Determination of Ensemble-Average Pairwise Root Mean-Square Deviation from Experimental B-Factors

Antonija Kuzmanic[†] and Bojan Zagrovic[‡*]
[†]Laboratory of Computational Biophysics, Mediterranean Institute for Life Sciences, Split, Croatia; and [‡]Department of Physics, Faculty of Science, University of Split, Split, Croatia

ABSTRACT Root mean-square deviation (RMSD) after roto-translational least-squares fitting is a measure of global structural similarity of macromolecules used commonly. On the other hand, experimental x-ray B-factors are used frequently to study local structural heterogeneity and dynamics in macromolecules by providing direct information about root mean-square fluctuations (RMSF) that can also be calculated from molecular dynamics simulations. We provide a mathematical derivation showing that, given a set of conservative assumptions, a root mean-square ensemble-average of an all-against-all distribution of pairwise RMSD for a single molecular species, $<RMSD^2>^{1/2}$, is directly related to average B-factors ($<B>$) and $<RMSF^2>^{1/2}$. We show this relationship and explore its limits of validity on a heterogeneous ensemble of structures taken from molecular dynamics simulations of villin headpiece generated using distributed-computing techniques and the Folding@Home cluster. Our results provide a basis for quantifying global structural diversity of macromolecules in crystals directly from x-ray experiments, and we show this on a large set of structures taken from the Protein Data Bank. In particular, we show that the ensemble-average pairwise backbone RMSD for a microscopic ensemble underlying a typical protein x-ray structure is ~1.1 Å, under the assumption that the principal contribution to experimental B-factors is conformational variability.

INTRODUCTION

The most frequently used measure for structure comparison in structural biology is, arguably, the atom-positional root mean-square deviation (RMSD) obtained after roto-translational least-squares fitting (1–5). Its applications are diverse and include monitoring structural changes in simulations of protein folding and dynamics (6–12), evaluating the quality of structure prediction schemes (13–16), comparing the diversity of model structures derived from experiments (17,18), assessing the properties of modeling approaches at different levels of resolution (19,20), and defining high-resolution shapes of polymers (21). Furthermore, structural diversity of an ensemble of biomolecular structures obtained through computer simulations is analyzed frequently by calculating an all-against-all distribution of RMSD values (pairwise RMSD) (22,23). Such calculation is also carried out commonly in NMR spectroscopy to assess the mutual similarity of the lowest energy structures in an ensemble produced by the refinement process (24–26). The resulting distribution of pairwise RMSD values captures the degree of structural heterogeneity of a given ensemble that can be due to either the intrinsic flexibility of a given structure or the uncertainties of the refinement procedure. The properties of this distribution, calculated typically for backbone atoms, are often summarized by reporting its arithmetic mean. Even though the calculations of pairwise RMSD values can be computationally demanding for large ensembles, they are also frequently used as an appropriate measure for clustering of structures (7,27–29).

A distribution of pairwise RMSD values provides information on the mutual similarity of members of a given ensemble when it comes to their global structure. However, to obtain information on local structural flexibility, thermal stability, and heterogeneity of macromolecules, root mean-square fluctuations (RMSF) are often studied (30–32). Most importantly, RMSF can be obtained through Debye-Waller or temperature factors (B-factors) in x-ray experiments using Eq. 1, where B-factors are usually defined as a measure of spatial fluctuations of atoms around their average position and where their motion is described as an isotropic Gaussian distribution of displacements about the average position (33). The inverse of this equation has often been used in the literature to calculate B-factors from various models (most often molecular dynamics simulations or Gaussian network models) and to compare them to experimental values (34–42):

$$RMSF_i^2 = \frac{3B_i}{8\pi^2}. \quad (1)$$

B-factors have also been used in a variety of studies to predict protein flexibility (43,44), assess their thermal stability (45–47), test for errors in protein structures (48), analyze active sites and binding pockets (49–51), correlate side-chain mobility with conformation (52,53), investigate crystal packing contacts (54), analyze and predict protein disordered regions (55–58), and study protein dynamics (37,40,59). However useful B-factors may be, one should always keep in mind that they include not only the positional variance of macromolecules that is due to local thermal

Submitted October 12, 2009, and accepted for publication November 3, 2009.

*Correspondence: zagrovic@medils.hr

Editor: Nathan Andrew Baker.

0006-3495/10/03/0861/11 $2.00

doi: 10.1016/j.bpj.2009.11.011

motion, but also the effects of noise due to refinement errors, lattice defects, crystal contacts, and rigid-body motions (36,41,60). Furthermore, they also contain components coming from both static and dynamic disorder (61,62) whose separation is nontrivial (36). Finally, RMSF can also be predicted from NMR chemical shifts via a measure called random coil index (63–65).

Because pairwise RMSD, B-factors (or RMSF) are all frequently used to give information on different aspects of biomolecular ensembles, we study their relationship. We present a derivation showing that, given a set of conservative assumptions, $<RMSD^2>^{1/2}$ is directly proportional to average experimental B-factors ($<B>$), i.e., $<RMSF^2>^{1/2}$ for a single molecular species. Our finding is illustrated and its limits of validity probed by calculations made on structures taken from molecular dynamic (MD) simulations of the native and unfolded state of the villin headpiece domain (10,66) generated using worldwide-distributed computing techniques. In particular, we use simulated ensembles to study the effects of the exact method of structure alignment on the derived relationship, and show that the influence is typically only marginal. Finally, the newly derived relation is used to calculate quadratic means of pairwise RMSD distributions for a set of x-ray structures, given the B-factors reported in the Protein Data Bank (PDB), to assess their heterogeneity in the crystal environment.

To foreshadow the derivation presented in this study, we would like to introduce a useful analogy between $<RMSD^2>^{1/2}$ and $<RMSF^2>^{1/2}$ on the one hand and the radius of gyration (R_g) on the other. The radius of gyration, a measure often used to describe the dimensions of biopolymers such as proteins (67–69), can be analytically calculated in two ways: one using the pairwise distances between monomers (Eq. 2, Fig. 1 *A*), and the other using the distances between each monomer and their center of mass, i.e., the average position of all monomers if they have the same mass (Eq. 3, Fig. 1 *B*).

$$R_g^2 = \frac{1}{2N_m^2}\sum_{i=1}^{N_m}\sum_{j=1}^{N_m}\|\vec{r}_i - \vec{r}_j\|^2, \tag{2}$$

$$R_g^2 = \frac{1}{N_m}\sum_{i=1}^{N_m}\|\vec{r}_i - \langle\vec{r}\rangle\|^2. \tag{3}$$

Indices i and j refer to different monomers, whereas N_m is a total number of monomers in a chain. Vector $\vec{r}$ represents spatial coordinates of a monomer, whereas $\langle\vec{r}\rangle$ is the average position of N_m monomers. Equations 2 and 3 are shown to be identical by modifications of the Lagrange's theorem (70).

Our derivation of the relationship between $<B>$, $<RMSF^2>^{1/2}$, and $<RMSD^2>^{1/2}$, which is the main result of this study, mirrors the relationship between these two definitions of the radius of gyration. Namely, the two definitions given for R_g can be applied easily to ensembles of biomolecular structures where monomers are replaced by structures and RMSD is used as a measure of distance between them

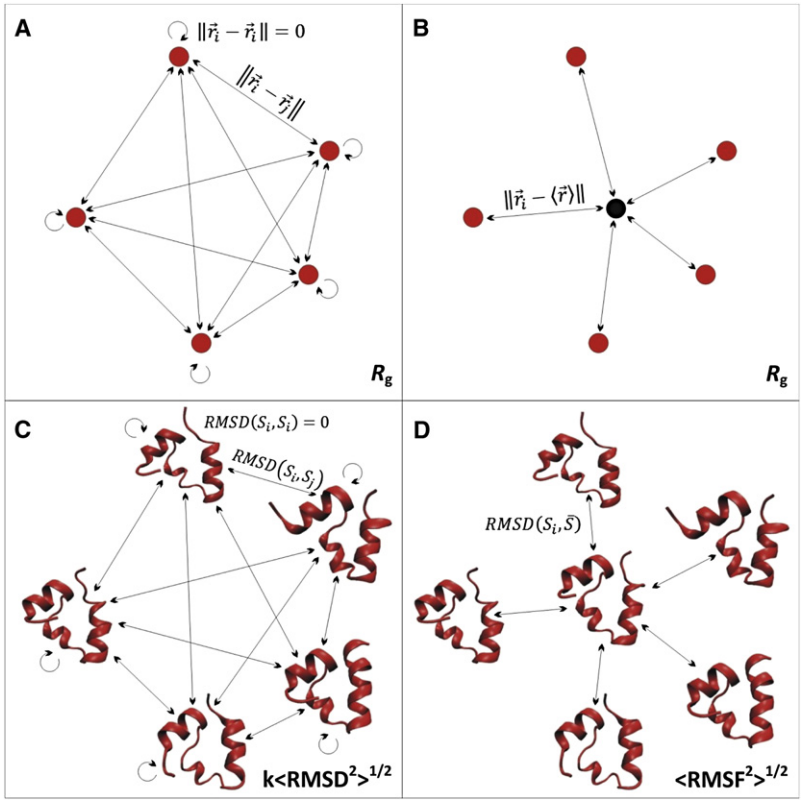

FIGURE 1 Analogy connecting $<RMSD^2>^{1/2}$ and $<RMSF^2>^{1/2}$ with the radius of gyration. (*A*) For a polymer consisting of N_m monomers, the radius of gyration can be calculated as a root mean-square average over all pairwise distances between monomers as shown in Eq. 2. (*B*) Another way of calculating the radius of gyration is through distances between monomers and their average position shown in Eq. 3. Analogously to the two ways of calculating R_g, there is equivalence between the root mean-square average of pairwise RMSD for a set of structures ($<RMSD^2>^{1/2}$) (*C*) and the root mean-square average deviation from the average structure ($<RMSF^2>^{1/2}$) (*D*), as shown in this study. k is a multiplicative factor that is a function of N_s (see Eq. 19). Villin structures in *C* and *D* have been prepared by VMD v1.8.6 (85).

(Eq. 2, Fig. 1, *C* and *D*). Following this analogy, there is equivalence between root mean-square average pairwise RMSD, $<RMSD^2>^{1/2}$ (recalling the first definition of R_g, see Eq. 2) and the root mean-square average deviation between each structure and the average structure of the ensemble, $<RMSF^2>^{1/2}$ (recalling the second definition of R_g, see Eq. 3). The exact relationship between $<RMSD^2>^{1/2}$, $<RMSF^2>^{1/2}$, and $<B>$ is explored below, together with an analysis of a novel measure of structural diversity in ensembles, the structural radius (R_{struct}), which can be thought of as a structural analog of R_g.

MATERIALS AND METHODS

Molecular dynamics simulations

Thousands of tens of nanoseconds long, independent trajectories for the villin headpiece domain were generated using a heterogeneous computer cluster as a part of the ongoing Folding@Home distributed computing project (10,66). The folding simulations were initiated from fully extended conformations ($\varphi = -135°$, $\psi = 135°$) with N-acetyl and C-amino caps. The equilibrium simulations were started from the experimental NMR structure of the molecule (PDB code 1VII, average structure) (66). The simulations, run using Tinker biomolecular simulation package, involved Langevin dynamics in implicit GB/SA solvent (71) (velocity damping parameter of $\gamma = 91$ ps^{-1}) with a 2-fs integration step, at 300 K. Bond lengths were constrained using RATTLE (72). No cutoffs were used for electrostatics. The protein was modeled using the OPLSua force field (73). The molecule in the equilibrium simulations was stable with respect to both secondary and tertiary structure (10,74).

The structures were divided into two data sets for calculations: one that included native-like structures (1543 structures taken from the same number of independent equilibrium simulations at $t = 20$ ns), whereas the other one contained unfolded structures 5213 structures taken from the same number of independent folding trajectories at $t = 27$ ns).

RMSD calculations—pairwise alignment

To illustrate the relationship between average RMSD and RMSF for ensembles spanning a large range of average RMSD values, we used a clustering procedure on the two villin data sets. The main purpose of this procedure was to derive a set of mutually different distributions of pairwise RMSD to help us illustrate and assess the properties of the derivation provided in this study. Backbone atoms for each pair of structures from both simulated data sets were optimally aligned (pairwise alignment (PA)) before RMSD calculations. Nonweighted pairwise RMSDs were then calculated for the aligned backbone atoms that included C, N, and C_α of every residue (108 atoms in total). A distribution of the calculated pairwise RMSD values was plotted and divided into 20 equal segments between the smallest and the largest RMSD value. The structure that appeared in the highest number of pairs in a given segment was chosen as the center of a cluster, and the structures paired with it were assigned to that particular cluster as well. Twenty clusters were obtained through such a procedure for each data set (number of structures in each cluster is listed in Table S1 in the Supporting Material). Nonweighted pairwise RMSD was calculated for each cluster in the same way as described above using backbone-based PA. Quadratic mean of pairwise RMSD for each cluster was calculated as well.

We have noticed that the choice of the reference structure for the alignment of all the structures before calculating RMSF does affect its quadratic value in the very heterogeneous data set as shown in the Results. Therefore, RMSF for each cluster was calculated by using every single structure from the cluster for the alignment before the calculations and then quadratically averaging the obtained values to get the RMSF value for each cluster. All the alignments and calculations were done by using GROMACS-3.3 and its routines (75).

RMSD calculations—reference structure alignment

Backbones of all the structures were aligned to the backbone of the native structure of the villin headpiece domain taken from the PDB (average NMR structure, PDB code 1VII) (reference structure alignment (RSA)) to rule out the alignment effect from the calculations. Fitted structures were then subjected to the same procedure described in the previous section to obtain clusters (number of structures in each cluster is also listed in Table S1) and quadratic averages of RMSD and RMSF values. Structures were aligned to the reference structure using the McLachlan algorithm (76) as implemented in the program ProFit v3.1 (Martin, A.C.R., http://www.bioinf.org.uk/software/profit/).

RESULTS

Demonstration of a direct proportionality between $<RMSD^2>^{1/2}$ and $<RMSF^2>^{1/2}$

RMSD is defined as the root mean-square-average distance between atoms of two optimally superimposed macromolecules (S_i and S_j) and is calculated as a minimum over all rotations and translations of one of the structures being compared (Eq. 4).

$$RMSD(S_i, S_j) = \min\left(\frac{1}{N_a}\sum_{k=1}^{N_a}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2\right)^{\frac{1}{2}}_{rot,trans}, \quad (4)$$

where N_a is the number of atoms in a structure and should not be confused with the Avogadro constant. Indices *i* and *j* refer to different structures, whereas the index *k* refers to the atom position in a given structure. Vector $\vec{r}$ represents spatial coordinates of a given atom.

To capture the properties of a distribution of pairwise RMSD, in this study we have used its quadratic mean calculated using Eq. 5 as it lends itself to better analytic manipulation compared to the arithmetic mean that is usually reported in NMR studies.

$$\langle RMSD^2\rangle^{1/2} = \sqrt{\frac{2}{N_s(N_s-1)}\sum_{i=1}^{N_s-1}\sum_{j>i}^{N_s} RMSD^2(S_i, S_j)}, \quad (5)$$

where N_s is the number of structures in an ensemble. Here and in the rest of the derivation we will assume that all the structures are aligned to the same reference structure (therefore, the notation from Eq. 4 was simplified). Note that for the derivation it is not relevant what the exact nature of the reference structure is, as long as the same structure is used for aligning the whole ensemble. This is to be contrasted with typical calculation of pairwise RMSD, where each pair of structures is mutually superimposed.

RMSF for a specific number of structures is defined as a root mean-square-average distance between an atom and its average position in a given set of structures (Eq. 6)

$$RMSF_k = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s} \|\vec{r}_{ik} - \langle\vec{r}\rangle_k\|^2}, \quad (6)$$

where $\langle\vec{r}\rangle_k$ is the average position of the atom k over N_s structures (Eq. 7)

$$\langle\vec{r}\rangle_k = \frac{1}{N_s}\sum_{i=1}^{N_s}\vec{r}_{ik}. \quad (7)$$

In the following, we have used the quadratic mean of RMSF calculated using Eq. 8

$$\langle RMSF^2\rangle^{1/2} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a} RMSF_k^2}. \quad (8)$$

If Eq. 6 is inserted into Eq. 8,

$$\langle RMSF^2\rangle^{1/2} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a}\frac{1}{N_s}\sum_{i=1}^{N_s}\|\vec{r}_{ik} - \langle\vec{r}\rangle_k\|^2}. \quad (9)$$

On the other hand, if Eq. 4 is inserted into Eq. 5,

$$\langle RMSD^2\rangle^{1/2} = \sqrt{\frac{2}{N_s(N_s-1)}\sum_{i=1}^{N_s-1}\sum_{j>i}^{N_s}\frac{1}{N_a}\sum_{k=1}^{N_a}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2}. \quad (10)$$

The sums in Eq. 10 can be rearranged

$$\langle RMSD^2\rangle^{1/2} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a}\frac{2}{N_s(N_s-1)}\sum_{i=1}^{N_s-1}\sum_{j>i}^{N_s}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2}. \quad (11)$$

The sums over N_a can now be written

$$\langle RMSD^2\rangle^{1/2} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a}\frac{2N_s^2}{N_s(N_s-1)}\frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2}. \quad (12)$$

For simplicity, let us define a new variable

$$R_k^2 = \frac{1}{N_s}\sum_{i=1}^{N_s}\|\vec{r}_{ik} - \langle\vec{r}\rangle_k\|^2 = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\vec{r}_{ik}^2 - 2\vec{r}_{ik}\langle\vec{r}\rangle_k + \langle\vec{r}\rangle_k^2\right)$$
$$= -\langle\vec{r}\rangle_k^2 + \frac{1}{N_s}\sum_{i=1}^{N_s}\vec{r}_{ik}^2. \quad (13)$$

The first term on the right-hand side of Eq. 13 can be separated by applying Eq. 7 and the second term can be represented as a double summation over the number of structures

$$R_k^2 = -\left(\frac{1}{N_s}\sum_{i=1}^{N_s}\vec{r}_{ik}\right)\left(\frac{1}{N_s}\sum_{j=1}^{N_s}\vec{r}_{jk}\right) + \frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\left(\vec{r}_{ik}^2 + \vec{r}_{jk}^2\right). \quad (14)$$

Terms in Eq. 14 can be added

$$R_k^2 = \frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\left(-2\vec{r}_{ik}\vec{r}_{jk} + \vec{r}_{ik}^2 + \vec{r}_{jk}^2\right)$$
$$= \frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2. \quad (15)$$

By combining Eqs. 13 and 15, we now see that

$$R_k^2 = \frac{1}{N_s}\sum_{i=1}^{N_s}\|\vec{r}_{ik} - \langle\vec{r}\rangle_k\|^2 = \frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2. \quad (16)$$

Using the former definition of R_k^2 in Eq. 16, we can express $<RMSF^2>^{1/2}$ (Eq. 9) as

$$\langle RMSF^2\rangle^{1/2} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a} R_k^2}. \quad (17)$$

Furthermore, we can express $<RMSD^2>^{1/2}$ (Eq. 12) as

$$\langle RMSD^2\rangle^{1/2} = \sqrt{\frac{1}{N_a}\frac{2N_s}{N_s-1}\sum_{k=1}^{N_a} R_k^2}. \quad (18)$$

Finally, combining Eqs. 1, 17, and 18, a formula is derived that proves that $<RMSD^2>^{1/2}$ is directly proportional to $<RMSF^2>^{1/2}$ and, subsequently, B-factors.

$$\langle RMSD^2\rangle^{1/2} = \sqrt{\frac{2N_s}{N_s-1}}\langle RMSF^2\rangle^{1/2}$$
$$= \sqrt{\frac{2N_s}{N_s-1}\frac{1}{N_a}\sum_{k=1}^{N_a}\frac{3B_k}{8\pi^2}}. \quad (19)$$

Finally, for $N_s \gg 1$,

$$\langle RMSD^2\rangle^{1/2} \approx \sqrt{\frac{2}{N_a}\sum_{k=1}^{N_a}\frac{3B_k}{8\pi^2}}. \quad (20)$$

Exclusion of the ensemble size effect and the derivation of identity

Typical calculations of the average pairwise RMSD for a given ensemble exclude the RMSDs between the same structures (equaling zero). One can show (see below) that this causes the relationship between the average RMSD and RMSF (and subsequently B-factors) to depend on the number of structures in an ensemble as in Eq. 19. This effect, as seen in Eq. 20, vanishes only for a large number of structures in the ensemble. One can define a new measure, R_{struct} (that we term structural radius), using the following equation instead of Eq. 5 to eliminate the aforementioned effect:

$$R_{\text{struct}} = \sqrt{\frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s} RMSD^2(S_i, S_j)}. \quad (21)$$

Combining Eqs. 4 and 21, it follows

$$R_{\text{struct}} = \sqrt{\frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\frac{1}{N_a}\sum_{k=1}^{N_a}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2}. \quad (22)$$

The sums in Eq. 22 can be rearranged

$$R_{\text{struct}} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a}\frac{1}{2N_s^2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\|\vec{r}_{ik} - \vec{r}_{jk}\|^2}. \quad (23)$$

Equation 16 can be inserted into Eq. 23

$$R_{\text{struct}} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a} R_k^2}. \quad (24)$$

The calculation of the average RMSD and RMSF remains the same as in the previous section, so by combining Eqs. 1, 17, 20, and 24, it follows

$$R_{\text{struct}} = \langle RMSF^2\rangle^{1/2} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a}\frac{3B_k}{8\pi^2}} \approx \frac{1}{\sqrt{2}}\langle RMSD^2\rangle^{1/2}. \quad (25)$$

With Eq. 25 we have shown that R_{struct} and $<\text{RMSF}^2>^{1/2}$ are identical and can be linked directly with both $<\text{RMSD}^2>^{1/2}$ and experimental B-factors. In this sense, R_{struct} serves as a measure of structural diversity of an ensemble of structures that is in an intuitively clear fashion directly related to B-factors, RMSF, and RMSD.

Illustrations of the relationship between the average RMSD and RMSF

To demonstrate the derived relationship between the average values of pairwise RMSD and RMSF, structures taken from distributed-computing MD simulations of villin headpiece domain were used to calculate the two measures. Each data set (native and unfolded) was divided into two sets of 20 clusters based on the distributions of pairwise RMSD values for two types of alignment (PA or reference structure alignment (RSA)). For RSA, all structures were first roto-translationally aligned to a common reference structure before their pairwise RMSD values were calculated. For PA, each individual pair of structures was first optimally aligned before calculating their RMSD. Fig. 2 shows these distributions, their arithmetic means and standard deviations. For the native data set, we can see that distributions of pairwise RMSD are very much alike regardless of the type of the alignment. Their arithmetic means and standard deviations are also very similar: 4.02 ± 1.95 Å for the PA curve and 4.07 ± 2.01 Å for the RSA curve. On the other hand, distributions associated with the unfolded data set and generated with different types of alignment are more different that can be seen from their arithmetic means and standard deviations: 7.38 ± 1.23 Å for the PA curve and 7.86 ± 1.39 Å for the RSA curve. From the given values, it can be seen that RMSD values calculated after aligning structures to a reference structure are higher than the ones calculated after the pairwise alignment. This is, of course, expected as for each individual pair of structures, PA gives by definition the lowest values over all possible roto-translational fittings. To further demonstrate this point, in the inset of Fig. 2 we compare RMSD values calculated using both types of alignment for several hundred randomly chosen pairs of structures from both data sets. It is clear from the inset of Fig. 2 that none of the RMSD values calculated after pairwise alignment of

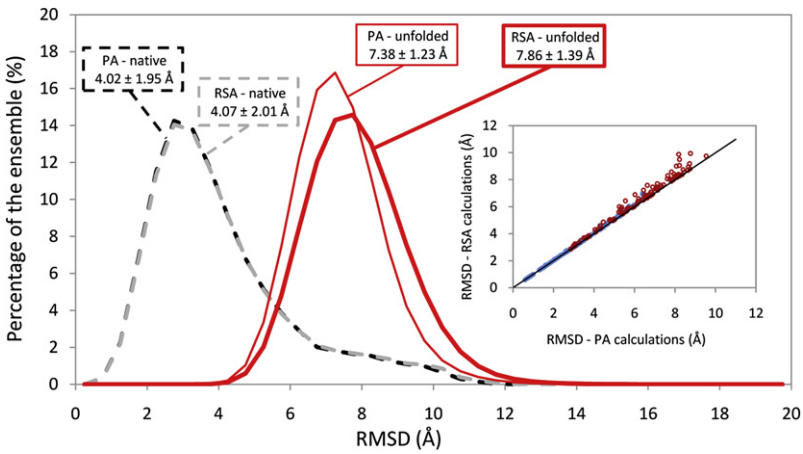

FIGURE 2 Distributions of backbone pairwise RMSD values of the native and unfolded ensembles of the villin headpiece domain. The dashed black curve represents backbone RMSD values calculated using PA for the native ensemble (<RMSD> = 4.02 ± 1.95 Å). Values for the unfolded ensemble are shown with the thin curve (<RMSD> = 7.38 ± 1.23 Å). Distributions of the corresponding RMSD values calculated using initial alignment to a reference structure (1VII, average structure) (RSA), are shown with the dashed gray curve for the native ensemble (<RMSD> = 4.07 ± 2.01 Å) and with the thick curve for the unfolded ensemble (<RMSD> = 7.86 ± 1.39 Å). The average values and standard deviations for every distribution are also given in the figure. All the values were binned in 0.5 Å bins to generate the distributions. *Inset*: Relationship between pairwise RMSD values calculated using PA and RSA. For clarity, several hundred randomly chosen points whose values have been taken from the native ensemble are shown as solid circles, whereas several hundred points whose values have been taken from the unfolded ensemble are shown as open circles. The identity line is shown in black.

structures is higher than the corresponding values calculated after aligning structures to a reference structure.

For every pair of structures belonging to a particular cluster, pairwise RMSD (Eq. 4) was calculated using either PA or RSA, and the quadratic mean of all the RMSD values (Eq. 5) in the given cluster was determined. RSA-calculations were used to study the effect of the optimal alignment on the derived relationship. RMSF for backbone atoms in the cluster was computed as well and its quadratic mean was calculated (Eq. 8). The average values of pairwise RMSD and RMSF for clusters of both data sets and types of alignment are presented in Fig. 3. As can be seen, in the case of RSA, the real data completely agrees with the analytical derivation above (Fig. 3 *A*). The slope of the trendline shown in Fig. 3 *A* is in complete agreement with the expected value of $2^{1/2}$ (seen from the derivation) and its squared correlation coefficient (R^2) equals 1. However, in the case of PA (Fig. 3 *B*), the slope of the trendline (1.2968) deviates from the expected value due to the pairwise alignment of structures before calculations, but the R^2 still has a high value of 0.9956. Deviations caused by the pairwise alignment are smaller for the native data set and the trendline applied to it would have a slope of 1.3862 with an R^2 of 0.9998 (not shown).

The relationship shown between RMSD and RMSF explored in Fig. 3 still depends on the number of structures contained in a cluster. That effect can be eliminated by using the structural radius (Eq. 21) as a measure of structural diversity. The calculated values of the structural radius and $<RMSF^2>^{1/2}$ for clusters of both data sets and types of alignment are presented in Fig. 4. In the case of RSA (Fig. 4 *A*), both slope of the applied trendline and its squared correlation coefficient equal 1 showing that the two measures

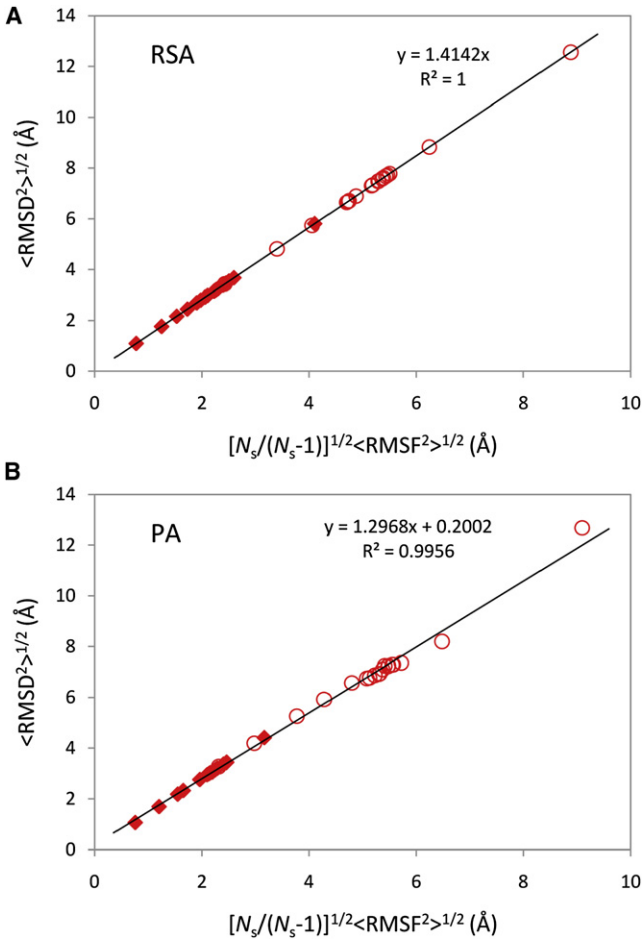

FIGURE 3 $<RMSD^2>^{1/2}$ versus $[N_s/(N_s\text{-}1)]^{1/2}<RMSF^2>^{1/2}$ for native and unfolded state clusters for the villin headpiece domain. $<RMSD^2>^{1/2}$ values were calculated using Eq. 5, whereas $<RMSF^2>^{1/2}$ values using Eq. 8. N_s is the number of structures contained in a given cluster. The values for the native ensemble are shown as solid diamonds, and the values for the unfolded ensemble are shown as open circles. In *A*, all the values have been calculated using RSA, whereas the values in *B* have been calculated using PA. Trendlines, their analytic expressions, and R^2 are also shown in the figure.

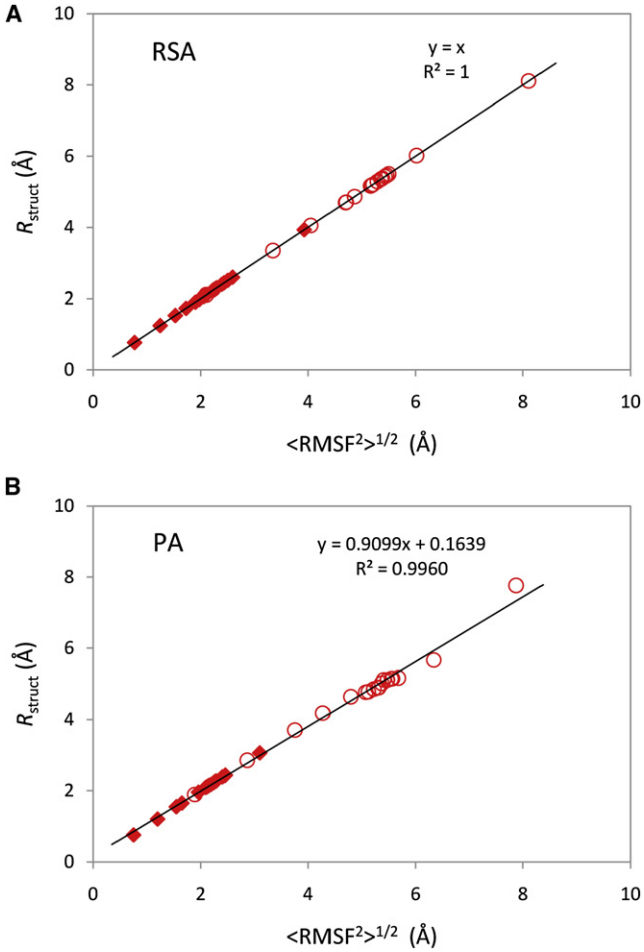

FIGURE 4 R_{struct} versus $<RMSF^2>^{1/2}$ for native and unfolded state clusters for the villin headpiece domain. R_{struct} values were calculated using Eq. 21, whereas $<RMSF^2>^{1/2}$ values using Eq. 8. The values for the native ensemble are shown as solid diamonds, and the values for the unfolded ensemble are shown as open circles. In *A*, all the values have been calculated using RSA. The values in *B* have been calculated using PA. Trendlines, their analytic expressions, and R^2 are also shown in the figure.

are identical and confirming the derivation (Eq. 25). The agreement of the two measures is slightly worse in Fig. 4 *B*due to the pairwise alignment of structures before calculations and the slope of the trendline (0.9099) differs from the expected value of 1, but still has a very high R^2 of 0.9960. Once more, deviations caused by the pairwise alignment are smaller for the native data set and the trendline applied to it would have a slope of 0.9804 with R^2 of 0.9999 (not shown). Altogether, one can claim that the effects of the alignment are negligible (<2%) for ensembles with $<RMSD^2>^{1/2}$ under 4 Å.

We also examined how the choice of the reference structure for alignment affects the $<RMSF^2>^{1/2}$ values, by using every single structure in a given cluster for the alignment before calculations and comparing the obtained distributions of $<RMSF^2>^{1/2}$ values for every cluster through their arithmetic means and standard deviations. We have also analyzed the maximal and the minimal quadratic means of every cluster to show how extreme the effects of the choice of the reference structure for the alignment can be (Fig. 5). As can be seen from the figure, arithmetic means of both data sets are very close to the minimal values of quadratic means. Their standard deviations are also quite small and they do not exceed the value of 0.26 Å, but they are higher for the unfolded data set. However, for some of the clusters, the difference between the maximal and the minimal value of $<RMSF^2>^{1/2}$ is more than twofold (e.g., clusters 15 and 18 in the native ensemble) that implies that the choice of the structure for the alignment can make a significant difference for a very diverse data set such as this.

Structural heterogeneity of proteins in crystals

The above relationship between experimental B-factors, $<RMSD^2>^{1/2}$ and the structural radius provided us with an opportunity to calculate the latter ($<RMSD^2>^{1/2}$ and R_{struct}) for an ensemble of structures in a crystal using Eq. 25 and the measured B-factors, thus assessing the heterogeneity of a given crystal. The calculations were made under the assumption that the crystal contains a very large number of structures ($N_s \gg 1$) that eliminated cluster size effect. Distributions of $<RMSD^2>^{1/2}$ and R_{struct} values for backbone and all atoms for a representative set of x-ray structures from the PDB with ~4800 structures are presented in Fig. 6 (see the Supporting Material for selection criteria). $<RMSD^2>^{1/2}$ values for the backbone and all atoms are quite similar: 1.07 ± 0.23 Å for the backbone and 1.13 ± 0.23 Å for all atoms with the maximum values that are <2.5 Å, but still with >6% of structures with $<RMSD^2>^{1/2}$ >1.5 Å for all atoms. The structural radius values are lower than $<RMSD^2>^{1/2}$ values, but are still similar: 0.76 ± 0.17 Å for the backbone and 0.80 ± 0.16 Å for all atoms with the maximum values that do not exceed the value of 1.8 Å.

DISCUSSION

To the best of our knowledge, the heterogeneity of biomolecular ensembles in crystals used in x-ray experiments has never been evaluated previously on the level of pairwise RMSD. Here, we have derived and illustrated a nontrivial relationship between ensemble-average RMSD and RMSF and, subsequently, isotropic B-factors that gave us the opportunity to evaluate the heterogeneity of the microscopic ensembles underlying the typical crystal structures deposited in the PDB. When these values (Fig. 6) are compared to the values for villin headpiece obtained through simulation (Figs. 3 and 4), we can easily see that even the highest values for proteins in a crystal are rather small and coincide with the values for the native data set in the villin graphs, for which the relationship between $<RMSD^2>^{1/2}$ and $<RMSF^2>^{1/2}$ is close to exact, regardless of the alignment. The crystal lattice aligns the protein structures to a significant degree such that the RSA would likely be a valid approximation, but it is reassuring to see that the typical values for $<RMSD^2>^{1/2}$ in crystals occupy the regime in which the choice of alignment makes very little, if any, difference. It is our estimate that if one calculates $<RMSD^2>^{1/2}$ from B-factors as in Eq. 20, the error committed is <2% on average compared to the ideal-case pairwise alignment. Namely, 2% is the average deviation between the $<RMSD^2>^{1/2}$ values obtained using RSA and PA for all villin structures below $<RMSD^2>^{1/2}$ of 4 Å (Fig. 3). It has been proposed recently

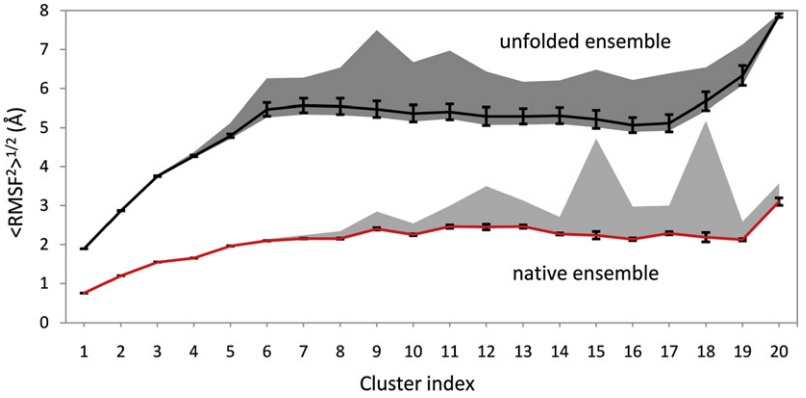

FIGURE 5 Arithmetic mean, standard deviation and extreme values of quadratic means of RMSF for every cluster. $<RMSF^2>^{1/2}$ values for every cluster were calculated using Eq. 8, using every structure in the cluster for the alignment before calculation. The lower curve shows arithmetic means of the distributions of quadratic means for native structure clusters. The upper line captures the values for the unfolded structures. Standard deviations of the distributions are shown with black bars. Extreme values and their differences are shown with a light gray area for the native ensemble and with a dark gray area for the unfolded one.

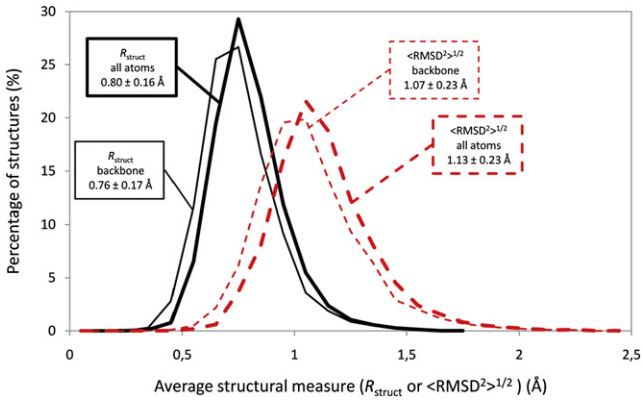

FIGURE 6 Distributions of root mean-square pairwise RMSD values calculated for x-ray structures. R_{struct} values calculated using Eq. 25 are represented by the thin solid curve for the backbone atoms with the average of 0.76 ± 0.17 Å and the thick solid curve for all the atoms in the structure (0.80 ± 0.16 Å). $<RMSD^2>^{1/2}$ values calculated using Eq. 20 are represented by the thin dashed curve for backbone atoms (1.07 ± 0.23 Å) and the thick dashed curve for all the atoms in the structure (1.13 ± 0.23 Å). Average values and standard deviations for every distribution are also shown in the figure. All the values were binned in 0.1 Å bins to generate the distributions.

that a single crystallographic structure deposited in the PDB is not enough to assess the heterogeneity of a crystal and that an ensemble of models would be a more suitable representation (77). Nevertheless, we feel that the identity we have shown in this study is a good starting point for the assessment of crystal heterogeneity that could be generalized to an ensemble of models in a straightforward manner. Finally, the derivation presented here could in future research potentially be generalized to include anisotropic B-factors as well. Due to the additional information present, the associated $<RMSD^2>^{1/2}$ values would likely be more informative in that case.

Given the analogy between the structural radius and the radius of gyration and our derivation, it becomes obvious that the structural radius can be used for an ensemble of macromolecular structures in the same sense as the radius of gyration is used for a single macromolecular structure. Although the radius of gyration provides information on a macromolecule's size, the structural radius tells how diverse structures in a given ensemble are on a global scale. The structural radius can be calculated either by using pairwise RMSD as a measure of distance between structures (Eq. 21), or by using RMSD values between each structure and the average structure of the ensemble. The latter corresponds to $<RMSF^2>^{1/2}$ as shown in Eq. 9, but with a difference of first summing over the number of atoms (N_a), and then over the number of structures (N_s), which is actually identical to Eq. 9 because the sums are interchangeable. Note that if the structural radius is calculated in this way, its usage should be restricted to molecular ensembles whose $<RMSD^2>^{1/2}$ is $\lesssim$6 Å (for which the alignment effects are $\lesssim$3%). For more heterogeneous ensembles, we would advise to either use the PA approach or simply exercise caution when interpreting results because of the potential deviations at high RMSD values. Here, it should be mentioned that it has been shown previously that the sum of squared distances for all atomic pairs equals the sum of squared distances to the average structure (78). Even though the authors suggested that this connection could be used for speeding up RMSD calculations (as the number of RMSD evaluations is reduced from $N_s(N_s - 1)/2$ to N_s) and improving algorithms in multiple structure alignment, they made no explicit link between RMSD, RMSF, and B-factors.

An important challenge in quantifying the relationship between RMSF and RMSD is the influence of optimal alignment of two structures on their mutual RMSD value. All the RMSD values calculated after the optimal pairwise alignment of two structures (PA) are lower than the ones calculated after the initial alignment of all the structures to a reference (RSA). Optimal alignment means that roto-translational fitting of the structures is carried out to minimize the RMSD value between them. Now, if one optimally aligns two structures to a reference structure, those two structures will not be optimally aligned with each other and their mutual RMSD will not be minimal, unless all three structures are mutually highly similar. For example, the effect of the optimal alignment is noticeably lesser in the native ensemble of villin than in the unfolded one, because 1), the unfolded ensemble is much more heterogeneous than the native one, and 2), the structure used as a reference is the native form of the villin headpiece domain taken from the PDB. However, as shown here, the effects of the alignment are typically only marginal. Parenthetically, one way of avoiding roto-translational alignment altogether would be to use internal coordinates to represent biomolecules and assess their structural heterogeneity. Nevertheless, as biomolecular structures (including the associated B-factors) are refined in Cartesian coordinates, we believe that the most natural measure for evaluating their global structural diversity directly from experiment should also involve atom-positional Cartesian representation, such as in the case of RMSD. In fact, the mathematical simplicity of the connection between B-factors, RMSF, and RMSD described in this study actually serves as indirect evidence supporting this claim.

In addition, RMSF values are also affected by the choice of the reference structure for the alignment. In Fig. 5, we show the arithmetic mean and standard deviation of distributions of $<RMSF^2>^{1/2}$ values calculated for every cluster, but differing in the choice of the structure used for the alignment before calculations. Even though the standard deviations throughout the clusters are quite small and they never exceed the value of 0.26 Å, the differences between the maximal and minimal values of quadratic means in some clusters are twofold (Fig. 5), which suggests that the choice of a reference structure in certain rare cases could indeed influence the outcome significantly. Contrary to this finding, Yang et al. (39) found no major effect of the choice of the alignment structure in their studies where they calculated residual

RMSD of the ensembles of NMR models. We explain this discrepancy with the greater diversity of structures in our data set compared to most ensembles of NMR models, and we would like to stress the necessity of evaluating the diversity of an ensemble before ruling out the possible effect of the choice of the reference structure.

Here, we would like to emphasize that all of our conclusions involving B-factors and other nonprimary data depend on several critical assumptions. The danger of using derived data, such as B-factors, lies in the inaccuracies of the refinement processes linking the primary data from x-ray crystallography and NMR with model structures. All the structures submitted to the PDB are based on time and ensemble-average signals that undoubtedly affect the nature of the models derived from them and potentially cause different artifacts to appear (79,80). Furthermore, it is hard to tell whether the refinement has been conducted using the state-of-the-art software at the time of deposition and whether the software has been used in an optimal manner (81). For that reason, there have been several re-refinement attempts that yielded improved structural models (81–83). Until all the artifacts are fully resolved and understood, making assumptions based on the comparison of simulations and secondary or derived data can result in overinterpreted or misinterpreted conclusions (84). Nonetheless, we find the interpretation of our results to be useful for determining the $<RMSD^2>^{1/2}$ of an ensemble of structures in a crystal from B-factors as long as the Eq. 1 holds, i.e., as long as the major contribution to the measured B-factors is indeed the structural heterogeneity of molecules.

SUPPORTING MATERIAL

One table is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(09)01738-X.

We thank Ivo F. Sbalzarini, Christian L. Müller, and the members of the Laboratory of Computational Biophysics at MedILS for useful comments on the manuscript. Contribution of Folding@Home members is gratefully acknowledged.

This work was supported in part by the National Foundation for Science, Higher Education and Technological Development of Croatia (EMBO Installation grant to B.Z.), the Unity Through Knowledge Fund (UKF 1A to B.Z.), and a National Institutes of Health R01-GM062868 grant (Folding@Home).

REFERENCES

1. McLachlan, A. D. 1972. Mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. A.* 28:656–657.
2. Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A.* 32:922–923.
3. Kabsch, W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A.* 34:827–828.
4. Kneller, G. R. 1991. Superposition of molecular structures using quaternions. *Mol. Simul.* 7:113–119.
5. Kneller, G. R. 2005. Comment on "Using quaternions to calculate RMSD" [J. Comp. Chem. 25, 1849 (2004)]. *J. Comput. Chem.* 26:1660–1662.
6. Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science.* 282:740–744.
7. Daura, X., B. Jaun, …, A. E. Mark. 1998. Reversible peptide folding in solution by molecular dynamics simulation. *J. Mol. Biol.* 280:925–932.
8. Daura, X., W. F. van Gunsteren, and A. E. Mark. 1999. Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. *Proteins.* 34:269–280.
9. Zagrovic, B., E. J. Sorin, and V. Pande. 2001. Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.* 313:151–169.
10. Zagrovic, B., C. D. Snow, …, V. S. Pande. 2002. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* 323:927–937.
11. Yang, J. S., W. W. Chen, …, E. I. Shakhnovich. 2007. All-atom ab initio folding of a diverse set of proteins. *Structure.* 15:53–63.
12. Verma, A., and W. Wenzel. 2009. A free-energy approach for all-atom protein simulation. *Biophys. J.* 96:3483–3494.
13. Schueler-Furman, O., C. Wang, …, D. Baker. 2005. Progress in modeling of protein structures and interactions. *Science.* 310:638–642.
14. Rangwala, H., and G. Karypis. 2008. fRMSDPred: predicting local RMSD between structural fragments using sequence information. *Proteins.* 72:1005–1018.
15. Zhang, Y. 2008. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* 18:342–348.
16. Bowman, G. R., and V. S. Pande. 2009. The roles of entropy and kinetics in structure prediction. *PLoS One.* 4:e5840.
17. Andrec, M., D. A. Snyder, …, R. M. Levy. 2007. A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins.* 69:449–465.
18. Saccenti, E., and A. Rosato. 2008. The war of tools: how can NMR spectroscopists detect errors in their structures? *J. Biomol. NMR.* 40:251–261.
19. Sullivan, D. C., and I. D. Kuntz. 2001. Conformation spaces of proteins. *Proteins.* 42:495–511.
20. Sullivan, D. C., and I. D. Kuntz. 2004. Distributions in protein conformation space: implications for structure prediction and entropy. *Biophys. J.* 87:113–120.
21. Müller, C. L., I. F. Sbalzarini, …, P. H. Hünenberger. 2009. In the eye of the beholder: inhomogeneous distribution of high-resolution shapes within the random-walk ensemble. *J. Chem. Phys.* 130:214904–214925.
22. Brüschweiler, R. 2003. Efficient RMSD measures for the comparison of two molecular ensembles. *Proteins.* 50:26–34.
23. Zagrovic, B., and W. F. van Gunsteren. 2007. Computational analysis of the mechanism and thermodynamics of inhibition of phosphodiesterase 5A by synthetic ligands. *J. Chem. Theory Comput.* 3:301–311.
24. Laurents, D., J. M. Pérez-Cañadillas, …, M. Bruix. 2001. Solution structure and dynamics of ribonuclease Sa. *Proteins.* 44:200–211.
25. Kövér, K. E., M. Bruix, …, M. Rico. 2008. The solution structure and dynamics of human pancreatic ribonuclease determined by NMR spectroscopy provide insight into its remarkable biological activities and inhibition. *J. Mol. Biol.* 379:953–965.
26. Zhou, Z., H. Q. Feng, …, Y. Bai. 2008. The high-resolution NMR structure of the early folding intermediate of the *Thermus thermophilus* ribonuclease H. *J. Mol. Biol.* 384:531–539.
27. Shortle, D., K. T. Simons, and D. Baker. 1998. Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. USA.* 95:11158–11162.
28. Betancourt, M. R., and J. Skolnick. 2001. Finding the needle in a haystack: educing native folds from ambiguous ab initio protein structure predictions. *J. Comput. Chem.* 22:339–353.

29. Zhang, Y., and J. Skolnick. 2004. SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem.* 25:865–871.

30. Król, M., I. Roterman, …, P. Spólnik. 2005. Analysis of correlated domain motions in IgG light chain reveals possible mechanisms of immunological signal transduction. *Proteins.* 59:545–554.

31. Yin, J., D. Bowen, and W. M. Southerland. 2006. Barnase thermal titration via molecular dynamics simulations: detection of early denaturation sites. *J. Mol. Graph. Model.* 24:233–243.

32. Sousa, S. F., P. A. Fernandes, and M. J. Ramos. 2009. Molecular dynamics simulations on the critical states of the farnesyltransferase enzyme. *Bioorg. Med. Chem.* 17:3369–3378.

33. Willis, B. T. M., and A. W. Pryor. 1975. Thermal Vibrations in Crystallography. Cambridge University Press, London; New York.

34. Phillips, Jr., G. N. 1990. Comparison of the dynamics of myoglobin in different crystal forms. *Biophys. J.* 57:381–383.

35. Halle, B. 2002. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA.* 99:1274–1279.

36. Meinhold, L., and J. C. Smith. 2005. Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease. *Biophys. J.* 88:2554–2563.

37. Lu, W. C., C. Z. Wang, …, K. M. Ho. 2006. Dynamics of the trimeric AcrB transporter protein inferred from a B-factor analysis of the crystal structure. *Proteins.* 62:152–158.

38. Glykos, N. M. 2007. On the application of molecular-dynamics simulations to validate thermal parameters and to optimize TLS-group selection for macromolecular refinement. *Acta Crystallogr. D Biol. Crystallogr.* 63:705–713.

39. Yang, L. W., E. Eyal, …, I. Bahar. 2007. Insights into equilibrium dynamics of proteins from comparison of NMR and x-ray data with computational predictions. *Structure.* 15:741–749.

40. Lu, C. H., S. W. Huang, …, J. K. Hwang. 2008. On the relationship between the protein structure and protein dynamics. *Proteins.* 72: 625–634.

41. Li, D. W., and R. Brüschweiler. 2009. All-atom contact model for understanding protein dynamics from crystallographic B-factors. *Biophys. J.* 96:3074–3081.

42. Hu, Z., and J. Jiang. 2010. Assessment of biomolecular force fields for molecular dynamics simulations in a protein crystal. *J. Comp. Chem.* 31:371–380.

43. Karplus, P. A., and G. E. Schulz. 1985. Prediction of chain flexibility in proteins—a tool for the selection of peptide antigens. *Naturwissenschaften.* 72:212–213.

44. Vihinen, M., E. Torkkila, and P. Riikonen. 1994. Accuracy of protein flexibility predictions. *Proteins.* 19:141–149.

45. Vihinen, M. 1987. Relationship of protein flexibility to thermostability. *Protein Eng.* 1:477–480.

46. Parthasarathy, S., and M. R. N. Murthy. 2000. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng.* 13:9–13.

47. Reetz, M. T., P. Soni, and L. Fernández. 2009. Knowledge-guided laboratory evolution of protein thermolability. *Biotechnol. Bioeng.* 102:1712–1717.

48. Stroud, R. M., and E. B. Fauman. 1995. Significance of structural changes in proteins: expected errors in refined protein structures. *Protein Sci.* 4:2392–2404.

49. Carugo, O., and P. Argos. 1998. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins.* 31:201–213.

50. Yuan, Z., J. Zhao, and Z. X. Wang. 2003. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng.* 16:109–114.

51. Mohan, S., N. Sinha, and S. J. Smith-Gill. 2003. Modeling the binding sites of anti-hen egg white lysozyme antibodies HyHEL-8 and HyHEL-26: an insight into the molecular basis of antibody cross-reactivity and specificity. *Biophys. J.* 85:3221–3236.

52. Carugo, O., and P. Argos. 1997. Correlation between side chain mobility and conformation in protein structures. *Protein Eng.* 10:777–787.

53. Eyal, E., R. Najmanovich, …, V. Sobolev. 2003. Protein side-chain rearrangement in regions of point mutations. *Proteins.* 50:272–282.

54. Carugo, O., and P. Argos. 1997. Protein-protein crystal-packing contacts. *Protein Sci.* 6:2261–2263.

55. Altman, R., C. Hughes, …, O. Jardetsky. 1994. Compositional characteristics of relatively disordered regions in proteins. *Protein Pept. Lett.* 1:120–127.

56. Romero, P., Z. Obradovic, …, A. K. Dunker. 1997. Identifying disordered regions in proteins from amino acid sequence. *The 1997 IEEE International Conference on Neural Networks Proc, Houston, TX.* 1:90–95.

57. Romero, P., Z. Obradovic, …, A. K. Dunker. 1998. Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* 3:437–448.

58. Radivojac, P., Z. Obradovic, …, A. K. Dunker. 2004. Protein flexibility and intrinsic disorder. *Protein Sci.* 13:71–80.

59. Navizet, I., R. Lavery, and R. L. Jernigan. 2004. Myosin flexibility: structural domains and collective vibrations. *Proteins.* 54:384–393.

60. Kuriyan, J., and W. I. Weis. 1991. Rigid protein motion as a model for crystallographic temperature factors. *Proc. Natl. Acad. Sci. USA.* 88:2773–2777.

61. Frauenfelder, H., G. A. Petsko, and D. Tsernoglou. 1979. Temperature-dependent x-ray diffraction as a probe of protein structural dynamics. *Nature.* 280:558–563.

62. Chong, S. H., Y. Joti, …, F. Parak. 2001. Dynamical transition of myoglobin in a crystal: comparative studies of x-ray crystallography and Mössbauer spectroscopy. *Eur. Biophys. J.* 30:319–329.

63. Berjanskii, M., and D. S. Wishart. 2006. NMR: prediction of protein flexibility. *Nat. Protoc.* 1:683–688.

64. Berjanskii, M. V., and D. S. Wishart. 2007. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. *Nucleic Acids Res.* 35(Web Server issue):W531–W537.

65. Berjanskii, M. V., and D. S. Wishart. 2008. Application of the random coil index to studying protein flexibility. *J. Biomol. NMR.* 40:31–48.

66. McKnight, C. J., P. T. Matsudaira, and P. S. Kim. 1997. NMR structure of the 35-residue villin headpiece subdomain. *Nat. Struct. Biol.* 4:180–184.

67. Eliezer, D., P. A. Jennings, …, H. Tsuruta. 1995. The radius of gyration of an apomyoglobin folding intermediate. *Science.* 270:487–488.

68. Bright, J. N., T. B. Woolf, and J. H. Hoh. 2001. Predicting properties of intrinsically unstructured proteins. *Prog. Biophys. Mol. Biol.* 76: 131–173.

69. Knott, M., and H. S. Chan. 2006. Criteria for downhill protein folding: calorimetry, chevron plot, kinetic relaxation, and single-molecule radius of gyration in chain models with subdued degrees of cooperativity. *Proteins.* 65:373–391.

70. Flory, P. J. 1989. Statistical Mechanics of Chain Molecules. Hanser Publishers, New York.

71. Qiu, D., P. S. Shenkin, …, W. C. Still. 1997. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem.* 101:3005–3014.

72. Andersen, H. C. 1983. Rattle—a velocity version of the shake algorithm for molecular-dynamics calculations. *J. Comput. Phys.* 52:24–34.

73. Jorgensen, W. L., and J. Tiradorives. 1988. The Opls potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* 110:1657–1666.

74. Zagrovic, B., C. D. Snow, …, V. S. Pande. 2002. Native-like mean structure in the unfolded ensemble of small proteins. *J. Mol. Biol.* 323:153–164.

75. Lindahl, E., B. Hess, and D. van der Spoel. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* 7:306–317.

76. McLachlan, A. D. 1982. Rapid comparison of protein structures. *Acta Crystallogr. A.* 38:871–873.

77. Furnham, N., T. L. Blundell, …, T. C. Terwilliger. 2006. Is one solution good enough? *Nat. Struct. Mol. Biol.* 13:184–185, discussion 185.

78. Wang, X., and J. Snoeyink. 2006. Multiple structure alignment by optimal RMSD implies that the average structure is a consensus. *LSS Computational Systems Bioinformatics Conference, Stanford, CA*. 79–87.

79. Bürgi, R., J. Pitera, and W. F. van Gunsteren. 2001. Assessing the effect of conformational averaging on the measured values of observables. *J. Biomol. NMR*. 19:305–320.

80. Zagrovic, B., and W. F. van Gunsteren. 2006. Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us? *Proteins*. 63:210–218.

81. Joosten, R. P., T. Womack, …, G. Bricogne. 2009. Re-refinement from deposited x-ray data can deliver improved models for most PDB entries. *Acta Crystallogr. D Biol. Crystallogr*. 65:176–185.

82. Nabuurs, S. B., A. J. Nederveen, …, C. A. Spronk. 2004. DRESS: a database of REfined solution NMR structures. *Proteins*. 55:483–486.

83. Joosten, R. P., and G. Vriend. 2007. PDB improvement starts with data deposition. *Science*. 317:195–196.

84. van Gunsteren, W. F., J. Dolenc, and A. E. Mark. 2008. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struct. Biol.* 18:149–153.

85. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph. Model.* 14:33–38.