JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 4929−4941

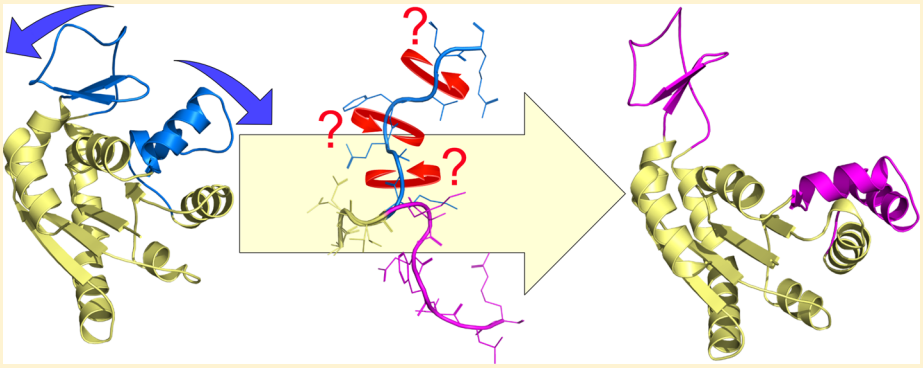
pubs.acs.org/jcim

Can Conformational Changes of Proteins Be Represented in Torsion Angle Space? A Study with Rescaled Ridge Regression

Ugo Bastolla* and Yves Dehouck*

Centro de Biologia Molecular "Severo Ochoa", CSIC-UAM Cantoblanco, 28049 Madrid, Spain

Supporting Information



ABSTRACT: Torsion angles are the natural degrees of freedom of protein structures. The ability to determine torsional variations corresponding to observed changes in Cartesian coordinates is highly valuable, notably to investigate the mechanisms of functional conformational changes or to develop computational models of protein dynamics. This issue is far from trivial in practice since the impact of modifying one torsion angle strongly depends on all other angles, and the compounding effects of small variations in bond lengths and valence angles can completely disrupt a protein fold. We demonstrate that naive strategies, such as directly comparing torsion angles between structures without correcting for variations in bond lengths and valence angles or fitting torsional variations without a proper regularization scheme, fail at producing an adequate representation of conformational changes in internal coordinates. In contrast, rescaled ridge regression, a method recently introduced to regularize multidimensional regressions with correlated explanatory variables, is shown to consistently identify a minimal set of torsion angles variations that closely reproduce changes in Cartesian coordinates. This torsional representation of conformational changes is shown to be robust to the choice of experimental structures. It also provides a better agreement with theoretical models of protein dynamics than the Cartesian representation, regarding notably the predominance of low-frequency normal modes in functional motions and the presence, in predicted equilibrium dynamics, of hints of natural selection for specific functional motions. The software is available at https://github.com/ugobas/tnm.

INTRODUCTION

The intrinsic dynamical properties of proteins often play a fundamental role in their functional activity, notably for catalysis,[1−3] allosteric regulation,[4−6] or molecular recognition.[7,8] A variety of computational approaches have been developed to study protein dynamics, ranging from detailed models with a high computing cost, such as molecular dynamics,[9,10] to coarse-grained models that can be more easily suited to large-scale studies.[11−14] Elastic network models (ENM) are well-known representants of the latter category. Taking as sole input the native structure of a protein, these models provide a global prediction of native protein dynamics at very low computational cost, under the assumption that the observed structure corresponds to the free-energy minimum and is minimally frustrated.[15−17]

The nature and level of detail of the structural representation is an important design choice for any computational model.[18] Instead of Cartesian coordinates, the structure of a protein can be described by its internal coordinates: the bond length, valence angle, and torsion angle. Since bond lengths and valence angles are strongly constrained by covalent forces, torsion angles are often considered as the natural degrees of freedom for describing protein structures and motions. This representation presents advantages for computations since the backbone conformation can essentially be described by 2 degrees of freedom per residue (the ϕ and ψ torsion angles, except for the rare cis/trans isomerizations of the peptide bond), instead of nine (the three-dimensional coordinates of the N, C_α, and C atoms). Torsional descriptions have long been part of the protein modeling scene and have been successfully used in a variety of applications, such as structure prediction or calculation from NMR constraints, protein−ligand and protein−protein docking, protein dynamics, or

Received: July 29, 2019
Published: October 10, 2019

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

protein design.[19−23] The torsional network model (TNM)[24] follows the same concept as the ENM, but in the space of torsion angles, generating thus normal modes of motion that automatically preserve the integrity of the bond geometries.

A challenging aspect of the development of computational procedures consists in the definition of reliable approaches to evaluate model performance. The quality of the reproduction of thermal fluctuations around the native state is often quantified by measuring the correlation with crystallographic B-factors[25−28] or by comparing the fluctuations of interresidue distances in the model and in experimental structural ensembles.[29] To evaluate more specifically the description of functional conformational changes, advantage can be taken of the availability of proteins, for which multiple structures have been resolved in different conformational states (e.g., open and closed states, ligand bound, and apo form). Such functional conformational changes have been shown to correlate well with the low-frequency normal modes of thermal dynamics, predicted by ENMs.[30−35] This observation, which is somewhat surprising since ENM computations are in principle only valid for very small deviations around the equilibrium, has tremendously boosted the recognition of the validity, and the popularity, of this type of computational model.

In this study, we address the question of how well functional conformational changes can be described using only torsion angles. We present and evaluate different approaches to determine appropriate variations in torsion angles from observed changes in Cartesian coordinates. And we investigate whether correlations between conformational changes and predicted thermal dynamics are enhanced when torsion angles are used as coordinates.

The translation of conformational changes from Cartesian to internal coordinates is not trivial since small local variations of torsion angles can produce large global Cartesian displacements and since an approximate match of the Cartesian coordinates can be obtained with very different values of the torsion angles. Moreover, the complexity of the problem is greatly increased if small variations in bond lengths and valence angles must be considered. Yet, these questions hold significant importance for the development and evaluation of computational models of protein dynamics and, more generally, for decoding functional motions in proteins. Indeed, the ability to identify a small set of residues that contribute most to a conformational change can bring valuable insights into the mechanisms of this functional motion and reveal new options for the design of altered functional dynamics, either via mutations or via small ligands targeted at critical regions.

RESULTS

We investigate how well a conformational change of the backbone of a protein can be described using only 2 degrees of freedom per residue: the ϕ and ψ torsion angles. We examine 31 pairs of PDB structures, representing the same protein chain in two different conformations. Each structure in a pair is, in turn, considered as the initial conformation (A), and the other as the final conformation (B). A total of 62 conformational changes are thus analyzed. The root-mean-square deviation RMSD(A,B) between the Cartesian coordinates $\mathbf{r}_i^A$ and $\mathbf{r}_i^B$ of two structures in a pair (for every backbone atom i) ranges from 0.35 to 34.4 Å, with a mean of 3.32 Å. The smaller conformational changes correspond to allosteric proteins, and the larger ones describe functional motions of molecular machines such as chaperones, polymerases, or transporters (see Methods).

For each conformational change, we modify the ϕ and ψ torsion angles in A, with the objective of creating a transformed structure A* that closely resembles B. The bond lengths l_a^A and valence angles θ_a^A (for every backbone bond a) are left unchanged during this transformation, which is thus uniquely defined by the differences in torsion angles $\Delta\varphi_a = \varphi_a^{A*} - \varphi_a^A$. The ω torsion angles are considered to be fixed as well ($\Delta\varphi_a \equiv 0$ for ω torsion angles), except for the few ones that undergo a cis/trans isomerization. We investigate different approaches to identify adequate values of $\Delta\varphi_a$ from the observed coordinates of the two conformations and evaluate the quality of the resulting description of the conformational change by computing RMSD(A*,B).

Switch Torsions (ST). The most naive and straightforward strategy consists in simply computing the differences of aligned torsion angles, $\Delta\varphi_a^{ST} = \varphi_a^B - \varphi_a^A$. The ϕ and ψ torsion angles in the transformed initial conformation A* are thus given identical values as in the final conformation B. This approach relies on the common assumption that differences in bond lengths, valence angles, and ω torsion angles (barring a cis/trans isomerization) between the two structures are small enough to be negligible.

Such an assumption appears to be very reasonable since those differences are indeed quite small. The RMSD between the lengths of corresponding backbone bonds in the initial and final structures is equal to 0.016 Å (i.e., about 1% of the bond length). The RMSD between corresponding valence angles and between corresponding ω torsion angles (excluding cis/trans isomerizations) is equal to 0.047 and 0.068 radians (2.7 and 3.9°), respectively. These small differences are not necessarily relevant to the conformational change, but can be consequences of the structure refinement procedure. They are indeed only slightly larger than those observed between different bonds within the same structure (Table 1).

Table 1. Variability of Internal Coordinates

	μ^a	σ^b	RMSD^c
	Bond Lengths (Å)		
C−N	1.33	0.012	0.017
N−C_α	1.46	0.011	0.014
C_α−C	1.52	0.013	0.017
all		0.012	0.016
	Valence Angles (rad)		
C−N−C_α	2.12	0.034	0.043
N−C_α−C	1.94	0.059	0.060
C_α−C−N	2.03	0.027	0.034
all		0.042	0.047
	Torsion Angles (rad)		
ϕ: C−N−C_α−C	−1.41	0.709	0.432
ψ: N−C_α-C−N	−0.33	1.686	0.461
ω: C_α−C−N−C_α	3.14	0.057	0.068

[a]Mean value of the internal coordinate, over all backbone bonds in the proteins of our data set (but ignoring cis ω bonds). [b]Standard deviation of the internal coordinate. For valence and torsion angles, we use the angular definition of the mean and standard deviation, as given by eq 4 (Methods). The values for "all" bond lengths and valence angles are calculated as the RMSD from the mean, where the mean depends on the bond type. [c]RMSD of the internal coordinate between corresponding backbone bonds in each pair of structures representing a conformational change.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

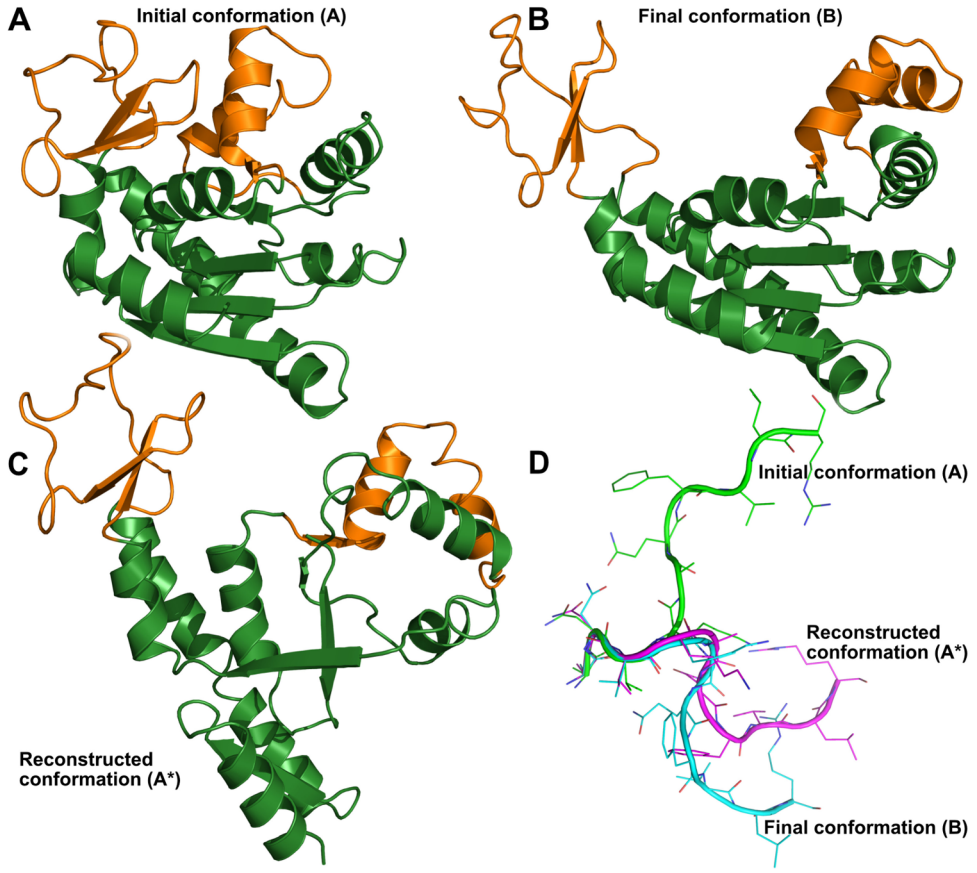


Figure 1. Examples of the quality of the modeling of conformational changes using the Switch torsions (ST) approach. (A, B) Chain A of adenylate kinase from *Escherichia coli*, in its closed (PDB: 1ank) and open conformation (PDB: 4ake). The RMSD between these two conformations is equal to 7.1 Å, and is mostly due to the displacement of the two regions depicted in orange, while the rest of the structure remains unaltered. (C) Conformational change modeled with ST: the backbone torsion angles in the initial conformation A were modified to match those in the final conformation B. This reconstruction is unsuccessful, as it yields a poorly folded structure that is very different from the target: RMSD(A*,B) = 11.0 Å. (D) Conformations of a 10-residue fragment of aspartate transcarbamoylase from *E. coli* (initial PDB: 1d09; final PDB: 1raa). The different conformations were aligned using the first four residues. On the scale of such a small peptide, the reconstruction can be somewhat successful in creating a conformation closer to the target. However, the small errors due to bond lengths and valence angles quickly propagate and create differences that are already notable a few residues down the chain.

However, as illustrated in Figure 1, small differences in internal coordinates create errors that quickly propagate and can have a major impact on the reconstructed Cartesian coordinates. As a result, the transformed structures A* end up being very different from the target structures B, with RMSD(A*,B) equal to 16.2 Å on average, and almost systematically much larger than RMSD(A,B). Modifying the torsion angles without any sort of compensation for the small differences in bond lengths and valence angles yields thus a very poor reconstruction of the conformational changes. Since the main issue is the propagation of small errors along the chain, the performances of this approach tend to deteriorate when longer protein chains are considered (Figure 2 and Table S1).

These results can obviously be improved by extending the approach to consider more than 2 degrees of freedom. If, in addition to ϕ and ψ, all ω torsion angles in the transformed structure A* are also given the same values as in the target structure B, the mean RMSD(A*,B) decreases to 14.2 Å. Valence angles are particularly important, and including them yields an average RMSD(A*,B) of 0.2 Å. If bond lengths are also included, the transformed structures are virtually identical

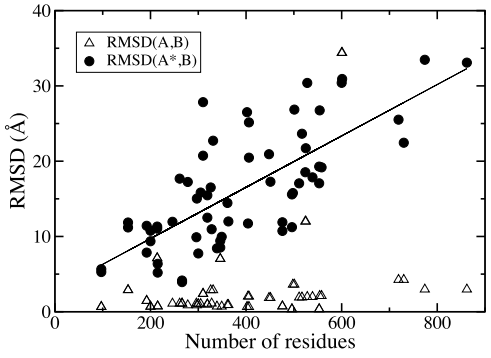


Figure 2. Performances of the switch torsions (ST) approach. The RMSD between the reconstructed structure A* and the target structure B is given as a function of the number of residues in the protein chain (filled circles). These values are almost systematically much larger than the RMSD between the initial structure A and the target (empty triangles), which emphasizes the poor quality of the reconstruction. RMSD(A*,B) correlates strongly with the number of residues ($r = 0.74$, $p < 10^{-11}$) and slightly with the amplitude of the conformational change, RMSD(A,B) ($r = 0.39$, $p = 0.002$). The latter is partially explained by the correlation between RMSD(A,B) and the number of residues ($r = 0.29$, $p = 0.02$).

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

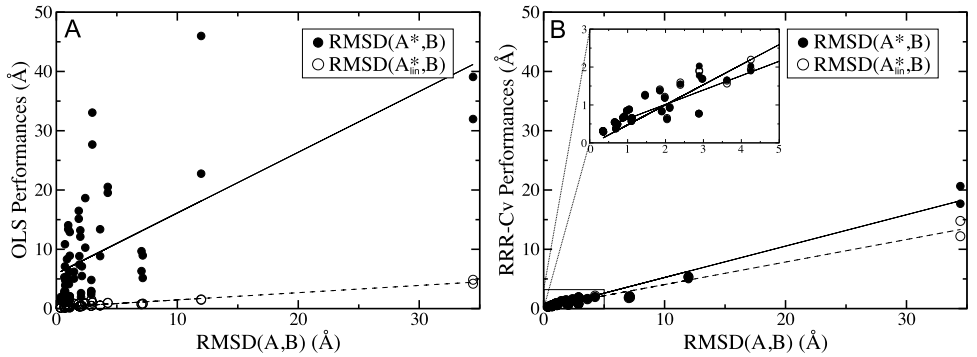


Figure 3. Performances of the regression approaches. RMSD between the reconstructed structure A^*_{lin} or A*, and the target structure B, as a function of the amplitude of the conformational change. (A) Performances of the OLS approach. RMSD(A^*_{lin},B) values are remarkably small, and are strongly correlated with RMSD(A,B) ($r = 0.95$, $p < 10^{-12}$): the error of the fit is proportional to the variation that we aim to fit, i.e., the relative error is essentially constant. However, deviations from the linear approximation are significant, and RMSD(A*,B) values tend to be much larger. RMSD(A*,B) correlates with RMSD(A,B) ($r = 0.66$, $p < 10^{-8}$) as well as with the number of residues ($r = 0.58$, $p < 10^{-6}$). (B) Performances of the rescaled ridge regression (RRR) approach, with the Cv criterion. The regularization of the fit produces larger RMSD(A^*_{lin},B) values than with OLS. However, deviations from the linear approximation are much smaller and, in most cases, RMSD(A*,B) remains practically indistinguishable from RMSD(A^*_{lin},B). Both values are strongly correlated with RMSD(A,B) ($r > 0.98$, $p < 10^{-12}$).

to the targets, except for numerical imprecisions, and RMSD(A*,B) ≈ 0.01 Å.

Ordinary Least-Squares Regression (OLS). Since the impact of small differences in bond lengths, valence angles, and ω torsion angles cannot be neglected, modeling a conformational change using only 2 degrees of freedom per residue requires the identification of appropriate values of the ϕ and ψ torsion angles, which may be different from those in the target structure.

For that purpose, we can exploit the analytical relationship between variations in Cartesian coordinates and variations in torsion angles, which at first order in $\Delta\varphi$ can be written as $\Delta r_i = \sum_a J_{ia}\Delta\varphi_a$, or $\Delta\mathbf{r} = \mathbf{J}\Delta\varphi$ in matrix notation, where $\mathbf{J}$ is the $(9N \times 2N)$ Jacobian matrix of the transformation between internal and Cartesian coordinates, and N is the number of residues. This relationship is underdetermined since we consider only two internal degrees of freedom per residue (ϕ and ψ torsion angles), while nine are necessary for the Cartesian description of the main backbone atoms (N, C_α, and C). The values of the torsion angles can, however, be obtained by linear regression, minimizing the square error $E = (\mathbf{J}\Delta\varphi - \Delta\mathbf{r})\cdot\mathbf{M}(\mathbf{J}\Delta\varphi - \Delta\mathbf{r})$, where each atom is weighted by its mass via the diagonal matrix $\mathbf{M}$, and $\mathbf{x}\cdot\mathbf{y}$ denotes the scalar product (see Methods).

In principle, this approach should identify the values of the ϕ/ψ torsion angles that yield an optimal reconstruction of the conformational change, except for the fact that it relies on a linear approximation of the relationship between Cartesian and internal coordinates. We denote by A^*_{lin} the artificial reconstruction of the conformational change, obtained within the context of this linear approximation: $\mathbf{r}_{A^*_{lin}} = \mathbf{r}_A + \mathbf{J}\Delta\varphi$. The RMSD($A^*_{lin}$,B) is on average equal to 0.65 Å, which indicates that the regression is successful in minimizing the deviation between the coordinates of the transformed starting structure and those of the target structure.

However, because of the linear approximation, A^*_{lin} is not a valid protein structure, as bond lengths and valence angles are not respected. A very different picture emerges when we compare the structure A*, properly reconstructed using the $\Delta\varphi$ from the linear regression, to the target structure B: the average value of RMSD(A*,B) is equal to 9.28 Å. The OLS approach is thus superior to the ST approach described above, but still provides a very poor reconstruction of the conformational changes. The performances tend to depend on both the length of the protein chain and the amplitude of the modeled conformational change (Figure 3A). Indeed, deviations from the linear approximation are sharper when RMSD(A,B) is large and create errors that propagate and get amplified along the protein chain.

Interestingly, extending the OLS approach by including the ω torsion angles does not improve the performances. Even though RMSD(A^*_{lin},B) decreases, the consequences of the deviations from the linear approximation are amplified, and the reconstructed structures end up even more dissimilar to the target structures, with RMSD(A*,B) equal to 14.9 Å on average.

Rescaled Ridge Regression (RRR). In the regression problem addressed here, the explanatory variables are highly correlated: the impact of modifying one torsion angle on the Cartesian coordinates strongly depends on the values of the other torsion angles. In such cases, OLS is known to be bad-behaved and prone to overfitting, i.e., fitting too closely to a set of data points affected by some level of noise or error, resulting in poor generalization performances, and often unphysical values of the parameters. Here, the noise results from the approximate nature of the linear relationship between $\Delta\varphi$ and Δr, and the poor generalization performances are demonstrated by the large differences between RMSD(A^*_{lin},B) and RMSD(A*,B). A popular solution to this problem is the Tykhonov regularization, or ridge regression (RR), which consists in penalizing large values of the fit parameters.[36] In this framework, the function to be minimized is $E + \Lambda(\Delta\varphi\cdot\Delta\varphi)$, where E is the error of the fit and Λ is the Tykhonov parameter.

In the limit of large Λ, the fitted parameters $\Delta\varphi$ tend to zero, and the regression becomes equal to the intercept of the fit, which is not penalized in ordinary ridge regression. In the present case, there is no intercept since it does not make physical sense to have nonzero $\Delta\mathbf{r}$ when the $\Delta\varphi$ vanish. We previously introduced a variant of the Tykhonov regularization scheme, called rescaled ridge regression (RRR), to deal with cases where the intercept is either absent or must be interpreted as a physical parameter and penalized like any

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

other.[27] In RRR, the fit parameters are rescaled by a factor that diverges in the large Λ limit so that they tend to finite values instead of vanishing. This is achieved by the use of a second Lagrange multiplier, as described in Methods. The fit parameters obtained from RRR are directly related to those obtained from ordinary RR: $\Delta\varphi^{RRR} = \nu(\Lambda)\Delta\varphi^{RR}$, where the factor $\nu(\Lambda)$ ensures that the fit respects the scale of the dependent variable(s).

A critical aspect is the determination of an adequate value of Λ. While different methods have been proposed for that purpose,[37−39] there is no consensus on an optimal approach. We previously showed that the quantity minimized in RR is formally equivalent to a free energy, where the error of the fit corresponds to the energy, $(\Delta\varphi\cdot\Delta\varphi)$ the entropy, and Λ the temperature. The error monotonically increases with Λ, and we can see $\partial E/\partial\Lambda$ as the specific heat. This interpretation suggests two criteria for selecting an appropriate value of Λ[27]

1. Cv criterion: Λ is chosen so as to maximize the "specific heat" Cv = $\partial E/\partial\Lambda$ since this maximum is expected to separate the small Λ regime dominated by overfitting from the large Λ regime dominated by the error.
2. MP criterion: Λ is chosen so as to maximize the penalty term $\Lambda(\Delta\varphi\cdot\Delta\varphi)$, which vanishes for both $\Lambda = 0$ and $\Lambda \to\infty$. In RRR, we maximize $\Lambda((\Delta\varphi - \Delta\varphi^{\infty})\cdot(\Delta\varphi - \Delta\varphi^{\infty}))$, where $\Delta\varphi^{\infty}$ are reference parameters obtained in the $\Lambda \to \infty$ limit.

Within the context of the linear approximation, the RRR method does, of course, produce a larger error than the OLS fit since the error is only one term of the minimized quantity. We find that RMSD(A^*_{lin},B) is on average equal to 1.84 Å with the MP criterion, and 1.52 Å with the Cv criterion, compared to 0.67 Å with OLS. However, the regularization of the fit drastically reduces overfitting and the occurrence of unphysical parameter values, entailing therefore much smaller deviations from the linear approximation. The transformed structures A*, reconstructed using the $\Delta\varphi$ from the regression while respecting the bond lengths and valence angles, remain very similar to the corresponding A^*_{lin} structures and provide a much more accurate approximation of the target structures B. The average value of RMSD(A*,B) is equal to 1.90 Å with RRR-MP, and 1.70 Å with RRR-Cv, compared to 9.28 Å with OLS.

Both RMSD(A^*_{lin},B) and RMSD(A*,B) are strongly correlated with RMSD(A,B), indicating that the performance of the RRR fit is poorer in the case of large conformational changes (Figure 3B and Table S1). Unlike OLS, including the ω torsion angles does not worsen the quality of the fit, albeit the improvement remains marginal: RMSD(A*,B) is reduced to 1.88 Å with RRR-MP and to 1.68 Å with RRR-Cv.

Greedy Iterative Search (GIS). In this method, we start from the initial conformation A and iteratively update the torsion angles (keeping the bond lengths and valence angles fixed) so as to reach a conformation as similar as possible to the final conformation B.

At each step, we perform the RRR described above, but with the scale factor ν modified in such a way that $\Delta\varphi_a \le \delta\forall a$ and that the RMSD from the previous configuration is smaller than a certain threshold. In other words, only small variations of the overall structure and of each individual torsion angle are allowed at each step, to ensure that $\Delta\varphi$ remains within the range of validity of the linear approximation, $\Delta\mathbf{r} \approx \mathbf{J}\Delta\varphi$. The Cartesian coordinates are then reconstructed, and the move is accepted if the RMSD from the target conformation decreases, otherwise δ is reduced. When δ becomes smaller than a predefined threshold, δ is reinitialized at $\delta = \delta_{ini}$ and a second phase of the iterative search is triggered. At each step, the torsion angles are changed one by one by exactly δ, and the move that most reduces the RMSD from the target conformation is accepted. If no such move exists, δ is further reduced, and the procedure stops when it reaches the predefined minimum value.

At the expense of a somewhat heavier computational burden, this greedy minimization provides a considerable improvement in the quality of the conformational change modeling, with an average RMSD(A*,B) as low as 0.79 Å. RMSD(A*,B) is correlated with the amplitude of the modeled conformational change RMSD(A,B) ($r = 0.86$, $p < 10^{-12}$) and, to a lower extent, with the number of residues in the protein ($r = 0.37$, $p = 0.003$). Yet, even for the largest conformational change in our data set (600 residues, RMSD(A,B) = 34 Å), the reconstructed conformation is still reasonably close to the target, with RMSD(A*,B) = 3.2 Å, instead of ∼30 Å with ST or OLS, or ∼20 Å with RRR (Table S1).

Besides the reconstruction of the final structure, GIS also produces a series a dynamical snapshots of the conformational change. Various morphing techniques have been developed to create such trajectories between different structural states of a protein.[41,42] We do not investigate here the quality and physical relevance of the trajectory itself, which is of course of paramount importance in morphing applications. We can, however, compare the final reconstructed structures to those produced by iMODS, a morphing technique similarly built in the space of torsion angles.[40] Since the iMODS server does not always deal with missing atoms and chain breaks, we could only retrieve results for a fraction of the structural transitions in our data set. When considering an extended backbone representation (with five atoms per residue), the average RMSD(A*,B) obtained with iMODS on 33 out of 62 transitions is equal to 1.07 Å, compared to 0.82 Å with GIS on the same subset (Table S1).

Analysis of the Modeled Torsional Conformational Changes. A comparative summary of the results obtained with the different approaches is given in Table 2. The RMSD(A*,B) between the reconstructed and the target structure measures the quality of the modeling of the conformational changes, while RMS($\Delta\varphi$) and $C(\Delta\varphi)$ quantify, respectively, the average amplitude of the torsional

Table 2. Summary of the Properties of the Torsional Conformational Changes

	RMSD(A*,B)[a]	RMS($\Delta\varphi$)[b]	$C(\Delta\varphi)$[c]	S_{MP}[d]
ST	16.2 Å	0.41	17%	3.10
OLS	9.28 Å	0.37	27%	3.10
RRR-MP	1.90 Å	0.01	4.3%	2.31
RRR-Cv	1.70 Å	0.03	2.7%	2.32
GIS	0.79 Å	0.08	3.2%	2.25

[a]Average RMSD between the reconstructed structure A* and the target conformation B. [b]Average RMS of the torsional variations $\Delta\varphi$ between the A and A* conformations. [c]Average collectivity of the torsional variations $\Delta\varphi$ between A and A* (see Methods). [d]Average MolProbity score[43] of the reconstructed structures A*, after optimization of the side-chain conformations using SIDEpro.[44] This composite score evaluates the quality of the structural model (steric clashes, unusual torsion angle values, etc.) and is normalized to approximate the resolution at which that score would be average.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

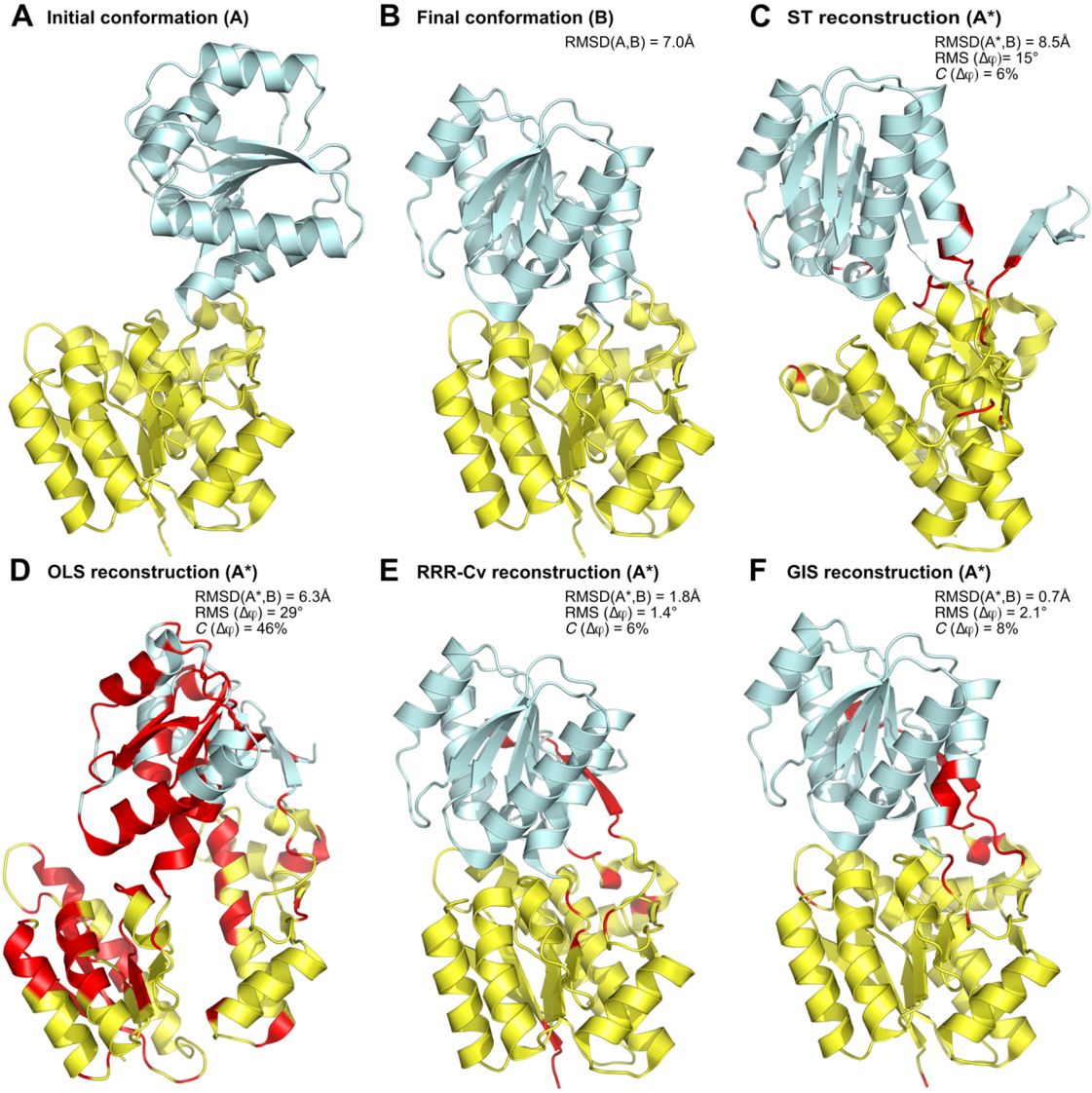


Figure 4. Example of the reconstruction of conformational changes with different approaches. (A, B) Chain A of L-leucine-binding protein from *E. coli*, in its initial apo state (PDB: 1usg) and final bound state (PDB: 1usk). The RMSD between these two structures is equal to 7.0 Å, and is mostly due to the relative motion of the two domains (in yellow and cyan), from an open to a closed conformation. (C−F) Conformational change modeled with different approaches: the backbone torsion angles φ_a in the initial conformation A were modified in an attempt to reproduce the final conformation B. A lower value of RMSD(A*,B) indicates a more successful reconstruction. For each method, we also indicate the root-mean-square value of the variations of the torsion angles, RMS($\Delta\varphi$), as well as their collectivity, $C(\Delta\varphi)$. The subset of residues with the highest $\Delta\varphi_a$ values, corresponding to the percentage of collectivity, is highlighted in red.

changes (between A and A*) and the torsional collectivity, i.e., the effective fraction of torsion angles that are modified in the conformational changes (see Methods).

An interesting observation is that the worst performing methods, ST and OLS, both produce large average deviations of the torsion angles (RMS($\Delta\varphi$) = 0.41 and 0.37 radians, respectively), spread over a large fraction of the residues ($C(\Delta\varphi)$ = 17 and 27%, respectively). In contrast, the regularized fits (RRR with the MP or Cv criterion) yield much smaller angular deviations, which is not surprising since ridge regression penalizes large values of the fit parameters $\Delta\varphi$. The torsional collectivity is also strongly reduced with the RRR fits ($C(\Delta\varphi)$ = 4.3 and 2.7%, respectively), i.e., only a small fraction of the residues contribute to the modeled conformational changes, which is consistent with the view that functional motions in proteins are often operated by a small number of hinge regions. In the case of conformational changes reconstructed via the iterative strategy GIS, the amplitude of the torsional deviations tends to be somewhat larger than with the RRR fits, but still much smaller than with ST or OLS.

These differences are illustrated by an example in Figure 4. Upon binding, L-leucine-binding protein undergoes a conformational change from an open to a closed state. The two domains of the protein move relatively to each other, but mostly conserve their internal structure (Figure 4A,B). As depicted in Figure 4C−F, all methods succeed in generating a closed conformation by modifying the torsion angles in the open structure. However, with ST and OLS, the large amplitudes of the torsional changes create errors that severely disrupt the internal structures of the individual domains. With RRR and GIS, the deviations of the torsion angles are much

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

smaller and tend to be localized in proximity of the hinge region, resulting in a significantly improved reconstruction of the conformational change.

The structural quality of the models created by the different approaches was evaluated by computing the average MolProbity score (S_{MP})[43] of the reconstructed structures A*, after optimization of the side-chain conformations using SIDEpro.[44] As shown in Table 2, the RRR and GIS approaches produce structures with fewer structural defects ($S_{MP} \approx 2.3$) than ST and OLS ($S_{MP} \approx 3.1$). For comparison, the MolProbity score of the 62 X-ray structures in our data set is on average equal to 2.69 without, or 2.04 with, optimization of the side-chain conformations.

Robustness of the Torsional Description of Conformational Changes. The results in Table 2 show that the target structures can be reconstructed with good precision, by applying $\Delta\varphi$ modifications that are more localized, and of much smaller amplitude, than those observed between the initial and final structures. Therefore, when comparing two structures representative of a conformational change, this begs the question of how meaningful the observed differences in internal coordinates can be considered, and whether they can be relied on to identify important residues and hinge regions.

We illustrate this issue with the example of adenylate kinase from *E. coli*, by considering five structures of the protein in its closed form and two structures in its open form. The 10 possible closed−open structural pairs give thus 10 distinct representations of the conformational change, which are all very consistent in terms of Cartesian coordinates, with a small RMSD between structures in the same form, and a large RMSD between closed and open structures (Table 3). The

Table 3. Variability of Cartesian/Internal Coordinates among Structures of a Protein

	closed−closed	open−open	closed−open
$RMSD_r$ (Å)[a]	[0.32−0.65]	1.11	[6.33−7.25]
$RMSD_l$ (Å)[b]	[0.01−0.02]	0.01	[0.01−0.02]
$RMSD_\theta$ (rad)[c]	[0.02−0.06]	0.04	[0.03−0.06]
$RMSD_\varphi$ (rad)[d]	[0.14−0.37]	0.26	[0.29−0.43]

[a]RMSD between the Cartesian coordinates of backbone atoms in adenylate kinase structures: interval of values obtained for 10 pairwise comparisons between five closed structures (PDB: 1ake, 1ank, 2eck, 3hpq, 4jzk) for the comparison between two open structures (PDB: 4ake and 6f7u) and for 10 comparisons between a closed and an open structure. [b]RMSD between corresponding backbone bond lengths. [c]RMSD between corresponding backbone valence angles. [d]RMSD between corresponding ϕ or ψ torsion angles.

two structural states are, however, much more difficult to differentiate in terms of internal coordinates. The small differences in bond lengths and valence angles ($RMSD_l$ and $RMSD_\theta$) are of similar amplitude whether the structures are in the same form or not. Variations in torsion angles ($RMSD_\varphi$) do tend to be larger between closed and open structures than between two closed or two open structures, but the difference is small and not systematic.

In Figure 5A, the differences in torsion angles that characterize the closed−open transition are given for each position in the sequence, using two different pairs of structures (either 1ank−4ake or 3hpq−6f7u). Although some similarities exist between the two $|\Delta\varphi|$ profiles, there is a clear lack of robustness in the identification of the residues that contribute

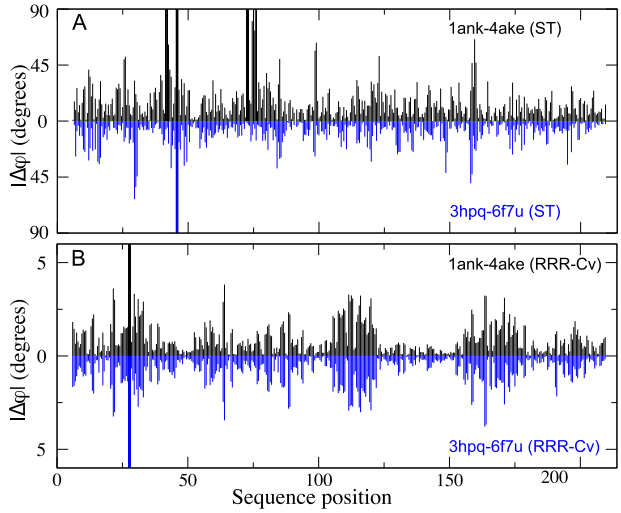


Figure 5. Torsional description of a conformational change using different pairs of structures. The absolute value of the variation in ϕ and ψ torsion angles between the initial closed form (A) and reconstructed open form (A*) of adenylate kinase is given for each sequence position, excluding five N- and C-terminal residues. Results obtained from the PDB structures 1ank and 4ake (black) are compared to those obtained from the structures 3hpq and 6f7u (blue). (A) The open form A* is reconstructed using the ST method, i.e., the plotted $|\Delta\varphi|$ are those observed between the original PDB structures of the closed and open forms. (B) The open form A* is reconstructed using the RRR-Cv method.

most to the conformational change. The overlap $O_{\Delta\varphi}$ between the subsets of torsion angles that undergo the largest modifications in each description of the conformational change is as low as 38%, on average over 40 similar pairwise comparisons (Tables 4 and S2).

Table 4. Robustness of the Torsional Representation

	$O_{\Delta\varphi}$[a]	$r_{\Delta\varphi}$[b]
ST	0.38 ± 0.01	0.34 ± 0.02
OLS	0.63 ± 0.01	0.83 ± 0.01
RRR-MP	0.71 ± 0.03	0.73 ± 0.04
RRR-Cv	0.72 ± 0.02	0.79 ± 0.03
GIS	0.61 ± 0.03	0.77 ± 0.02

[a]Overlap between the 10% of torsion angles that undergo the largest modifications in two different descriptions of a conformational change. We report the mean and standard error of $O_{\Delta\varphi}$ over 40 pairwise comparisons (20 for the closed−open, and 20 for the open−closed transition of adenylate kinase). In each comparison, the two descriptions are based on different PDB structures for both the open and closed conformations, excluding the five N- and C-terminal residues. [b]Mean and standard error of the angular correlation between the $\Delta\varphi$ from different descriptions of a conformational change, over 40 pairwise comparisons.

In contrast, Figure 5B shows that the $|\Delta\varphi|$ profile obtained with RRR-Cv is almost identical for both structural pairs. This is quite impressive considering the large differences in internal coordinates between the two closed structures and between the two open structures. The RRR approach manages thus to capture—in internal coordinates—the essential motions required to transition from the closed to the open form, in a way that is very robust with respect to the choice of experimental structures. Important regions can be identified

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

in a much more reliable manner, as witnessed by an average overlap $O_{\Delta\varphi}$ of 72% (Tables 4 and S2).

Correlation with Predicted Fluctuations. Finally, we investigate how well the torsional descriptions of the conformational changes, created by the different geometrical methods described above, compare with theoretical predictions based on physical considerations. More precisely, we use the TNM, an elastic network model in torsion angles space (see Methods), which provides as output the set of normal modes of motion $\mathbf{u}_\alpha$ (each with its associated frequency ω_α), as well as the mean-square fluctuations of the torsion angles $(\Delta\varphi_a^{\text{therm}})^2$, in the thermal equilibrium ensemble. The correlation, over all torsion angles a, between the variations associated with the conformational change $(\Delta\varphi_a)^2$ and those predicted in the thermal ensemble $(\Delta\varphi_a^{\text{therm}})^2$ is significant but limited: $r_{\text{fluct}} \approx 0.15-0.20$ for all methods (Figure 6A). This is not surprising

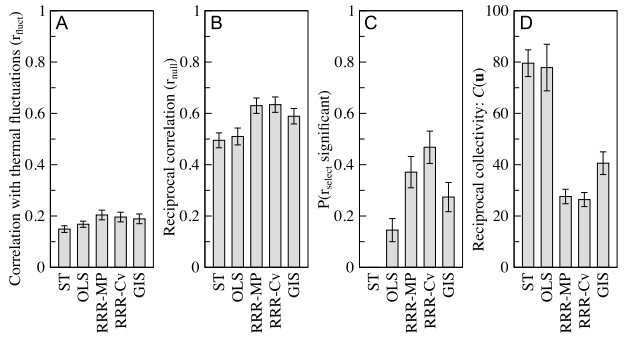


Figure 6. Correlations between conformational changes and predicted equilibrium dynamics. The torsional conformational changes, modeled by each of the five considered methods, are compared to the thermal dynamics at equilibrium, as predicted by the torsional network model (TNM). Several measures are presented, as described in the text. In each case, the error bars correspond to the standard error. (A) Correlation r_{fluct} between the torsion angles variations in the conformational change, $(\Delta\varphi_a)^2$, and the mean-square fluctuations of the torsion angles in the equilibrium dynamics predicted by the TNM, $(\Delta\varphi_a^{\text{therm}})^2$. (B) Correlation r_{null} between the contributions of torsional normal modes to the conformational change, c_α^2, and to thermal dynamics, ω_α^{-2}. This correlation reflects the agreement with the null model of protein response. (C) Fraction of conformational changes for which the correlation r_{select} (between $c_\alpha^2\omega_\alpha^2$ and ω_α^{-2}) is positive and significant, indicating a lowered free-energy barrier as a sign of natural selection. (D) Number of normal modes that effectively participate to the conformational change (collectivity).

since the $(\Delta\varphi_a^{\text{therm}})^2$ values predicted by the TNM describe the ensemble of thermal fluctuations around equilibrium, while the $(\Delta\varphi_a)^2$ values correspond to one specific functional motion, away from equilibrium.

However, it has been observed that low-frequency normal modes, which contribute more to thermal dynamics, tend to also contribute more to functional conformational changes.[30] This is consistent with linear response theory. It was proposed that the response of a protein to a generic perturbation can be modeled as a conformational change in which the contribution c_α^2 of each normal mode α is proportional to its contribution to thermal dynamics, i.e., the inverse of the squared mode frequency, ω_α^{-2}. This null model of protein response (see Methods) was verified on a large ensemble of functional and nonfunctional conformational changes.[33] Accordingly, this "reciprocal" correlation in the space of normal modes, $r_{\text{null}} = r(c_\alpha^2\omega_\alpha^{-2})$, is observed here for all methods (Figure 6B). On average, it is lowest with the ST method ($\langle r_{\text{null}}\rangle = 0.50$) and largest with RRR ($\langle r_{\text{null}}\rangle = 0.63$), providing further evidence of the superiority of the latter methodology.

The correlation r_{null} between the contribution of normal modes to predicted thermal fluctuations, and to experimentally observed conformational changes, is an important point supporting the validity and practical usefulness of elastic network models. In this context, our results also give the opportunity to compare the merits of the torsional and Cartesian representations. Indeed, it can be shown that the mode contributions c_α^2 evaluated from torsional conformational changes fitted with OLS are identical to the mode contributions evaluated from Cartesian conformational changes (see Methods). The agreement with the null model of protein response is thus improved from the perspective of torsional degrees of freedom, with respect to the Cartesian representation: $\langle r_{\text{null}}\rangle$ is equal to 0.63 with RRR, versus 0.51 with OLS. These results are consistent with previous studies showing that predicted low-frequency torsional modes present a better agreement with experimentally observed conformational changes, in comparison to predicted Cartesian modes of motions.[32,34] However, this improvement was shown to be conditional to the translation of torsional modes into Cartesian space via a nonlinear optimization scheme. Here, by creating a robust description of the conformational change in torsion angles space, adapted for small variations in bond lengths and valence angles, we remove the need to translate torsional modes into Cartesian space and are able to directly compare the torsional modes to the torsional representation of the conformational change.

The null model of protein response, $c_\alpha^2 \propto \omega_\alpha^{-2}$, holds for both functional and nonfunctional conformational changes, simply because of linear response theory. Deviations from this null model may be random or due to imperfections in the prediction scheme, but may also indicate specificities of functional conformational changes. In particular, if the contribution of low-frequency normal modes is larger than expected from the null model, the free-energy barrier opposed to the conformational change would be reduced, at least within the harmonic approximation. Such cases can be identified by computing the correlation $r_{\text{select}} = r(c_\alpha^2\omega_\alpha^2\omega_\alpha^{-2})$. A positive and significant r_{select} suggests that natural selection may have acted on the dynamical properties of the equilibrium ensemble so as to lower the energy barrier to overcome during a functional conformational change.[24,33] As shown in Figure 6C, this is indeed the case for almost half of the considered proteins, if the $\Delta\varphi_a$ values are identified by the RRR-Cv method. This fraction is much smaller with the other methods: as low as 14 and 0% with OLS and ST, respectively.

The reciprocal collectivity of the conformational change, i.e., the effective number of normal modes that contribute to it (see Methods), is given in Figure 6D for each method. This collectivity is significantly smaller when conformational changes are fitted by the RRR method than when using ST or OLS, which is consistent with the above results indicating that the predominance of low-frequency modes is best recovered with RRR. Note that, here again, the reciprocal collectivity with respect to Cartesian displacements is identical to the value obtained with torsional displacements fitted by OLS.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

DISCUSSION AND CONCLUSIONS

Deciphering the mechanisms that underlie functional conformational changes in proteins requires an understanding of what happens both at the level of single residues and at the level of the tertiary/quaternary structure. The former aspect is best described with torsional angles, which are the natural degrees of freedom of protein chains and are well suited to evaluate how, and how much, each residue contributes to the overall conformational change. On the other hand, Cartesian coordinates, which are the typical output of structure determination methods, provide a more intuitive global picture of protein structures and conformational variations. The translation from one set of coordinates to the other is trivial in principle, but not in practice.

In this paper, we evaluate and compare different approaches to identify variations in torsion angles that best reproduce the observed variations in Cartesian coordinates, in a set of experimentally characterized conformational change of proteins. The naive strategy ST, which consists in simply computing the differences of aligned torsion angles between the two conformations, largely fails to provide satisfactory results. Indeed, despite the fact that variations in bond lengths and valence angles are very small, these variations create errors that quickly propagate and get amplified along the chain, completely disrupting the protein fold. Linear regression with OLS allows in principle to retrieve the variations in torsion angles that minimize the differences between the reconstructed and target Cartesian coordinates, correcting for any variations in bond lengths and valence angles. However, the large correlations between explanatory variables lead to overfitting issues. The minimal error ends up being achieved by forcing large amplitude variations on a large number of torsion angles, which breaks the linear approximation on which the fit relies, and causes reconstructed structures to be almost as poor as those obtained from the ST method.

A first satisfactory solution is found in rescaled ridge regression (RRR), a variant of ridge regression that we recently introduced,[27] which is suited to the regularization of fits even when there is no intercept, or when the intercept holds physical meaning and must be penalized like the other fit parameters. This method allows to modify the torsion angles in the initial structure A and obtain a transformed structure A* that is very similar to the target structure B, effectively modeling the experimentally observed conformational change in the space of torsion angles. With RRR and the Cv criterion, RMSD(A*,B) is as low as 1.7 Å on average, compared to 16 Å with ST and 9 Å with OLS. The reconstruction of the conformational change with RRR is achieved by small-amplitude variations of a limited number of torsion angles, which is more consistent with functional motions operated by hinge regions. The fraction of torsion angles that effectively contribute to the conformational change is smaller than 5%, on average with RRR, compared to 17% with ST and 27% with OLS. We also evaluated a greedy iterative procedure based on RRR (GIS), which yields improved results, with RMSD(A*,B) equal to 0.8 Å on average, but at the cost of a more significant computational burden.

Interestingly, the conformational changes fitted via the RRR approach also provide a better agreement with the predictions of the TNM and the null model of protein response:[24,33] the contributions of the normal modes to the conformational changes are more strongly correlated with their contributions to the predicted thermal fluctuations. In addition, evidence that natural selection may have acted to reduce the energy barrier that must be overcome during the conformation change is observed in 47% of the cases when the RRR-Cv approach is used. This fraction goes down to 15% if the conformational changes are compared to Cartesian modes of motion (or torsional modes with the OLS approach) and 0% if we consider the changes in torsion angles directly obtained from the two structures (with ST). Note that the improvement achieved by the RRR method is not only due to the fact that it reduces the amplitude of the change of torsion angles, thus reducing the violations of the linear approximation on which the fit is based. In fact, RRR also improves measures that do not depend on the amplitude, such as the number of relevantly modified angles, the correlation with the normal modes predicted through the TNM, and the number of relevant modes.

Since the GIS method creates a trajectory between the two structural states of the protein, it could potentially evolve into a proper morphing technique in torsion angles space. This would require further research, notably concerning the addition of a potential function to ensure that the trajectory follows low-energy paths and avoids steric clashes. It is interesting to note that RR is equivalent to minimizing the quantity $(\Delta\varphi \cdot T\Delta\varphi) + \Lambda(\Delta\varphi \cdot \Delta\varphi) - 2(\Delta\varphi \cdot \Delta\varphi^*)$, where the first two terms describe a kind of harmonic energy associated with the conformational change $\Delta\varphi$, and $\Delta\varphi^* = \mathbf{J}^T\mathbf{M}\Delta\mathbf{r}$ is the target conformational change. However, this pseudoenergy is not sufficient for preventing steric clashes. If this limitation is overcome, GIS could present certain benefits as an interpolation technique. Indeed, most current methods are built in Cartesian coordinates, often relying on (linear or nonlinear) interpolation between atomic coordinates or between interatomic distances.[41,42] Despite the advantage of preserving the integrity of bond geometries, internal coordinates have been less exploited in this context, in part because the interplay of soft (torsion angles) and supposedly hard (bond lengths and valence angles) degrees of freedom complicates any attempt at interpolation. Another potential advantage of GIS is that the iterative steps are guided by the RRR method, which we showed to be very robust to the choice of experimental structures and which correlates well with ENM-predicted low-frequency normal modes, without being limited to a small arbitrary number of these modes. However, like most other interpolation schemes, RRR cannot be interpreted as a dynamical equation, and it is therefore unclear how well the trajectory produced by GIS coincides with the actual transition path.

In summary, our results indicate that the comparison of backbone torsion angle values between pairs of structures corresponding to conformational changes may not necessarily give a reliable picture of the mechanisms at play at the residue level. This is due to compensations between relevant modifications of torsion angles and small variations in bond lengths and valence angles, which may occur in the structure refinement process. In contrast, the RRR methodology yields a minimal set of torsional variations that is sufficient to describe the conformational change in a robust manner, allowing to easily pinpoint the most important residues and presenting a strongly improved agreement with theoretical models.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

METHODS

Data Set. The data set considered in this study consists of 31 pairs of protein structures determined by X-ray crystallography, extracted from the Protein Data Bank.[45] Each pair corresponds to two distinct structures of the same protein chain, representing a conformational change that is relevant for the protein's function. These include large molecular machines that were previously studied in ref 46, allosteric proteins that were investigated in ref 47, as well as others that were studied in our previous work.[33] The root-mean-square deviation (RMSD) between structural pairs ranges from 0.35 to 34.4 Å. Each structure in a pair is, in turn, considered as the initial conformation, and the other as the final conformation. A total of 62 conformational changes are thus analyzed.

For each conformational change, we consider only aligned atoms that are present in the two structures. Even though both structures represent the same protein, it may happen that some residues or atoms are unresolved (disordered) in one of the structures, but not in the other. Small differences in protein sequence are also possible, due to amino acid replacements, insertions, or deletions. Breaks in the considered protein chains, in the case of disordered regions or sequence mismatch, are represented by pseudobonds. For these pseudobonds, we consider all three internal coordinates as degrees of freedom, i.e., the pseudobond lengths and angles are also adjusted in the same way as torsion angles. Residues at the N- or C-terminal end of the protein chain are often minimally constrained and thus very flexible, which has little effect on the overall structure but can induce a significant bias in some measurements. Therefore, residues at the chain ends, with a B-factor (in the initial structure) at least 50% larger than the average over all residues, were not considered in the analysis of the results (e.g., calculation of the torsional collectivity, correlation with predicted fluctuations).

Ordinary Least-Squares Regression. At first order in $\Delta\varphi$, the analytical relationship between variations in Cartesian coordinates and variations in torsion angles can be written as $\Delta r_i = \sum_a J_{ia}\Delta\varphi_a$, or $\Delta\mathbf{r} = \mathbf{J}\Delta\varphi$ in matrix notation, where $\mathbf{J}$ is the $(9N \times 2N)$ Jacobian matrix of the transformation between internal and Cartesian coordinates, with $J_{ia} = \partial r_i/\partial\varphi_a$, and N is the number of residues.

An estimation of the variations of torsion angles $\Delta\varphi$ corresponding to the observed variations in Cartesian coordinates $\Delta\mathbf{r}$ can be obtained by linear regression. We minimize the square error $E = (\mathbf{J}\Delta\varphi - \Delta\mathbf{r})\cdot\mathbf{M}(\mathbf{J}\Delta\varphi - \Delta\mathbf{r})$, where each atom is weighted by its mass via the diagonal matrix $\mathbf{M}$ and $\mathbf{x}\cdot\mathbf{y}$ denotes the scalar product. The quantity to be minimized can be rewritten as

$$\begin{aligned} F &= (\mathbf{J}\Delta\varphi\cdot\mathbf{M}\mathbf{J}\Delta\varphi) - 2(\mathbf{M}\Delta\mathbf{r}\cdot\mathbf{J}\Delta\varphi) \\ &= (\Delta\varphi\cdot\mathbf{T}\Delta\varphi) - 2(\mathbf{J}^T\mathbf{M}\Delta\mathbf{r}\cdot\Delta\varphi) \end{aligned} \quad (1)$$

where $\mathbf{J}^T$ is the transpose of the Jacobian matrix and $\mathbf{T} = \mathbf{J}^T\mathbf{M}\mathbf{J}$ is the kinetic energy matrix, which plays the role of the correlation matrix of explanatory variables. The solution of this minimization problem is given by $\Delta\varphi^{OLS} = \mathbf{T}^{-1}(\mathbf{J}^T\mathbf{M}\Delta\mathbf{r})$.

Rescaled Ridge Regression. Ordinary least-square regression with correlated explanatory variables (like the components of the Jacobian matrix in this problem) is known to be bad-behaved and to lead to overfitting problems and unphysical parameters unless the fit is regularized. One of the most widely used regularization schemes, the Tykhonov regularization or ridge regression (RR), consists in penalizing large values of the fit parameters.[36] In this framework, the function to be minimized is

$$\begin{aligned} F &= E + \Lambda(\Delta\varphi\cdot\Delta\varphi) \\ &= (\Delta\varphi\cdot\mathbf{T}\Delta\varphi) - 2(\mathbf{J}^T\mathbf{M}\Delta\mathbf{r}\cdot\Delta\varphi) + \Lambda(\Delta\varphi\cdot\Delta\varphi) \end{aligned} \quad (2)$$

where E is the error of the fit, as defined above, and Λ is the Tykhonov parameter. The solution of this minimization problem is given by $\Delta\varphi^{RR} = (\mathbf{T} + \Lambda\mathbf{I})^{-1}(\mathbf{J}^T\mathbf{M}\Delta\mathbf{r})$. In the limit of large Λ, the fitted parameters $\Delta\varphi$ tend to zero and the regression sets the dependent variable equal to the intercept of the fit, which is not penalized in ordinary ridge regression.

Rescaled ridge regression (RRR) allows to deal with situations where the intercept is either absent or must be interpreted as a physical parameter and penalized like any other. To avoid the vanishing fit parameters that RR would yield in such cases, the fit parameters are rescaled by a factor that diverges in the large Λ limit so that they tend to a finite limit.[27] This is achieved via a second Lagrange multiplier μ ensuring the respect of the scaling condition $(\mathbf{J}^T\mathbf{M}\Delta\mathbf{r}\cdot\Delta\varphi) - (\Delta\varphi\cdot\mathbf{T}\Delta\varphi) = 0$. The function to be minimized in RRR becomes then

$$\begin{aligned} F &= E + (1-\mu)\Lambda(\Delta\varphi\cdot\Delta\varphi) + \mu((\mathbf{J}^T\mathbf{M}\Delta\mathbf{r}\cdot\Delta\varphi) \\ &\quad -(\Delta\varphi\cdot\mathbf{T}\Delta\varphi)) \\ &= (\Delta\varphi\cdot\mathbf{T}\Delta\varphi) - 2\nu(\mathbf{J}^T\mathbf{M}\Delta\mathbf{r}\cdot\Delta\varphi) + \Lambda(\Delta\varphi\cdot\Delta\varphi) \end{aligned} \quad (3)$$

where $\nu = (1-\mu/2)/(1-\mu)$. The optimal fit parameters of the RRR are a rescaled version of the parameters of ordinary ridge regression, i.e., $\Delta\varphi^{RRR} = \nu\Delta\varphi^{RR}$. The factor ν, which depends on Λ, ensures the proper scaling of the fit parameters $\Delta\varphi$. It is given by $\nu(\Lambda) = (\Delta\varphi^{RR}\cdot\mathbf{J}^T\mathbf{M}\Delta\mathbf{r})/(\Delta\varphi^{RR}\cdot\mathbf{T}\Delta\varphi^{RR})$.

Torsional Network Model. To compare the experimentally observed conformational changes with the thermal dynamics at equilibrium, we rely on the predictions of the torsional network model (TNM), an ENM in torsion angle space that preserves the bond lengths and valence angles of the protein.[24]

The main degrees of freedom of the TNM are the ϕ and ψ torsion angles of the protein backbone. In this study, we also include the ω torsion angles corresponding to cis/trans isomerizations observed in the conformational changes. It is possible to further extend the model and consider all ω angles, as well as the torsion angles of the side chains. The kinetic energy is computed here from the atoms of the extended backbone (N, C_α, C_β, C, and O) since we focus on conformational changes of the protein backbone and do not consider side chain rotamers. Native interactions between pair of residues are identified by considering the smallest spatial distance between any pair of heavy atoms belonging to the two residues. If that distance is smaller than a predefined cutoff C = 4.5 Å, the corresponding atoms are joined by a spring of stiffness $\kappa(r) = \kappa_0(r_0/r)^E$, where r is the equilibrium distance between the two atoms, r_0 = 3.5 Å, is a reference distance, E = 6, and κ_0 is the force constant obtained from the fit of B-factors. In addition, a harmonic potential was also associated with the rotation around each torsion angle in the protein, with a uniform value of the torsional force constant κ_φ = 0.1. The values of these parameters, which define the TNM force field, were previously determined as optimal for reproducing experimentally observed individual and pairwise residue fluctuations.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

The TNM provides as output the normal modes of motion $\mathbf{u}_\alpha$ of the molecule, along the internal coordinates corresponding to the ϕ and ψ torsion angles. These modes describe the dynamic behavior of the protein around equilibrium and can be used to evaluate the mean-square fluctuations of the torsion angles in the thermal ensemble: $(\Delta\varphi_a^{\text{therm}})^2 = \sum_\alpha \omega_\alpha^{-2} u_{\alpha a}^2$. The torsional modes can also easily be converted into Cartesian modes of motion $\mathbf{x}_\alpha = \mathbf{J}\mathbf{u}_\alpha$.

We define the contribution c_α^2 of each normal mode to a given conformational change as the normalized squared mass-weighted projection of the conformational change onto the normal mode, that is, $c_\alpha^2 = p_\alpha^2/\sum_\beta p_\beta^2$ with $p_\alpha = (\mathbf{T}^{1/2}\Delta\varphi\cdot\mathbf{T}^{1/2}\mathbf{u}_\alpha) = (\Delta\varphi\cdot\mathbf{T}\mathbf{u}_\alpha)$. Note that, when the OLS method is used to determine the $\Delta\varphi$ values, i.e., $\Delta\varphi^{\text{OLS}} = \mathbf{T}^{-1}\mathbf{J}^{\text{T}}\mathbf{M}\Delta\mathbf{r}$, the projections p_α are exactly equal to the mass-weighted projections of the Cartesian conformational change onto the Cartesian normal mode $\mathbf{J}\mathbf{u}_\alpha$, i.e., $(\Delta\mathbf{r}\cdot\mathbf{M}\mathbf{J}\mathbf{u}_\alpha) = (\mathbf{J}^{\text{T}}\mathbf{M}\Delta\mathbf{r}\cdot\mathbf{u}_\alpha) = (\Delta\varphi^{\text{OLS}}\cdot\mathbf{T}\mathbf{u}_\alpha)$.

Null Model of Protein Response. The contribution of each normal mode $\mathbf{u}_\alpha$ (or $\mathbf{x}_\alpha$ in Cartesian coordinates) to the thermal fluctuations around equilibrium is inversely proportional to the squared mode frequency ω_α^2. It has been observed that the low-frequency normal modes, which contribute more to thermal dynamics, also tend to contribute more to functional conformational changes.[30] This behavior is consistent with linear response theory.

One of us and co-workers have recently proposed that the response of a protein to a generic perturbation can be modeled as a conformational change, in which the contribution of each normal mode is proportional to its contribution to thermal dynamics, i.e., $c_\alpha^2 \propto \omega_\alpha^{-2}$. This null model of protein response was verified in an exhaustive data set of experimental conformational changes.[33] Note that this "reciprocal" correlation in the space of normal modes, $r(c_\alpha^2, \omega_\alpha^{-2})$, may be large even if the direct correlation between the conformational change, $(\Delta\varphi_a)^2$, and the thermal fluctuations, $(\Delta\varphi_a^{\text{therm}})^2$, is small. Indeed, a large reciprocal correlation requires a good prediction of the amplitude (ω_α^{-2}) of the contribution of each normal mode, but not of the direction of the motion along each normal mode. In the mean-square thermal fluctuations $(\Delta\varphi_a^{\text{therm}})^2$, the amplitude of the contributions of each mode is also proportional to ω_α^{-2}, but motions along each mode are averaged over both directions.

The null model holds for both functional and nonfunctional conformational changes, such as those observed between structures of the same protein in different experimental conditions. However, functional conformational changes often exhibit systematic deviations from the null model, attested by a positive correlation between $(c_\alpha\omega_\alpha)^2$ and ω_α^{-2}.[24,33] Such a correlation indicates that low-frequency normal modes contribute to the conformational change more than expected based on linear response theory. This has the effect of reducing the free energy barrier that must be overcome during the conformational change, as computed within the harmonic approximation and can therefore be seen as a sign of natural adaptation favoring specific functional motions.

Collectivity of Conformational Changes. The collectivity $C(\Delta\varphi)$ of a conformational change measures the number of torsional degrees of freedom that significantly contribute to it. We evaluate the fractional contribution of each torsion angle φ_a as $p_a^\varphi = (\Delta\varphi_a)^2/\sum_b(\Delta\varphi_b)^2$ and the Shannon entropy of p_a^φ, which is given by $\mathcal{S}(\Delta\varphi) = -\sum_a p_a^\varphi \log p_a^\varphi$. The collectivity is defined as the exponential of the Shannon entropy, normalized by the total number of torsional degrees of freedom N_φ, i.e., $C(\Delta\varphi) = \exp(\mathcal{S}(\Delta\varphi))/N_\varphi$. For example, $C(\Delta\varphi) = 0.3$ means that the effective fraction of torsion angles modified during the conformational change is equal to 30%, or more precisely that the Shannon entropy of p_a^φ is equivalent to a case, where $\Delta\varphi_a$ is equal to a constant value for 30% of the torsion angles and null for the remaining 70%.

The reciprocal collectivity $C(\mathbf{u})$ of a conformational change measures the number of normal modes that significantly contribute to it, and is defined in a similar way to the torsional collectivity. The fractional contribution of each normal mode u_α is given by the normalized squared projection c_α^2, as defined above. And $C(\mathbf{u}) = \exp(\mathcal{S}(\mathbf{u}))$, where $\mathcal{S}(\mathbf{u}) = -\sum_\alpha c_\alpha^2 \log c_\alpha^2$. Note that we did not normalize the reciprocal collectivity by the number of degrees of freedom since no notable correlation was observed between $\exp(\mathcal{S}(\mathbf{u}))$ and N_φ.

Angular Statistics. In case of angular data, the mean μ_φ and standard deviation σ_φ are calculated as follows[48]

$$\begin{aligned} C_\varphi &= \frac{1}{N}\sum_{a=1}^{N}\cos(\varphi_a) \\ S_\varphi &= \frac{1}{N}\sum_{a=1}^{N}\sin(\varphi_a) \\ \mu_\varphi &= \tan^{-1}\left(\frac{S_\varphi}{C_\varphi}\right)(+\pi \text{ if } C_\varphi<0) \\ R_\varphi &= \sqrt{C_\varphi^2 + S_\varphi^2} \\ \sigma_\varphi &= \sqrt{-2\ln R_\varphi}. \end{aligned} \tag{4}$$

The correlation between two sets of angles, φ_1 and φ_2, is defined as[49]

$$r_\varphi = \frac{\sum_{a=1}^{N}\sin(\varphi_{1a}-\mu_{\varphi_1})\sin(\varphi_{2a}-\mu_{\varphi_2})}{\sqrt{\sum_{a=1}^{N}\sin^2(\varphi_{1a}-\mu_{\varphi_1})\sum_{a=1}^{N}\sin^2(\varphi_{2a}-\mu_{\varphi_2})}} \tag{5}$$

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00627.

Quality of the reconstruction of the conformational changes (Table S1); robustness of the torsional representation of the conformational changes (Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: ubastolla@cbm.csic.es (U.B.).
*E-mail: ydehouck@cbm.csic.es (Y.D.).

ORCID

Ugo Bastolla: 0000-0001-9342-4678
Yves Dehouck: 0000-0002-7401-104X

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

■ ACKNOWLEDGMENTS

U.B. acknowledges financial support from the Spanish Ministery of Economy, grant BIO2016-79043-P. Research at the CBMSO is facilitated by the Fundacion Ramon Areces.

■ REFERENCES

(1) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. Intrinsic Dynamics of an Enzyme Underlies Catalysis. *Nature* **2005**, *438*, 117−121.
(2) Callender, R.; Dyer, R. B. The Dynamical Nature of Enzymatic Catalysis. *Acc. Chem. Res.* **2015**, *48*, 407−413.
(3) Petrović, D.; Risso, V. A.; Kamerlin, S. C. L.; Sanchez-Ruiz, J. M. Conformational Dynamics and Enzyme Evolution. *J. R. Soc. Interface* **2018**, *15*, No. 20180330.
(4) Goodey, N. M.; Benkovic, S. J. Allosteric Regulation and Catalysis Emerge via a Common Route. *Nat. Chem. Biol.* **2008**, *4*, 474−482.
(5) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The Ensemble Nature of Allostery. *Nature* **2014**, *508*, 331−339.
(6) DuBay, K. H.; Bowman, G. R.; Geissler, P. L. Fluctuations within Folded Proteins: Implications for Thermodynamic and Allosteric Regulation. *Acc. Chem. Res.* **2015**, *48*, 1098−1105.
(7) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* **2009**, *5*, 789−796.
(8) Gibbs, A. C. Elements and Modulation of Functional Dynamics. *J. Med. Chem.* **2014**, *57*, 7819−7837.
(9) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(10) Nymeyer, H.; Gnanakaran, S.; Garcia, A. E. Atomic Simulations of Protein Folding, Using the Replica Exchange Algorithm. *Methods Enzymol.* **2004**, *383*, 119−149.
(11) Li, W.; Wolynes, P. G.; Takada, S. Frustration, Specific Sequence Dependence, and Nonlinearity in Large-Amplitude Fluctuations of Allosteric Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 3504−3509.
(12) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48*, 414−422.
(13) Takada, S.; Kanada, R.; Tan, C.; Terakawa, T.; Li, W.; Kenzaki, H. Modeling Structural Dynamics of Biomolecular Complexes by Coarse-Grained Molecular Simulations. *Acc. Chem. Res.* **2015**, *48*, 3026−3035.
(14) Togashi, Y.; Flechsig, H. Coarse-Grained Protein Dynamics Studies Using Elastic Network Models. *Int. J. Mol. Sci.* **2018**, *19*, No. E3899.
(15) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(16) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505−515.
(17) Bahar, I.; Rader, A. J. Coarse-Grained Normal Mode Analysis in Structural Biology. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586−592.
(18) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and their Applications. *Chem. Rev.* **2016**, *116*, 7898−7936.
(19) Go, N.; Noguti, T.; Nishikawa, T. Dynamics of a Small Globular Protein in Terms of Low-Frequency Vibrational Modes. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 3696−3700.
(20) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM—A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.
(21) Güntert, P.; Mumenthaler, C.; Wüthrich, K. Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. *J. Mol. Biol.* **1997**, *273*, 283−298.
(22) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol.* **1997**, *268*, 209−225.
(23) Parsons, J.; Holmes, J. B.; Rojas, J. M.; Tsai, J.; Strauss, C. E. Practical Conversion from Torsion Space to Cartesian Space for In Silico Protein Synthesis. *J. Comput. Chem.* **2005**, *26*, 1063−1068.
(24) Méndez, R.; Bastolla, U. Torsional Network Model: Normal Modes in Torsion Angle Space Better Correlate with Conformation Changes in Proteins. *Phys. Rev. Lett.* **2010**, *104*, No. 228103.
(25) Riccardi, D.; Cui, Q.; Philips, G. N., Jr. Evaluating Elastic Network Models of Crystalline Biological Molecules with Temperature Factors, Correlated Motions, and Diffuse X-Ray Scattering. *Biophys. J.* **2010**, *99*, 2616−2625.
(26) Fuglebakk, E.; Echave, J.; Reuter, N. Measuring and Comparing Structural Fluctuation Patterns in Large Protein Datasets. *Bioinformatics* **2012**, *28*, 2431−2440.
(27) Dehouck, Y.; Bastolla, U. The Maximum Penalty Criterion for Ridge Regression: Application to the Calibration of the Force Constant in Elastic Network Models. *Integr. Biol.* **2017**, *9*, 627−641.
(28) Sun, Z.; Liu, Q.; Qu, G.; Feng, Y.; Reetz, M. T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* **2019**, *119*, 1626−1665.
(29) Dehouck, Y.; Mikhailov, A. S. Effective Harmonic Potentials: Insights into the Internal Cooperativity and Sequence-Specificity of Protein Dynamics. *PLoS Comput. Biol.* **2013**, *9*, No. e1003209.
(30) Tama, F.; Sanejouand, Y. H. Conformational Change of Proteins Arising From Normal Mode Calculations. *Protein Eng.* **2001**, *14*, 1−6.
(31) Meireles, L.; Gur, M.; Bakan, A.; Bahar, I. Pre-Existing Soft Modes of Motion Uniquely Defined by Native Contact Topology Facilitate Ligand Binding to Proteins. *Protein Sci.* **2011**, *20*, 1645−1658.
(32) Bray, J. K.; Weiss, D. R.; Levitt, M. Optimized Torsion-Angle Normal Modes Reproduce Conformational Changes more Accurately than Cartesian Modes. *Biophys. J.* **2011**, *101*, 2966−2969.
(33) Dos Santos, H. G.; Klett, J.; Méndez, R.; Bastolla, U. Characterizing Conformation Changes in Proteins Through the Torsional Elastic Response. *Biochim. Biophys. Acta, Proteins Proteomics* **2013**, *1834*, 836−846.
(34) Frezza, E.; Lavery, R. Internal Coordinate Normal Mode Analysis: a Strategy to Predict Protein Conformational Transitions. *J. Phys. Chem. B* **2019**, *123*, 1294−1301.
(35) Rueda, M.; Chacón, P.; Orozco, M. Thorough Validation of Protein Normal Mode Analysis: a Comparative Study with Essential Dynamics. *Structure* **2007**, *15*, 565−575.
(36) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55−67.
(37) Mallows, C. Some Comments on C_p. *Technometrics* **1973**, *15*, 661−675.
(38) Golub, G. H.; Heath, M.; Wahba, G. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* **1979**, *21*, 215−223.
(39) Hansen, P. C. Analysis of Discrete Ill-Posed Problems by Means of the L-Curve. *SIAM Rev.* **1992**, *34*, 561−580.
(40) López-Blanco, J. R.; Aliaga, J. I.; Quintana-Ortí, E. S.; Chacón, P. iMODS: Internal Coordinates Normal Mode Analysis Server. *Nucleic Acids Res.* **2014**, *42*, W271−W276.
(41) Weiss, D. R.; Koehl, P. Morphing Methods to Visualize Coarse-Grained Protein Dynamics. *Methods Mol. Biol.* **2014**, *1084*, 271−282.
(42) Zheng, W.; Wen, H. A Survey of Coarse-Grained Methods for Modeling Protein Conformational Transitions. *Curr. Opin. Struct. Biol.* **2017**, *42*, 24−30.
(43) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B., 3rd.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941

MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **2018**, *27*, 293−315.
(44) Nagata, K.; Randall, A.; Baldi, P. SIDEpro: a Novel Machine Learning Approach for the Fast and Accurate Prediction of Side-Chain Conformations. *Proteins* **2012**, *80*, 142−153.
(45) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(46) Zheng, W.; Brooks, B. R.; Thirumalai, D. Low-Frequency Normal Modes that Describe Allosteric Transitions in Biological Nanomachines are Robust to Sequence Variations. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7664−7669.
(47) Guarnera, E.; Berezovsky, I. N. Structure-Based Statistical Mechanical Model Accounts for the Causality and Energetics of Allosteric Communication. *PLoS Comput. Biol.* **2016**, *12*, No. e1004678.
(48) Mardia, K. V.; Jupp, P. E. *Directional Statistics*; Wiley Series in Probability and Statistics; John Wiley & Sons Ltd.: Chichester (England), 2000.
(49) Jammalamadaka, S. R.; Sengupta, A. *Topics in Circular Statistics*; Series on Multivariate Analysis; World Scientific Publishing Co. Pte. Ltd.: Singapore, 2001; Vol. 5.

DOI: 10.1021/acs.jcim.9b00627
J. Chem. Inf. Model. 2019, 59, 4929−4941