between each 300 K simulation and unfolding trajectory carried out at 500 K (dotted lines). We employed either DCCM or LMI methods and 50 ns or 10 ns time windows for the averaging. We then calculated the density of ΔF data of each data set by mean of the Kernel Density Estimation algorithm, using a Gaussian kernel function and a bandwidth value of 0.4. On the bottom of the plot, we also highlighted the data points with filled dots (intrareplica comparison) or crosses (comparison with the unfolding simulation). We noticed that a clear distinction is observed between ΔF values that represent the protein native dynamics and the ones that are referred to the comparison with the unfolding simulation. The upper limit of ΔF of ∼0.12 observed for the ARID domain is here confirmed by the inspection of a larger number of simulations.

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

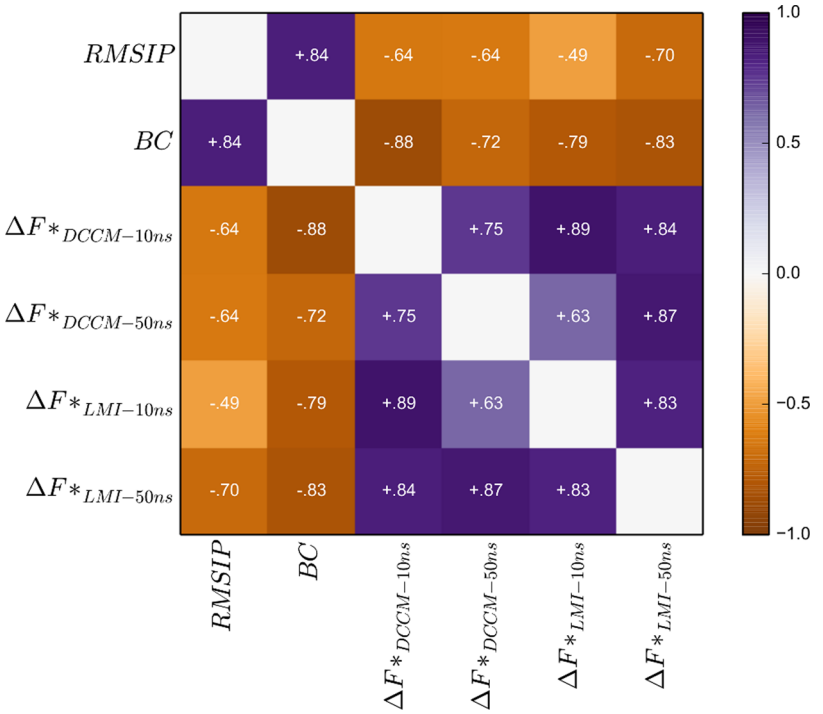


Figure 5. Comparison between RMSIP, BC, and ΔF. We employed RMSIP, BC, ΔF-LMI, and ΔF-DCCM for intrareplicate comparisons, i.e. estimating the overlap between the two halves of each replicate of the six target proteins of this study. For this analysis ΔF has been truncated at the upper limit 0.12 (i.e., all the values greater than 0.12 have been set equal to 0.12). The truncated ΔF values have been then rescaled between 0 and 1, calculating the ΔF^* shown in the figure.

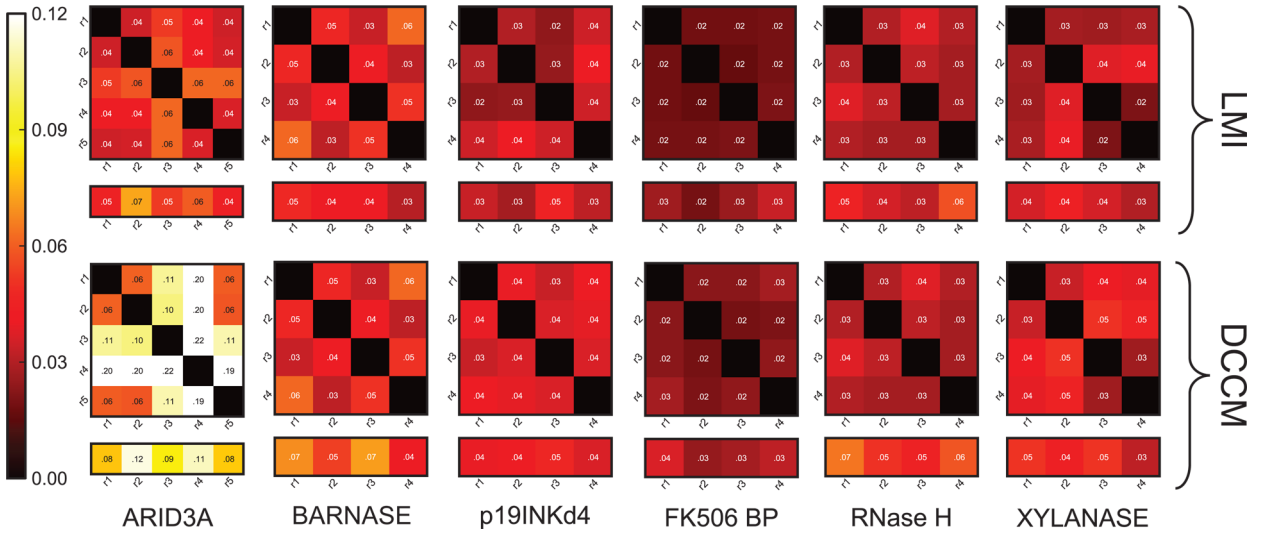


Figure 6. ΔF to compare correlated motions of independent MD replicates of the same system. We show the pairwise comparisons between four or five independent 100 ns MD replicates of each of the six protein targets using a 10 ns time window (i.e., 10-fold shorter than the whole MD trajectory) with either the LMI and DCCM methods. The results are represented in the form of heat maps, and each pairwise ΔF value is explicitly indicated in the plot. We also reported the ΔF between the two halves of each replicate (intrareplica) as a reference of inherent variability for each trajectory in the box on the bottom of of each panel.

described in a robust way. Overall, our analyses point out the need of more than one index to evaluate and compare MD simulations, in agreement with other studies where other methods to compare MD ensembles have been proposed.[9,82−85,66,86] In particular, we noticed the importance of developing quantitative metrics for the comparison of correlated motions when they are the main focus of the postprocessing of MD simulations.

We then assessed the robustness of DCCM and LMI approaches in describing correlated motions across different independent replicates of the same system. In Figure 6 we reported the pairwise ΔF inter-replica comparisons i.e., between the different 100 ns MD replicates of each protein, calculated using a time window of 10 ns (i.e., 10-fold shorter than the whole trajectory) with both the LMI and DCCM methods. We also reported the ΔF between two halves of each replicate (intrareplicate comparisons, discussed above in Figure

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

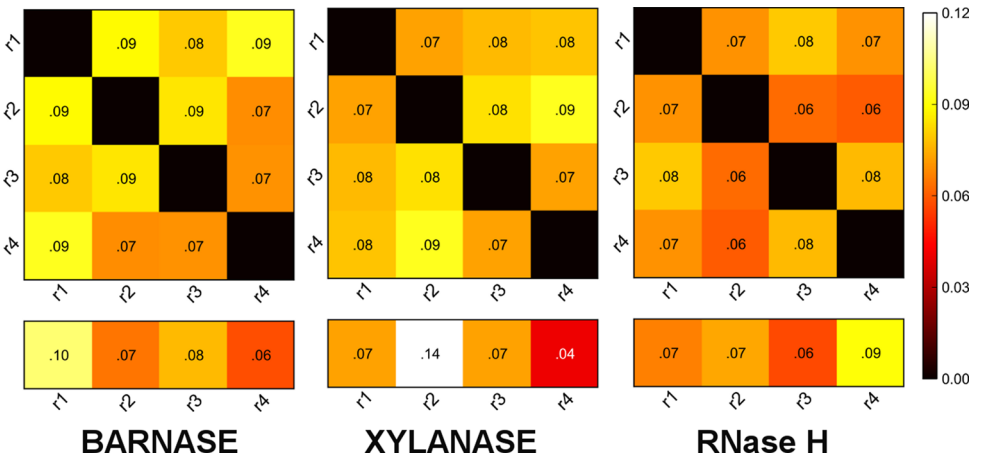


Figure 7. ΔF to compare correlated motions of independent MD replicates of the same system without averaging over a 10 ns time window. We show the results of the pairwise comparisons between four independent 100 ns MD replicates of three protein targets with either the LMI and DCCM methods, using heat maps. Each ΔF value is explicitly indicated in the plot. In this analysis, we employed only an average correlation map estimated from the whole trajectory without averaging over shorter time windows.

4) as a reference of inherent variability within each trajectory. We noticed that all the ΔF values are remarkably below the upper limit of 0.12 and generally comparable or lower than the intrareplicate values. The only exception is ARID3A, which is the example that we here report for an unstable MD simulation (Figure S1). ARID domains are indeed characterized by long unstructured loops, of which two loops in particular are the regions responsible for DNA interaction and are highly flexible in the free protein in solution. On the contrary, if no averaging over small time windows is employed, we failed at identifying a robust and consistent description of correlated motions in independent replicates of the same system (Figure 7), also for three proteins characterized by a fairly stable main-chain RMSD profile (Figure S1). In these cases, the ΔF values are indeed very close to the upper limit of 0.12 (Figure 7).

CONCLUSIONS

We here applied an index, which is based on a normalized Frobenius norm of difference matrix, to quantitatively compare correlation maps estimated with different metrics from protein conformational ensembles, such as the one derived by MD simulations. We have illustrated the applications for this index not only to quantify the similarity or dissimilarity between correlated motions but also more in general to complement other methods to compare MD ensembles. We also encouraged the usage of similar analyses to evaluate the convergence and significance of the average correlation map used to analyze MD data. Our analyses suggest that with microsecond simulations in hand, only correlated motions occurring on 10-fold shorter time scales can be estimated with a reasonable confidence, confirming results previously achieved on shorter time scales.[14] We have then compared different replicates of the same system simulated using the same force field. Our results suggest that the description of LMI and DCCM is robust across multiple replicates when the correlation maps are averaged over at least 10-fold shorter time scales with respect to the whole simulation length, at least with the force field here employed. We also noticed that all the replicates of the same system do not necessarily provide an identical description of the correlation maps, since we observed some exceptions even when short time windows are used for the averaging. This is especially true for highly dynamic proteins, such as the ARID domains here investigated. ΔF analysis can be extended to compare different force-field simulations or even different variants of the same protein, such as wild type and mutant variants of a protein to discriminate between mutations that have an effect on correlated motions and neutral mutations. Moreover, we should notice that we here illustrated the usage of ΔF to compare correlated motions across the whole protein structure. The same approaches can, in principle, be applied to a subset of residues to provide local or residue-specific comparisons. The usage of this index is not limited to the comparison of DCCM- or MI-derived correlation maps, but it can be more broadly employed for a quantitative comparison of any metrics that can be represented as a list of quantitative pairwise relationship between protein residues.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00512.

Figures S1 and S2 (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: elena.papaleo@unimib.it, elenap@cancer.dk.

Present Address

[†]Unit of Statistics, Bioinformatics and Registry, Danish Cancer Society Research Center, Strandboulevarden 49, 2100, Copenhagen, Denmark.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by HPC Caspur Grants 2011, ISCRA-Cineca Grant 2013 (HP10CS5VQ7) and 2014 (HP10CN2H3L).

REFERENCES

(1) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120−127.

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(2) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450*, 964−972.
(3) Karplus, M.; Kuriyan, J. Molecular Dynamics and Protein Function. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6679−6685.
(4) Ramanathan, A.; Savol, A.; Burger, V.; Chennubhotla, C. S.; Agarwal, P. K. Protein Conformational Populations and Functionally Relevant Substates. *Acc. Chem. Res.* **2014**, *47*, 149−156.
(5) Kern, D.; Zuiderweg, E. R. P. The Role of Dynamics in Allosteric Regulation. *Curr. Opin. Struct. Biol.* **2003**, *13*, 748−757.
(6) Manley, G.; Loria, J. P. NMR Insights into Protein Allostery. *Arch. Biochem. Biophys.* **2012**, *519*, 223−231.
(7) Popovych, N.; Sun, S.; Ebright, R. H.; Kalodimos, C. G. Dynamically Driven Protein Allostery. *Nat. Struct. Mol. Biol.* **2006**, *13*, 831−838.
(8) Long, D.; Brüschweiler, R. Structural and Entropic Allosteric Signal Transduction Strength via Correlated Motions. *J. Phys. Chem. Lett.* **2012**, *3*, 1722−1726.
(9) Fuglebakk, E.; Echave, J.; Reuter, N. Measuring and Comparing Structural Fluctuation Patterns in Large Protein Datasets. *Bioinformatics* **2012**, *28*, 2431−2440.
(10) Dodson, G. G.; Lane, D. P.; Verma, C. S. Molecular Simulations of Protein Dynamics: New Windows on Mechanisms in Biology. *EMBO Rep.* **2008**, *9*, 144−150.
(11) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429−452.
(12) 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.
(13) Berendsen, H. J.; Hayward, S. Collective Protein Dynamics in Relation to Function. *Curr. Opin. Struct. Biol.* **2000**, *10*, 165−169.
(14) Hünenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. Fluctuation and Cross-Correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J. Mol. Biol.* **1995**, *252*, 492−503.
(15) Lange, O. F.; Grubmüller, H. Generalized Correlation for Biomolecular Dynamics. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 1053−1061.
(16) Rod, T. H.; Radkiewicz, J. L.; Brooks, C. L. Correlated Motion and the Effect of Distal Mutations in Dihydrofolate Reductase. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 6980−6985.
(17) Doruker, P.; Atilgan, A. R.; Bahar, I. Dynamics of Proteins Predicted by Molecular Dynamics Simulations and Analytical Approaches: Application to Alpha-Amylase Inhibitor. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 512−524.
(18) Gohlke, H.; Kuhn, L. A.; Case, D. A. Change in Protein Flexibility upon Complex Formation: Analysis of Ras-Raf Using Molecular Dynamics and a Molecular Framework Approach. *Proteins: Struct., Funct., Genet.* **2004**, *56*, 322−337.
(19) Morra, G.; Verkhivker, G.; Colombo, G. Modeling Signal Propagation Mechanisms and Ligand-Based Conformational Dynamics of the Hsp90 Molecular Chaperone Full-Length Dimer. *PLoS Comput. Biol.* **2009**, *5*, e1000323.
(20) Grant, B. J.; Gorfe, A. A.; McCammon, J. A. Ras Conformational Switching: Simulating Nucleotide-Dependent Conformational Transitions with Accelerated Molecular Dynamics. *PLoS Comput. Biol.* **2009**, *5*, e1000325.
(21) Lukman, S.; Lane, D. P.; Verma, C. S. Mapping the Structural and Dynamical Features of Multiple p53 DNA Binding Domains: Insights into Loop 1 Intrinsic Dynamics. *PLoS One* **2013**, *8*, e80221.
(22) Papaleo, E.; Pasi, M.; Tiberti, M.; De Gioia, L. Molecular Dynamics of Mesophilic-like Mutants of a Cold-Adapted Enzyme: Insights into Distal Effects Induced by the Mutations. *PLoS One* **2011**, *6*, e24214.
(23) Papaleo, E.; Renzetti, G. Coupled Motions during Dynamics Reveal a Tunnel toward the Active Site Regulated by the N-Terminal A-Helix in an Acylaminoacyl Peptidase. *J. Mol. Graphics Modell.* **2012**, *38*, 226−234.
(24) Wu, S.; Acevedo, J. P.; Reetz, M. T. Induced Allostery in the Directed Evolution of an Enantioselective Baeyer-Villiger Monooxygenase. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 2775−2780.
(25) Masetti, M.; Falchi, F.; Recanatini, M. Protein Dynamics of the HIF-2α PAS-B Domain upon Heterodimerization and Ligand Binding. *PLoS One* **2014**, *9*, e94986.
(26) Liu, J.; Nussinov, R. Flexible Cullins in Cullin-RING E3 Ligases Allosterically Regulate Ubiquitination. *J. Biol. Chem.* **2011**, *286*, 40934−40942.
(27) Liu, J.; Nussinov, R. Rbx1 Flexible Linker Facilitates Cullin-RING Ligase Function before Neddylation and after Deneddylation. *Biophys. J.* **2010**, *99*, 736−744.
(28) Karaca, E.; Tozluoğlu, M.; Nussinov, R.; Haliloğlu, T. Alternative Allosteric Mechanisms Can Regulate the Substrate and E2 in SUMO Conjugation. *J. Mol. Biol.* **2011**, *406*, 620−630.
(29) Aykaç Fas, B.; Tutar, Y.; Haliloğlu, T. Dynamic Fluctuations Provide the Basis of a Conformational Switch Mechanism in Apo Cyclic AMP Receptor Protein. *PLoS Comput. Biol.* **2013**, *9*, e1003141.
(30) Ichiye, T.; Karplus, M. Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 205−217.
(31) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *J. Chem. Theory Comput.* **2009**, *5*, 2486−2502.
(32) Brandman, R.; Lampe, J. N.; Brandman, Y.; De Montellano, P. R. O. Active-Site Residues Move Independently from the Rest of the Protein in a 200 Ns Molecular Dynamics Simulation of Cytochrome P450 CYP119. *Arch. Biochem. Biophys.* **2011**, *509*, 127−132.
(33) Brandman, R.; Brandman, Y.; Pande, V. S. Sequence Coevolution between RNA and Protein Characterized by Mutual Information between Residue Triplets. *PLoS One* **2012**, *7*, e30022.
(34) Fataftah, H.; Karain, W. Detecting Protein Atom Correlations Using Correlation of Probability of Recurrence. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 2180−2189.
(35) Seeber, M.; Felline, A.; Raimondi, F.; Muff, S.; Friedman, R.; Rao, F.; Caflisch, A.; Fanelli, F. Wordom: A User-Friendly Program for the Analysis of Molecular Structures, Trajectories, and Free Energy Surfaces. *J. Comput. Chem.* **2011**, *32*, 1183−1194.
(36) Angelova, K.; Felline, A.; Lee, M.; Patel, M.; Puett, D.; Fanelli, F. Conserved Amino Acids Participate in the Structure Networks Deputed to Intramolecular Communication in the Lutropin Receptor. *Cell. Mol. Life Sci.* **2011**, *68*, 1227−1239.
(37) Ghosh, A.; Vishveshwara, S. A Study of Communication Pathways in Methionyl- tRNA Synthetase by Molecular Dynamics Simulations and Structure Network Analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 15711−15716.
(38) Ghosh, A.; Sakaguchi, R.; Liu, C.; Vishveshwara, S.; Hou, Y.-M. Allosteric Communication in Cysteinyl tRNA Synthetase: A Network of Direct and Indirect Readout. *J. Biol. Chem.* **2011**, *286*, 37721−37731.
(39) Papaleo, E.; Lindorff-Larsen, K.; De Gioia, L. Paths of Long-Range Communication in the E2 Enzymes of Family 3: A Molecular Dynamics Investigation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12515−12525.
(40) Mariani, S.; Dell'Orco, D.; Felline, A.; Raimondi, F.; Fanelli, F. Network and Atomistic Simulations Unveil the Structural Determinants of Mutations Linked to Retinal Diseases. *PLoS Comput. Biol.* **2013**, *9*, e1003207.
(41) Papaleo, E.; Renzetti, G.; Tiberti, M. Mechanisms of Intramolecular Communication in a Hyperthermophilic Acylaminoacyl Peptidase: A Molecular Dynamics Investigation. *PLoS One* **2012**, *7*, e35686.
(42) Invernizzi, G.; Tiberti, M.; Lambrughi, M.; Lindorff-Larsen, K.; Papaleo, E. Communication Routes in ARID Domains between Distal Residues in Helix 5 and the DNA-Binding Loops. *PLoS Comput. Biol.* **2014**, *10*, e1003744.

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(43) Papaleo, E.; Renzetti, G.; Invernizzi, G.; Asgeirsson, B. Dynamics Fingerprint and Inherent Asymmetric Flexibility of a Cold-Adapted Homodimeric Enzyme. A Case Study of the *Vibrio* Alkaline Phosphatase. *Biochim. Biophys. Acta, Gen. Subj.* **2013**, *1830*, 2970−2980.
(44) Feher, V. A.; Durrant, J. D.; Van Wart, A. T.; Amaro, R. E. Computational Approaches to Mapping Allosteric Pathways. *Curr. Opin. Struct. Biol.* **2014**, *25*, 98−103.
(45) Vanwart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput.* **2012**, *8*, 2949−2961.
(46) Amaro, R. E.; Sethi, A.; Myers, R. S.; Davisson, V. J.; Luthey-schulten, Z. A. A Network of Conserved Interactions Regulates the Allosteric Signal in a Glutamine Amidotransferase. *Biochemistry* **2007**, *46*, 2156−2173.
(47) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical Networks in tRNA:protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6620−6625.
(48) Blacklock, K.; Verkhivker, G. M. Differential Modulation of Functional Dynamics and Allosteric Interactions in the Hsp90-Cochaperone Complexes with p23 and Aha1: A Computational Study. *PLoS One* **2013**, *8*, e71936.
(49) Blacklock, K.; Verkhivker, G. M. Computational Modeling of Allosteric Regulation in the Hsp90 Chaperones: A Statistical Ensemble Analysis of Protein Structure Networks and Allosteric Communications. *PLoS Comput. Biol.* **2014**, *10*, e1003679.
(50) LeVine, M. V.; Weinstein, H. NbIT−a New Information Theory-Based Analysis of Allosteric Mechanisms Reveals Residues That Underlie Function in the Leucine Transporter LeuT. *PLoS Comput. Biol.* **2014**, *10*, e1003603.
(51) Armenta-Medina, D.; Pérez-Rueda, E.; Segovia, L. Identification of Functional Motions in the Adenylate Kinase (ADK) Protein Family by Computational Hybrid Approaches. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 1662−1671.
(52) Estabrook, R. A.; Luo, J.; Purdy, M. M.; Sharma, V.; Weakliem, P.; Bruice, T. C.; Reich, N. O. Statistical Coevolution Analysis and Molecular Dynamics: Identification of Amino Acid Pairs Essential for Catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 994−999.
(53) Silvestre-Ryan, J.; Lin, Y.; Chu, J.-W. Fluctuograms" Reveal the Intermittent Intra-Protein Communication in Subtilisin Carlsberg and Correlate Mechanical Coupling with Co-Evolution. *PLoS Comput. Biol.* **2011**, *7*, e1002023.
(54) Strafford, J.; Payongsri, P.; Hibbert, E. G.; Morris, P.; Batth, S. S.; Steadman, D.; Smith, M. E. B.; Ward, J. M.; Hailes, H. C.; Dalby, P. A. Directed Evolution to Re-Adapt a Co-Evolved Network within an Enzyme. *J. Biotechnol.* **2012**, *157*, 237−245.
(55) Roy, A.; Post, C. B. Detection of Long-Range Concerted Motions in Protein by a Distance Covariance. *J. Chem. Theory Comput.* **2012**, *8*, 3009−3014.
(56) Kamberaj, H.; van der Vaart, A. Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations. *Biophys. J.* **2009**, *97*, 1747−1755.
(57) Ribeiro, A.; Ortiz, V. Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology. *J. Chem. Theory Comput.* **2014**, *10*, 1762−1769.
(58) Hess, B. Convergence of Sampling in Protein Simulations. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2002**, *65*, 031910.
(59) Papaleo, E.; Mereghetti, P.; Fantucci, P.; Grandori, R.; De Gioia, L. Free-Energy Landscape, Principal Component Analysis, and Structural Clustering to Identify Representative Conformations from Molecular Dynamics Simulations: The Myoglobin Case. *J. Mol. Graphics Modell.* **2009**, *27*, 889−899.
(60) 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., Genet.* **1999**, *36*, 419−424.
(61) Fuglebakk, E.; Tiwari, S. P.; Reuter, N. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850*, 911−922.
(62) Tiwari, S. P.; Fuglebakk, E.; Hollup, S. M.; Skjærven, L.; Cragnolini, T.; Grindhaug, S. H.; Tekle, K. M.; Reuter, N. WEBnm@ v2.0: Web Server and Services for Comparing Protein Flexibility. *BMC Bioinf.* **2014**, *15*, 427.
(63) Fuglebakk, E.; Reuter, N.; Hinsen, K. Evaluation of Protein Elastic Network Models Based on an Analysis of Collective Motions. *J. Chem. Theory Comput.* **2013**, *9*, 5618−5628.
(64) Perica, T.; Kondo, Y.; Tiwari, S. P.; McLaughlin, S. H.; Kemplen, K. R.; Zhang, X.; Steward, A.; Reuter, N.; Clarke, J.; Teichmann, S. A. Evolution of Oligomeric State through Allosteric Pathways That Mimic Ligand Binding. *Science* **2014**, *346*, 1254346.
(65) Skjærven, L.; Yao, X.-Q.; Scarabelli, G.; Grant, B. J. Integrating Protein Structural Dynamics and Evolutionary Analysis with Bio3D. *BMC Bioinf.* **2014**, *15*, 399.
(66) Yang, S.; Salmon, L.; Al-Hashimi, H. M. Measuring Similarity between Dynamic Ensembles of Biomolecules. *Nat. Methods* **2014**, *11*, 552−554.
(67) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100*, L47−L49.
(68) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926.
(69) Liu, G.; Huang, Y. J.; Xiao, R.; Wang, D.; Acton, T. B.; Montelione, G. T. Solution NMR Structure of the ARID Domain of Human AT-Rich Interactive Domain-Containing Protein 3A: A Human Cancer Protein Interaction Network Target. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 2170−2175.
(70) Iwahara, J.; Clubb, R. T. Solution Structure of the DNA Binding Domain from Dead Ringer, a Sequence-Specific AT-Rich Interaction Domain (ARID). *EMBO J.* **1999**, *18*, 6084−6094.
(71) Martin, C.; Richard, V.; Salem, M.; Hartley, R.; Mauguen, Y. Refinement and Structural Analysis of Barnase at 1.5 A Resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 386−398.
(72) Baumgartner, R.; Fernandez-Catalan, C.; Winoto, A.; Huber, R.; Engh, R. A.; Holak, T. A. Structure of Human Cyclin-Dependent Kinase Inhibitor p19INK4d: Comparison to Known Ankyrin-Repeat-Containing Structures and Implications for the Dysfunction of Tumor Suppressor p16INK4a. *Structure* **1998**, *6*, 1279−1290.
(73) Itoh, S.; DeCenzo, M. T.; Livingston, D. J.; Pearlman, D. A.; Navia, M. A. Conformation of FK506 in X-Ray Structures of Its Complexes with Human Recombinant FKBP12 Mutants. *Bioorg. Med. Chem. Lett.* **1995**, *5*, 1983−1988.
(74) Katayanagi, K.; Miyagawa, M.; Matsushima, M.; Ishikawa, M.; Kanaya, S.; Nakamura, H.; Ikehara, M.; Matsuzaki, T.; Morikawa, K. Structural Details of Ribonuclease H from *Escherichia coli* as Refined to an Atomic Resolution. *J. Mol. Biol.* **1992**, *223*, 1029−1052.
(75) Brinda, K. V.; Vishveshwara, S. A Network Representation of Protein Structures: Implications for Protein Stability. *Biophys. J.* **2005**, *89*, 4159−4170.
(76) Pasi, M.; Tiberti, M.; Arrigoni, A.; Papaleo, E. xPyder: A PyMOL Plugin To Analyze Coupled Residues and Their Networks in Protein Structures. *J. Chem. Inf. Model.* **2012**, *52*, 1865−1874.
(77) Tiberti, M.; Invernizzi, G.; Lambrughi, M.; Inbar, Y.; Schreiber, G.; Papaleo, E. PyInteraph: A Framework for the Analysis of Interaction Networks in Structural Ensembles of Proteins. *J. Chem. Inf. Model.* **2014**, *54*, 1537−1551.
(78) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential Dynamics of Proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412−425.
(79) 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.
(80) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* **2012**, *7*, e32131.
(81) 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.

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(82) Lindorff-Larsen, K.; Ferkinghoff-Borg, J. Similarity Measures for Protein Ensembles. *PLoS One* **2009**, *4*, e4203.
(83) Fraccalvieri, D.; Pandini, A.; Stella, F.; Bonati, L. Conformational and Functional Analysis of Molecular Dynamics Trajectories by Self-Organising Maps. *BMC Bioinf.* **2011**, *12*, 158.
(84) Fraccalvieri, D.; Tiberti, M.; Pandini, A.; Bonati, L.; Papaleo, E. Functional Annotation of the Mesophilic-like Character of Mutants in a Cold-Adapted Enzyme by Self-Organising Map Analysis of Their Molecular Dynamics. *Mol. BioSyst.* **2012**, *8*, 2680−2691.
(85) McClendon, C. L.; Hua, L.; Barreiro, A.; Jacobson, M. P. Comparing Conformational Ensembles Using the Kullback-Leibler Divergence Expansion. *J. Chem. Theory Comput.* **2012**, *8*, 2115−2126.
(86) Kruschel, D.; Zagrovic, B. Conformational Averaging in Structural Biology: Issues, Challenges and Computational Solutions. *Mol. BioSyst.* **2009**, *5*, 1606−1616.

DOI: 10.1021/acs.jctc.5b00512
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX