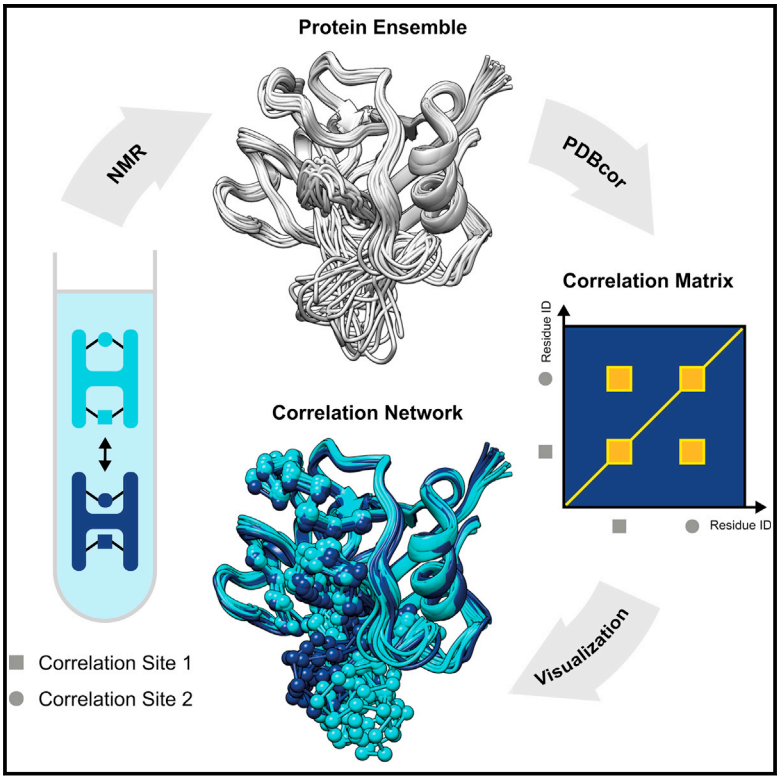
Structure

Resource

PDBcor: An automated correlation extraction calculator for multi-state protein structures

Graphical abstract



Highlights

- PDBcor algorithm extracts protein-correlated motion from protein ensembles
- PDBcor is based on GMM clustering and information theory
- High sensitivity to correlated motion comes from the use of protein distances
- PDBcor is unbiased, as the structure superposition step is not required

Authors

Dzmitry Ashkinadze, Piotr Klukowski, Harindranath Kadavath, Peter Güntert, Roland Riek

Correspondence

roland.riek@phys.chem.ethz.ch (R.R.), peter.guentert@phys.chem.ethz.ch (P.G.)

In brief

Ashkinadze et al. present an unbiased algorithm, PDBcor, for the extraction of protein-correlated motion from protein structural ensembles. Using clustering and mutual information, this algorithm is based on the statistical analysis of protein interresidual distances. The authors validate it on three model proteins with known structural correlations

Ashkinadze et al., 2022, Structure *30*, 646–652
April 7, 2022
https://doi.org/10.1016/j.str.2021.12.002

CellPress

CellPress

Structure

Resource

PDBcor: An automated correlation extraction calculator for multi-state protein structures

Dzmitry Ashkinadze,[1] Piotr Klukowski,[1] Harindranath Kadavath,[1] Peter Güntert,[1,2,3,*] and Roland Riek[1,4,*]
[1]Laboratory of Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland
[2]Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, 60438 Frankfurt am Main, Germany
[3]Department of Chemistry, Tokyo Metropolitan University, Hachioji, Tokyo 1920397, Japan
[4]Lead contact
*Correspondence: roland.riek@phys.chem.ethz.ch (R.R.), peter.guentert@phys.chem.ethz.ch (P.G.)
https://doi.org/10.1016/j.str.2021.12.002

SUMMARY

Allostery and correlated motion are key elements linking protein dynamics with the mechanisms of action of proteins. Here, we present PDBCor, an automated and unbiased method for the detection and analysis of correlated motions from experimental multi-state protein structures. It uses torsion angle and distance statistics and does not require any structure superposition. Clustering of protein conformers allows us to extract correlations in the form of mutual information based on information theory. With PDBcor, we elucidated correlated motion in the WW domain of PIN1, the protein GB3, and the enzyme cyclophilin, in line with reported findings. Correlations extracted with PDBcor can be utilized in subsequent assays including nuclear magnetic resonance (NMR) multi-state structure optimization and validation. As a guide for the interpretation of PDBcor results, we provide a series of protein structure ensembles that exhibit different levels of correlation, including non-correlated, locally correlated, and globally correlated ensembles.

INTRODUCTION

Protein dynamics is key for understanding enzymatic activity, protein-protein interactions, target recognition, ligand binding, and signaling (Ishima and Torchia, 2000). A particularly complex example is a ligand-induced correlated motion of two distant sites, termed allostery. Several mechanisms for such motions have been proposed including the population shift model (Monnot et al., 1996) and the dynamic allostery model (Cooper and Dryden, 1984). The population shift model is based on ligand-induced structural rearrangements between two distinct protein conformations. The dynamic allostery model is based on a statistical thermodynamics model able to quantify allosteric communication in the absence of a conformational change by investigating the effect of ligand-binding on thermal fluctuations within a protein.

In order to elucidate motion, including the correlated motion of a protein at atomic resolution, multi-state protein structures are determined by experimental methods including NMR using a plethora of experimental restraints (Clore et al., 1999; Orts et al., 2012; Palmer, 2004; Riek et al., 1999), by different class selections in cryoelectron microscopy (cryo-EM)-derived structure determination (Banerjee et al., 2016), or by the presence of distinct X-ray structures due to different crystal packings or the same crystals exposed to a strong electric field (Hekstra et al., 2016). Alternatively, such protein ensemble structures could be generated with molecular dynamics (MD) canonical ensemble simulations in the presence or absence of experimental data (Bouvignies et al., 2005; Hummer et al., 2004; Nosé, 1984). Conventionally, correlated motion is extracted in the form of residue-based cross-correlation matrices from MD trajectories (La Sala et al., 2017; Long and Brüschweiler, 2011; McClendon et al., 2009; Zhang et al., 2021) or, alternatively, from the superimposed structural ensembles either with principal-component analysis (PCA)-(Theobald and Wuttke, 2006; Zhang et al., 2021) or normal mode analysis (NMA)-based (Tiwari et al., 2014) approaches.

In this work, we present an alternative, highly sensitive method for the correlation extraction from structural ensembles that does not require any structure superposition and therefore is unbiased due to the fact that it is based solely on distance and angle statistics of individual structural entities. PDBcor performs an objective and automated correlation analysis of multi-state protein structures, which can be used for the elucidation of biologically important correlated motion. With the help of information theory, it is possible to extract residue-based protein correlations in a fully automated fashion. Information about such biologically relevant correlations is vital for our understanding of proteins. PDBcor is publicly available as a Python executable (https://github.com/dzmitryashkinadze/PDBcor) or as a server (https://www.pdbcor.ethz.ch/).

RESULTS

Theory

The workflow of the correlation extraction procedure with PDBcor is shown in Figure 1. First, an input structure bundle is subjected to significance thresholding that filters out spurious

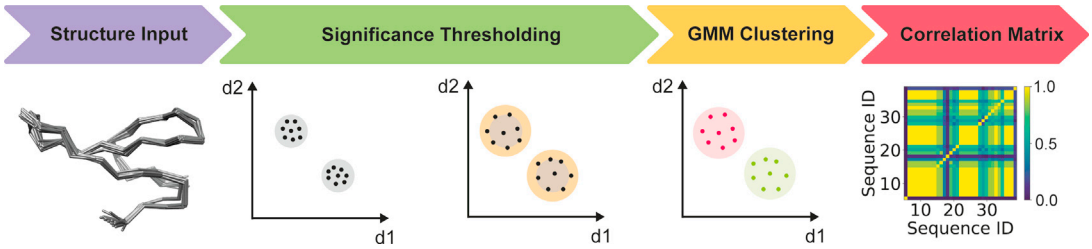


Figure 1. Overview of the correlation extraction procedure

First, an input structure bundle (PDB: 6SVC; Strotz et al., 2020) is subjected to a significance thresholding that filters out spurious insignificant correlations. Here, an illustrative example is depicted, where conformers existing in two states are shown as points in a scatter plot of two arbitrary distances (for example, the first is a distance between residues X and Y, and the second is a distance between residues X and Z). During significance thresholding, the random displacement of atoms broadens the edges of states so that states are separated by less than the amplitude of the noise loose separation. Then, interresidual distances are used to cluster conformers for each residue with GMM (in this case, it would be residue X). Finally, a pairwise comparison of the resulting clustering vectors based on their mutual information yields an interpretable correlation matrix with a scalebar.

small-amplitude correlations. Second, interresidual distances are used to cluster conformers. Finally, residue clusterings are compared to obtain a correlation matrix.

Objective extraction of correlated motion

PDBcor relies on a structure comparison based on a statistical analysis of interresidual distances or dihedral angles within individual conformers that does not require any superpositions. Conventionally a superimposed ensemble of protein conformations is visually sorted based on certain local protein features. For example, if protein conformers are sorted according to the relative position of a particular α-helix, neighboring regions might be sorted correctly and therefore correlate to the α-helix, but such sorting is typically not coherent throughout the whole protein scaffold (Privalov, 1989). In order to systematically study those correlations, an ensemble of multi-state protein conformations is repeatedly clustered for each residue with the aim to extract correlations between protein residues. Residue correlations are evaluated by computing a similarity between two arbitrary conformer clusterings.

Significance thresholding

Correlations extracted with PDBCor are based exclusively on the similarity between residue clusterings (see below). As such, they are largely independent of the degree of separation between states. In some well-defined structural bundles, individual states might therefore be identified that are closer to each other than the amplitudes of random thermal motion. This might lead to spurious distance correlations. To avoid such artifacts, a small amount of Gaussian noise is added to the atomic coordinates:

$$r'^{(j)}_{im} = r^{(j)}_{im} + \delta^{(j)}_{im}, \qquad \text{(Equation 1)}$$

where $r^{(j)}_{im}$ is the position of atom m in residue i of conformer j, which is obtained with Biopython (Cock et al., 2009), and $\delta^{(j)}_{im}$ is a vector of three independent, normally distributed random numbers with zero mean and standard deviation σ. This leads to the random mixing of insignificantly separated protein states and suppression of spurious distance correlations.

The noise amplitude σ should be set such that it is sufficient to remove background correlations with amplitudes below that of thermal motions and experimental uncertainties but does not exceed the separation between significantly different protein states that would remove correlations of interest. A standard value of 0.5 Å was used for all presented experiments as a value that resembles the fast (ps) order parameter of 0.8 that has been measured in proteins by NMR (Kay et al., 1989). However, PDBcor allows also to switch off the noise generator completely.

Residue-based conformer clustering

For the purpose of clustering, each residue i is represented by a single point, given by its centroid coordinates in conformer j:

$$x^{(j)}_i = \frac{1}{M_i}\sum_{m=1}^{M_i} r'^{(j)}_{im}, \qquad \text{(Equation 2)}$$

where M_i is the number of atoms of residue i that are considered for the correlation calculation. The scope of input atoms can be predefined to be either the backbone atoms, the sidechain atoms, or all atoms of the residue (see below). From the centroid coordinates, we construct a distance matrix D with elements

$$D^{(j)}_{ik} = \left| x^{(j)}_i - x^{(j)}_k \right|. \qquad \text{(Equation 3)}$$

Each row of the distance matrix contains the distances between the center of a given residue i and the centers of the other residues k and thus defines the relative location of the residue that can be used as a fingerprint of a given conformer. In the case of N distinct residue-based protein conformations, we expect that interresidual distances of all conformers from a given structure ensemble can be grouped into N clusters. Using this assumption, conformers are clustered based on their interresidual distances into N groups for each residue using the Gaussian mixture model (GMM) algorithms (Reynolds, 2009). This yields, for each residue i, a distance clustering vector, c_i, with elements $c_{ij} \in \{1, \ldots, N\}$ that stores the cluster labels of all conformers $j = 1, \ldots, N$. The total set of protein interresidual distances that is used as input to the PDBcor is highly redundant, as the number of distances is proportional to the number of residues squared. However, conformers are clustered independently for each residue, and for a selected residue, a non-redundant set of distances from the selected residue to the rest of the protein is used.

As an alternative to distance-based clustering, the clustering can also be based on the backbone φ, ψ, ω and side-chain $\chi_1, \chi_2, \chi_3, \chi_4, \chi_5$ torsion angles. For residues with less than five side-chain torsion angles, the undefined χ values are set to zero. As in the distance case, an angular matrix, $\Phi^{(j)}$, is formed by the eight dihedral angle values of each residue in the conformers $j = 1,\ldots,N$. It is used to cluster conformers into N groups using the GMM. In complete analogy to the distance-based case, this yields, for each residue i, an angular clustering vector, c_i^a, with elements $c_{ij}^a \in \{1,\ldots,N\}$ that stores the cluster labels of all conformers $j = 1,\ldots,N$.

Evaluation of correlated motion

Correlation extraction from the clustering matrix is possible using information theory (Cover and Thomas, 1991; Kullback, 1997; Shannon and Weaver, 1949). Two arbitrary clustering results are represented by two discrete variable vectors, X and Y. One of the most extensively studied measures specifying the amount of correlation between two discrete variable vectors is the mutual information $I(X,Y)$ (Kraskov et al., 2004):

$$I(X,Y) = \sum_{x,y=1}^{N} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \qquad \text{(Equation 4)}$$

where x and y are cluster labels of clusterings X and Y with probabilities $p(x) = p(X = x)$, $p(y) = p(Y = y)$ and joint probability $p(x,y) = p(X = x, Y = y)$. The mutual information tells us how much the conformer clustering of one residue tells us about the conformer clustering of another residue. A variant of the mutual information that was specifically developed for clustering comparison is the adjusted mutual information $I^{\star}(X,Y)$ (Vinh et al., 2010):

$$I^{\star}(X,Y) = \frac{I(X,Y) - E\{I(X',Y')\}}{max\{H(X),H(Y)\} - E\{I(X',Y')\}}, \qquad \text{(Equation 5)}$$

where $E\{I(X',Y')\}$ is the expected value of the mutual information for an ensemble of random, uncorrelated vectors X' and Y', and $H(X)$ is the entropy of the variable X:

$$H(X) = -\sum_{x} p(x)\log p(x), \qquad \text{(Equation 6)}$$

where $p(x)$ is the probability of cluster x. Note that $I^{\star}(X,Y) = I^{\star}(Y,X)$ is symmetric for any pair of clusterings and $I^{\star}(X,Y) \approx 0$ vanishes approximately between two random clusterings. The adjusted mutual information yields a correctly normalized value measured in bits that is a suitable measure for the correlation between protein residues.

Given a clustering matrix ––, all residue pair combinations are compared using the adjusted mutual information, describing a similarity between residues. The adjusted mutual information scores for residues i and j form a symmetric correlation matrix A with elements $A_{ij} = I^{\star}(c_i, c_j)$ for distance-based clustering or $A_{ij}^a = I^{\star}(c_i^a, c_j^a)$ for torsion-angle-based clustering. A visual inspection of the correlation matrix heatmap (Figure 1) provides information about residues or subdomains that are involved in correlated motion. In addition, the mean value of the elements of the matrix A yields an overall correlation parameter for the structure ensemble.

Both distance and angular correlation analyses are able to detect correlated motion. Nevertheless, distance correlation extraction is more sensitive to the protein motion.

Global conformer clustering

For visualization purposes, it is useful to get an optimal global (rather than residue-specific) clustering of conformers that can be used for highlighting state-specific features in a protein ensemble superposition view. For example, the two sets of clustered conformers within a two-state structure ensemble can then be colored differently, as shown in Figure 2.

To this end, we cluster the conformers according to the clustering c_i of the residue i that has the highest average correlation to the other residues of the protein. Since the protein ensemble superposition is made according to the protein coordinates, the distance correlation matrix A is used to calculate the average residue correlations.

Versatility of PDBcor for backbone and side-chain correlations

The correlation extraction procedure allows us to control the protein region from which correlations are extracted by filtering the input data. In particular, backbone correlations can be extracted by utilizing only backbone atom coordinates and backbone dihedral angles. Similarly, the side-chain or total (backbone and side chain) correlations can be extracted. This possibility might be particular interesting for some experimental methods including NMR, for which the backbone structure is better resolved than side chains. Therefore, extraction of backbone correlations could be beneficial for the resolution and sensitivity of protein correlations.

Spatial correlations in protein structures

Three different protein ensembles from the Protein Data Bank that have been determined by liquid-state NMR act as examples for a non-correlated protein ensemble (Figure 2A [Vanwetswinkel et al., 2003]), a locally correlated protein ensemble (Figure 2B [Sheftic et al., 2012]), and a globally correlated protein ensemble (Figure 2C [Crespo-Flores et al., 2019]). The structure bundles were analyzed by PDBcor with the assumption that an ensemble of structures samples the conformational space of a protein with residue-based, two-state dynamics, regardless of the structure origin.

Distance correlation matrix heatmaps of non-correlated systems do not show any significant correlations (visualized by yellow spots in the heatmap, Figure 2D). Optimally clustered conformers of non-correlated systems are typically non-balanced, with one state dominating the other one. The most probable explanation for the absence of correlations in such structure ensembles is a violation of the two-state model assumption.

As opposed to non-correlated systems, distance correlation matrix heatmaps of locally correlated systems show correlations that are localized to distinct regions of the protein structure. Optimally clustered conformers of locally correlated systems can be visually separated into two states in their corresponding protein correlation sites. Correlation lights up as yellow spots in the heatmap (Figure 2E). This correlation between α-helix 2 (residues 42–51) and α-helix 3 (residues 70–78) can also be seen in the

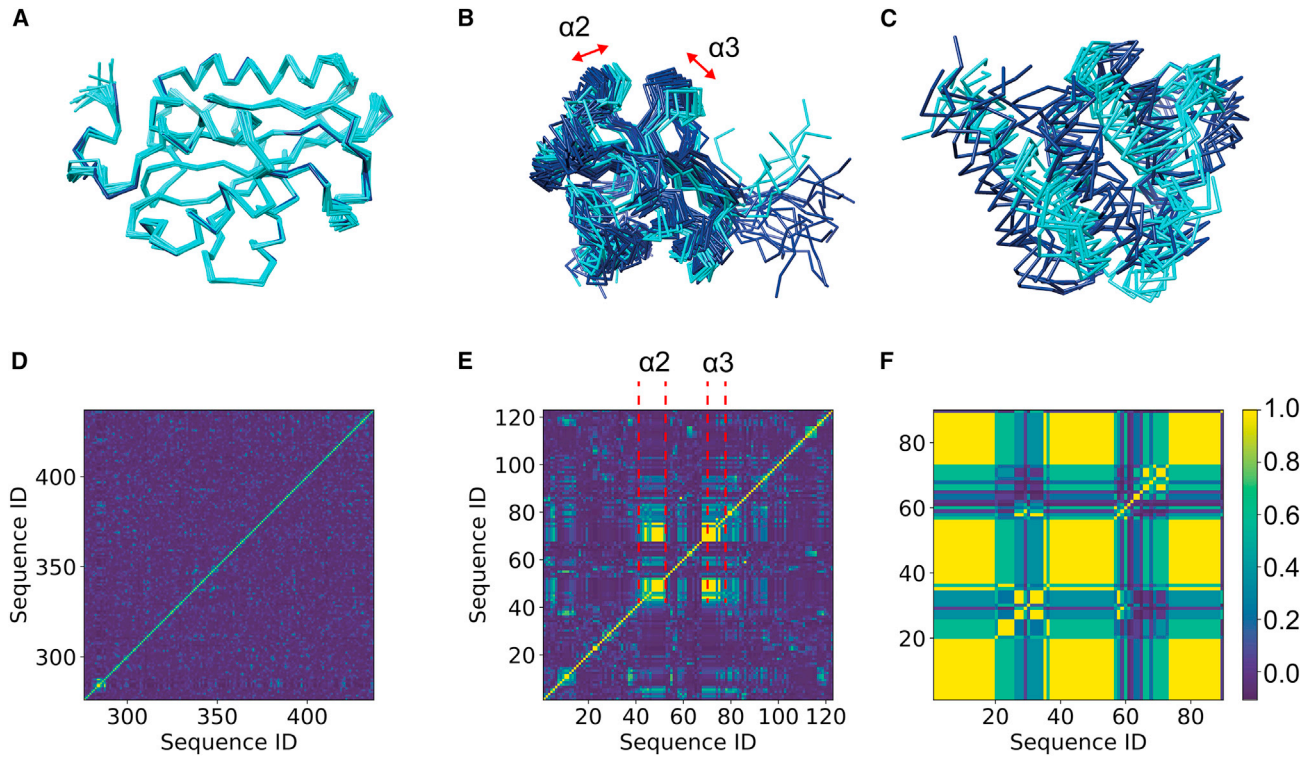


Figure 2. Distance correlation matrix heatmaps with a scalebar and optimally clustered bundles of proteins sorted in ascending order of structural correlations

(A) PDB: 1PBU is depicted as an example of a non-correlated protein system (Vanwetswinkel et al., 2003). The distance correlation matrix heatmap (D) does not show any significant correlations (yellow spots), and a single state (cyan) dominates among the optimally clustered conformers.

(B) PDB: 2LPM is depicted as an example of a locally correlated system (Sheftic et al., 2012). The distance correlation matrix heatmap (E) shows correlations that are localized to α-helices 2 and 3, whereas its optimally clustered conformers correlate also only in the regions of α2 and α3.

(C) PDB: 6P6C is depicted as an example of a globally correlated system (Crespo-Flores et al., 2019). Its distance correlation matrix heatmap (F) is fully correlated, and conformers are unambiguously separable. The conformer separation can be easily visually confirmed due to significant differences between the protein states.

structure superposition and coloring according to the global conformer clustering (Figure 2B).

Conformers from globally correlated protein ensembles can be unambiguously separated. It can be easily visually confirmed, as protein states do not overlap well due to significant differences between the protein states (Figure 2C). Since a global separation does not depend on the choice of the residue, there are pairwise correlations between most residues, and consequently, most of the distance correlation heatmap turns yellow (Figure 2F).

Correlations of WW domain, protein GB3 and cyclophilin

PDBcor was benchmarked on three model systems: the WW domain of PIN1 (Figure 3A; PDB: 6SVC [Strotz et al., 2020]), the protein GB3 (Figure 3B; PDB: 2LUM [Vögeli et al., 2012]), and cyclophilin A (Figure 3C; PDB: 2MZU [Chi et al., 2015]). For all three systems, multi-state structure ensembles were determined by solution-state NMR based on exact nuclear Overhauser effects (NOEs) (Vögeli et al., 2012). The detailed time-intensive study of the multi-state structures using subjective superpositions of conformers and objective angular correlations yielded the presence of correlated motion at atomic resolution in all three systems (Chi et al., 2015; Strotz et al., 2020; Vögeli et al., 2012).

The automated evaluation of the WW domain with PDBcor identifies a globally correlated network (Figure 3D). This shows that experimental restraints were able to separate two WW states.

The automated evaluation of the protein GB3 with PDBcor reveals a system that is (weakly) correlated everywhere except for the α-helix of residues 23–37 (Figure 3E). This finding confirms the previously reported observation of correlated motion across the β-sheet and a lack of correlated motion between the β-sheet and the α-helix (Vögeli et al., 2012). It is noted that the GB3 protein is reported to comprise three states which were successfully analyzed with PDBcor, as it generalizes to an arbitrary number of conformational states.

As an example of a larger system, the protein cyclophilin A was evaluated. According to the distance correlation matrix heatmap (Figure 3F), five previously reported correlations in regions 1 (residues 9–16), 2 (residues 54–57), 3 (residues 64–78), 4 (residues 101–107), and 5 (residues 118–127) were confirmed (Chi et al., 2015). PDBcor did not only find all reported correlation sites but also found an extension of the correlation system to an additional region in the protein, site 6 (residues 137–155). Notably, sites 2–6 form a fully connected correlation network, whereas

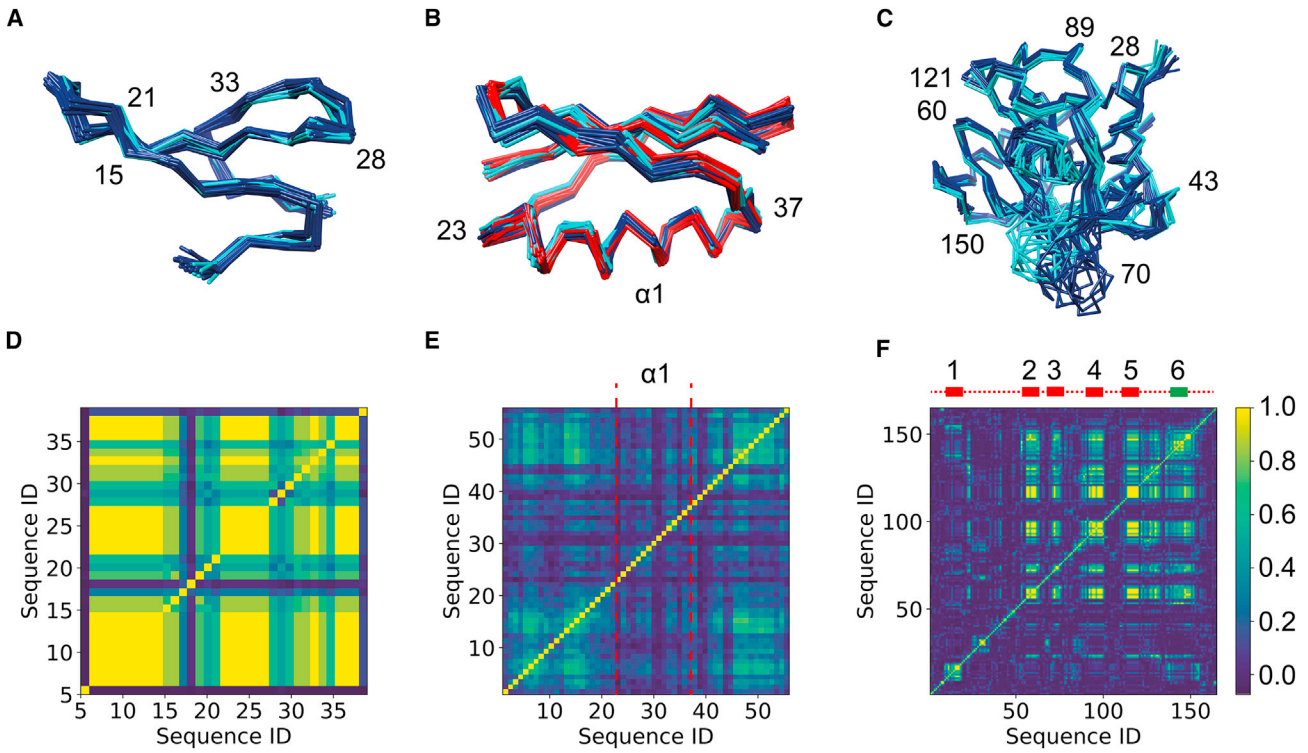


Figure 3. Correlations in multi-state NMR protein structures

(A–F) Automated correlations in multi-state NMR structures for the WW domain of PIN1 (A and D; PDB: 6SVC [Strotz et al., 2020]), protein GB3 (B and E; PDB: 2LUM [Vögeli et al., 2012]), and cyclophilin A (C and F; PDB: 2MZU [Chi et al., 2015]). The top panels (A, B, and C) illustrate the superimposed bundles of conformers and are colored according to the optimal global distance-based clustering. The bottom panels (D, E, and F) illustrate the backbone distance correlation matrix heatmaps with a salebar on the right. For the WW domain, the optimally colored backbone bundle (A) and its distance correlation matrix heatmap (D) both identify a globally correlation network. The distance correlation matrix heatmap of GB3 (E) identifies a system that is weakly correlated everywhere except for the α-helix of residues 23–37, highlighted with a pair of red dashed lines, as it was reported previously (Vögeli et al., 2012). The backbone distance correlation matrix heatmap for cyclophilin (F) confirms seven previously reported correlation sites, including site 1 (residues 9–16), site 2 (residues 54–57), site 3 (residues 64–78), site 4 (residues 101–107), and site 5 (residues 118–127) highlighted in red (Chi et al., 2015). Additionally, PDBcor identifies a previously undetected correlation site 6 (residues 137–155), highlighted in green.

site 1 correlates only to site 6. In the case of cyclophilin A, the strength of PDBcor is apparent: first, it elucidates all statistically significant structural correlations, yielding an extension of the correlation network that had been found manually. Second, in contrast to a tiresome selection by manual inspection, it is fully automated, objective, and reproducible.

DISCUSSION

PDBcor can be used to get an optimal conformer separation for the further analysis of protein states. Alternatively, further interpretation of PDBcor correlation matrices allows us to quantify correlations, identify which part of the protein is involved in correlated motion, and pinpoint the most prominent correlations between protein sites. Careful examination of the correlation matrix may provide information about the localization of correlated subsystems for a given protein.

PDBcor correlation amplitude can be interpreted as an information flow between residue pairs. Therefore, PDBcor is not only able to localize the correlation of interest but also to quantify it. Strong correlation of a residue pair, as in Figure 2F, means that by knowing the state of the first residue, we know the state of the second residue. Weak correlation of a residue pair, as in Figure 3E, means that by knowing the state of the first residue, we can predict with some certainty the state of the second residue.

Any protein structure ensemble can be analyzed with PDBcor. Nevertheless, meaningful correlations can only be extracted from structure bundles that have been generated with the aim to incorporate information about multiple protein states. A cautious use is indicated for proteins with disordered regions. In a limited number of cases, protein-flexible loops account for spurious correlations and should be manually removed from the PDBcor analysis. However, the low resolution of the protein structure caused by the lack of experimental restraints typically does not lead to spurious correlations.

The number of protein ensemble structures grows together with a rapid advancement in the field of structural biology (Levitt, 2007). A fraction of such deposited ensemble structures contains information about correlated motion. The knowledge about such protein correlations is vital for the understanding of protein mechanisms of action and should be systematically studied.

The PDBcor correlation extraction algorithm is sensitive as, unlike PCA and NMA-based correlation extraction algorithms, it relies on interresidual distances and does not require structure

superposition (see Figure S2). It is also versatile, as it can be applied not only to the protein ensembles but also to the MD trajectories (see Figure S1).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Application of PDBcor to MD trajectories
  - Comparison of PDBcor to PCA- and NMA-based methods
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2021.12.002.

ACKNOWLEDGMENTS

We would like to thank the Swiss National Science Foundation (SNF) for financial support.

AUTHOR CONTRIBUTIONS

D.A. developed PDBcor P.K. improved the machine learning and visualization aspects of the PDBcor and implemented the PDBcor server. H.K. proposed to validate PDBcor on previously deposited eNOE structures. P.G. developed significance thresholding. P.G. and R.R. supervised the project. D.A, P.G., and R.R. wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 9, 2021
Revised: October 15, 2021
Accepted: December 1, 2021
Published: December 27, 2021

REFERENCES

Banerjee, S., Bartesaghi, A., Merk, A., Rao, P., Bulfer, S.L., Yan, Y., Green, N., Mroczkowski, B., Neitz, R.J., and Wipf, P. (2016). 2.3 Å resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. Science *351*, 871–875.

Bouvignies, G., Bernado, P., Meier, S., Cho, K., Grzesiek, S., Brüschweiler, R., and Blackledge, M. (2005). Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. Proc. Natl. Acad. Sci. U S A *102*, 13885–13890.

Chi, C.N., Strotz, D., Riek, R., and Vögeli, B. (2015). Extending the eNOE data set of large proteins by evaluation of NOEs with unresolved diagonals. J. Biomol. NMR *62*, 63–69.

Clore, G.M., Starich, M.R., Bewley, C.A., Cai, M., and Kuszewski, J. (1999). Impact of residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints. J. Am. Chem. Soc. *121*, 6513–6514.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., and Wilczynski, B. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Cooper, A., and Dryden, D. (1984). Allostery without conformational change. Eur. Biophys. J. *11*, 103–109.

Cover, T.M., and Thomas, J.A. (1991). Entropy, relative entropy and mutual information. Elem. Inf. Theor. *2*, 12–13.

Crespo-Flores, S.L., Cabezas, A., Hassan, S., and Wei, Y. (2019). PEA-15 C-Terminal tail allosterically modulates death-effector domain conformation and facilitates protein–protein interactions. Int. J. Mol. Sci. *20*, 3335.

Hekstra, D.R., White, K.I., Socolich, M.A., Henning, R.W., Šrajer, V., and Ranganathan, R. (2016). Electric-field-stimulated protein mechanics. Nature *540*, 400–405.

Hummer, G., Schotte, F., and Anfinrud, P.A. (2004). Unveiling functional protein motions with picosecond x-ray crystallography and molecular dynamics simulations. Proc. Natl. Acad. Sci. U S A *101*, 15330–15334.

Ishima, R., and Torchia, D.A. (2000). Protein dynamics from NMR. Nat. Struct. Biol. *7*, 740–743.

Kay, L.E., Torchia, D.A., and Bax, A. (1989). Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. Biochemistry *28*, 8972–8979.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. Phys. Rev. E *69*, 066138.

Kullback, S. (1997). Information Theory and Statistics (Courier Corporation).

La Sala, G., Decherchi, S., De Vivo, M., and Rocchia, W. (2017). Allosteric communication networks in proteins revealed through pocket crosstalk analysis. ACS Cent. Sci. *3*, 949–960.

Levitt, M. (2007). Growth of novel protein structural data. Proc. Natl. Acad. Sci. U S A *104*, 3183–3188.

Long, D., and Brüschweiler, R. (2011). Atomistic kinetic model for population shift and allostery in biomolecules. J. Am. Chem. Soc. *133*, 18999–19005.

Luque, F., and Orozco, M. (2007). PCA Suite: Software Package for Lossy Trajectory Compression Using Principle Component Analysis Techniques (Molecular Recognition and Bioinformatics Group (University of Barcelona).

McClendon, C.L., Friedland, G., Mobley, D.L., Amirkhani, H., and Jacobson, M.P. (2009). Quantifying correlations between allosteric sites in thermodynamic ensembles. J. Chem. Theor. Comput. *5*, 2486–2502.

McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernández, C.X., Schwantes, C.R., Wang, L.-P., Lane, T.J., and Pande, V.S. (2015). MDTraj: A modern open library for the analysis of molecular dynamics trajectories. Biophysical J. *109*, 1528–1532.

Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., Carrillo, O., Camps, J., Fenollosa, C., and Repchevsky, D. (2010). MoDEL (molecular dynamics extended library): A database of atomistic molecular dynamics trajectories. Structure *18*, 1399–1409.

Monnot, C., Bihoreau, C., Conchon, S., Curnow, K.M., Corvol, P., and Clauser, E. (1996). Polar residues in the transmembrane domains of the type 1 angiotensin II receptor are required for binding and coupling: reconstitution of the binding site by co-expression of two deficient mutants. J. Biol. Chem. *271*, 1507–1513.

Nosé, S. (1984). A molecular dynamics method for simulations in the canonical ensemble. Mol. Phys. *52*, 255–268.

Orts, J., Vögeli, B., and Riek, R. (2012). Relaxation matrix analysis of spin diffusion for the NMR structure calculation with eNOEs. J. Chem. Theor. Comput. *8*, 3483–3492.

Palmer, A.G., III (2004). NMR characterization of the dynamics of biomacromolecules. Chem. Rev. *104*, 3623–3640.

Privalov, P. (1989). Thermodynamic problems of protein structure. Annu. Rev. Biophys. Biophys. Chem. *18*, 47–69.

Reynolds, D.A. (2009). Gaussian mixture models. Encyclopedia Biometrics *741*, 659–663.

Riek, R., Wider, G., Pervushin, K., and Wüthrich, K. (1999). Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. Proc. Natl. Acad. Sci. U S A *96*, 4918–4923.

Shannon, C.E., and Weaver, W. (1949). The Mathematical Theory of Information, *97* (University of Illinois Press).

Sheftic, S.R., Garcia, P.P., White, E., Robinson, V.L., Gage, D.J., and Alexandrescu, A.T. (2012). Nuclear magnetic resonance structure and dynamics of the response regulator Sma0114 from Sinorhizobium meliloti. Biochemistry *51*, 6932–6941.

Strotz, D., Orts, J., Kadavath, H., Friedmann, M., Ghosh, D., Olsson, S., Chi, C.N., Pokharna, A., Güntert, P., and Vögeli, B. (2020). Protein allostery at atomic resolution. Angew. Chem. Int. Ed. *59*, 22132–22139.

Theobald, D.L., and Wuttke, D.S. (2006). THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics *22*, 2171–2172.

Tiwari, S.P., Fuglebakk, E., Hollup, S.M., Skjærven, L., Cragnolini, T., Grindhaug, S.H., Tekle, K.M., and Reuter, N. (2014). WEBnm@ v2. 0: Web server and services for comparing protein flexibility. BMC Bioinformatics *15*, 1–12.

Vanwetswinkel, S., Kriek, J., Andersen, G.R., Dijk, J., and Siegal, G. (2003). 1H, 15N and 13C resonance assignments of the highly conserved 19 kDa C-terminal domain from human elongation factor 1Bgamma. J. Biomolecular NMR *26*, 189–190.

Vinh, N.X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. *11*, 2837–2854.

Vögeli, B., Kazemi, S., Güntert, P., and Riek, R. (2012). Spatial elucidation of motion in proteins by ensemble-based structure calculation using exact NOEs. Nat. Struct. Mol. Biol. *19*, 1053–1057.

Zhang, S., Krieger, J.M., Zhang, Y., Kaya, C., Kaynak, B., Mikulska-Ruminska, K., Doruker, P., Li, H., and Bahar, I. (2021). ProDy 2.0: Increased scale and scope after 10 years of protein dynamics modelling with Python. Bioinformatics *37*, 3657–3659.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
WW domain	(Strotz et al., 2020)	PDB: 6SVC
Protein GB3	(Vögeli et al., 2012)	PDB: 2LUM
Protein cyclophilin A	(Chi et al., 2015)	PDB: 2MZU
C-terminal domain of the human eEF1Bgamma subunit	(Vanwetswinkel et al., 2003)	PDB: 1PBU
Sma0114	(Sheftic et al., 2012)	PDB: 2LPM
PEA-15 Death Effector Domain in complex with ERK2	(Crespo-Flores et al., 2019)	PDB: 6P6C
Software and algorithms		
PDBcor	DOI: 10.5281/zenodo.5710842	https://github.com/dzmitryashkinadze/PDBCor
UCSF Chimera	https://www.cgl.ucsf.edu/chimera/	https://www.cgl.ucsf.edu/chimera/download.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Prof. Dr. Roland Riek (roland.riek@phys.chem.ethz.ch).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table.

All original code has been deposited at https://github.com/dzmitryashkinadze/PDBcor and is publicly available as of the data of publication. PDBcor server is available at https://www.pdbcor.ethz.ch/.

DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All data are generated from the datasets provided in the KRT.

METHOD DETAILS

This manuscript describes automated algorithm for the extraction of the correlated motion from the protein ensemble structures. All protein structures were downloaded from RCSB PDB Data bank and visualized with UCSF Chimera. All correlation matrix heatmaps were calculated and visualized with PDBcor software. PDB accession codes for all structures used in the figures are given in the figure captions and summarized in the key resource table.

Application of PDBcor to MD trajectories

In order to illustrate that PDBcor-based analysis can be applied to protein structure ensembles originating from techniques other than NMR, we analyzed a series of molecular dynamics (MD) trajectories. MD trajectories were downloaded from the (Molecular Dynamics Extended Library) (Meyer et al., 2010), https://mmb.irbbarcelona.org/MoDEL/. Compressed backbone MD trajectories for WW domain (PDB ID 1i6c), protein GB3 (PDB ID 2igd) and cyclophilin A (PDB ID 2cpl), each consisting of 10,000 frames simulating 10 ns, 10 ns and 80.5 ns, respectively, were downloaded, uncompressed with PCAsuite (Luque and Orozco, 2007), sliced down to 100 conformations with MDTraj (McGibbon et al., 2015) and loaded into PDBcor. Those MD trajectories were selected as they are corresponding to the structures analyzed in Figure 3. Resulting structural correlations are summarized in Figure S1.

Comparison of PDBcor to PCA- and NMA-based methods

In order to illustrate high sensitivity of the PDBcor we compared it to the conventional PCA-based technique THESEUS (Theobald and Wuttke, 2006) and NMA-based technique WEBnm@ (Tiwari et al., 2014). THESEUS performs structure alignment with maximum likelihood algorithm followed by PCA of the aligned protein coordinates that optimizes a correlation matrix. Unlike PDBcor, PCA-based approaches require structure superposition and are therefore biased by the way superposition was done. Furthermore, PCA-based approaches calculate correlations between Cartesian coordinates of individual residues, whereas in PDBcor we use interresidual distances that are more sensitive to the less pronounced, but statistically significant protein rearrangements. In turn, WEBnm@ approach is based on the analysis of torsion angles, whereas PDBcor is based largely on the interresidual distances and therefore PDBcor by design is more sensitive to correlated motion of secondary structure elements or protein domains.

Structural correlations of the cyclophilin A, a known and reported allosteric protein, were analyzed with PDBcor, THESEUS and WEBnm@ and compared in Figure S2. Whereas PDBcor results overlap with reported findings as shown in Figure 3, THESEUS and WEBnm@ techniques failed to reproduce them.

QUANTIFICATION AND STATISTICAL ANALYSIS

Detailed description of the statistical analysis for the PDBcor is described in the dedicated Theory section.