

# Similarity Measures for Protein Ensembles

Kresten Lindorff-Larsen<sup>1‡\*</sup>, Jesper Ferkinghoff-Borg<sup>2</sup>

**1** Department of Molecular Biology, University of Copenhagen, Copenhagen Ø, Denmark, **2** Ørsted DTU, Technical University of Denmark, Ørsted Plads, Lyngby, Denmark

## Abstract

Analyses of similarities and changes in protein conformation can provide important information regarding protein function and evolution. Many scores, including the commonly used root mean square deviation, have therefore been developed to quantify the similarities of different protein conformations. However, instead of examining individual conformations it is in many cases more relevant to analyse ensembles of conformations that have been obtained either through experiments or from methods such as molecular dynamics simulations. We here present three approaches that can be used to compare conformational ensembles in the same way as the root mean square deviation is used to compare individual pairs of structures. The methods are based on the estimation of the probability distributions underlying the ensembles and subsequent comparison of these distributions. We first validate the methods using a synthetic example from molecular dynamics simulations. We then apply the algorithms to revisit the problem of ensemble averaging during structure determination of proteins, and find that an ensemble refinement method is able to recover the correct distribution of conformations better than standard single-molecule refinement.

**Citation:** Lindorff-Larsen K, Ferkinghoff-Borg J (2009) Similarity Measures for Protein Ensembles. PLoS ONE 4(1): e4203. doi:10.1371/journal.pone.0004203

**Editor:** Mark Isalan, Center for Genomic Regulation, Spain

**Received:** October 28, 2008; **Accepted:** November 25, 2008; **Published:** January 15, 2009

**Copyright:** © 2009 Lindorff-Larsen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors have no support or funding to report.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: Lindorff@DEShawResearch.com

‡ Current address: D.E. Shaw Research, New York, New York, United States of America

## Introduction

Protein structures play an important role in molecular biology, and are used for example in protein engineering studies, in drug design and as basis for understanding biological mechanisms at the molecular level. However, not only the average structure but also the dynamics around this structure plays a role in function [1]. These dynamical features of proteins may be highlighted by presenting ensembles of conformations instead of individual representative structures. Such ensembles can either be determined experimentally [2–8] or using methods such as molecular dynamics (MD) simulations [1]. In addition to the intrinsic dynamics of protein molecules, additional variability in experimental structures may also arise from uncertainties and statistical noise during experiments and structure determination. In order to reflect this variability, experimental structures should in general be presented as ensembles of conformations [4,9,10].

The ability to compare different protein conformations is an important tool in structural biology [11]. For example, the structural changes that occur between the apo- and ligand-bound forms of a protein may provide important clues as to the mechanism of binding. Also, structural similarity between proteins is often used as a starting point for determining and understanding function. Such comparisons typically rely on the calculation of the root mean square deviation (RMSD), or other structural similarity measures, between the atomic coordinates of the different conformations [12].

In a pioneering study, Brüschweiler extended the RMSD measure to ensembles of conformations [13]. In particular, he defined an inter-ensemble RMSD (eRMSD), whose square value is the average mean square deviation between conformations in two ensembles. A very similar measure has recently been used to

compare multiple sets of MD trajectories [14] and different methods for protein structure determination [7]. One drawback of the eRMSD is that in general it is non-zero even when the two ensembles are identical, making it difficult to use the eRMSD quantitatively. Also, as the calculations involve the (isotropically distributed) covariance matrices, the eRMSD potentially neglects important contributions from higher order moments of the distributions of conformations. A related measure that has been used to examine convergence of molecular simulations is also based on the covariance matrix [15,16], and may hence suffer from the same problems.

We here describe three alternative methods to compare ensembles of conformations. The methods are all based on the idea that two ensembles can be compared by estimating the probability densities underlying the ensembles. Using distance measures for probability densities that are based in information theory we can then provide a direct measure of the similarity between protein ensembles.

We first explore the properties of the methods and validate them by examining approximately normally distributed conformations obtained from MD simulations. We then use the methods to revisit the problem of ensemble averaging during protein structure determination from NMR data [17,18]. In particular, we explore to what extent structure determination methods are able to recover the correct distribution of conformations using experimental data as input. This is an important problem that has so far received very little attention due to the lack of methods for estimating and comparing distributions of conformations. Using synthetic data we find that an ensemble that has been determined by ensemble-refinement is more similar to a reference ensemble than an ensemble determined using standard single-molecule refinement.

## Methods

### Overall strategy for ensemble comparison

Our strategy towards a quantitative comparison of conformational ensembles is first to estimate a probability density for each ensemble and subsequently to compare these densities. We thus view a particular set of conformations as a sample from an underlying distribution and aim to model this distribution based on the sample at hand. By comparing the probability densities we are not only able to test whether the conformations in the two ensembles are similar, but also whether they occur with the same frequency. The latter point is an important criterion in many applications such as for example in the analysis of convergence of MD simulations [19] as well as of methods for protein structure determination [7,9].

The problem of comparing two ensembles is thus broken down into two steps [20]. We first estimate the densities from the ensembles and secondly we compare the individual densities. The three methods described further below present three complementary methods for estimating the densities, and we here first focus on how to compare the individual probability densities.

Given two ensembles,  $A$  and  $B$ , and an estimate of the corresponding density functions,  $p_A$  and  $p_B$ , the similarity between  $A$  and  $B$  is given as the distance between  $p_A$  and  $p_B$ . We here use the word distance in a more general meaning including for example measures that may not be a metric in the mathematical sense. Several methods exist for comparing probability densities [21]. Because we want to be able to apply the methods generally we chose measures from information theory that do not rely on a particular form of the probability density. A standard measure of the similarity of two probability densities is the Kullback-Leibler divergence [22,23]:

$$D_{KL}(p_A, p_B) = \int p_A(x) \log \frac{p_A(x)}{p_B(x)} dx \quad (1)$$

While the Kullback-Leibler divergence is not a true metric it is commonly used as a measure of the similarity between two distributions. In particular  $D_{KL}(p_A, p_B)$  is zero only when  $p_A = p_B$ , and is positive otherwise.

In an information theory context,  $D_{KL}(p_A, p_B)$  is also known as the relative entropy, and is related to the information lost about  $p_A$  if only  $p_B$  is known [23]. In thermodynamics, the Kullback-Leibler divergence can be thought of as the (non-equilibrium) free-energy difference between the two ensembles [24,25]. Thus,  $D_{KL}$  is a natural choice for a method for comparing two protein ensembles. If the score is zero the two ensembles represent the same amount of information and have been derived from the same free energy surface.

A noteworthy property of  $D_{KL}$  is that it is not symmetric, i.e. in general  $D_{KL}(p_A, p_B) \neq D_{KL}(p_B, p_A)$ . One approach to obtain a symmetric measure is to use the so-called  $\mathcal{J}$ -divergence defined as the average of  $D_{KL}(p_A, p_B)$  and  $D_{KL}(p_B, p_A)$  [22,26].

Another symmetrized, and ‘smoothed’, version of  $D_{KL}$  is the related Jensen-Shannon divergence [27,28]:

$$D_{JS}(p_A, p_B) = 0.5 \cdot (D_{KL}(p_A, (p_A + p_B)/2) + D_{KL}(p_B, (p_A + p_B)/2)) \quad (2)$$

In addition to being symmetric, it can be shown that the Jensen-Shannon divergence is the square of a metric [29]. We in general prefer the Jensen-Shannon divergence because it is well-defined even in the situation where one density is zero in regions where the other is not.

For probability distributions of discrete variables the integrals in the calculations of  $D_{KL}$  and  $D_{JS}$  are substituted by summations.

### Calculation of ensemble similarities

We devised three complementary methods for calculating ensemble similarities. These three methods (harmonic ensemble similarity (HES), a clustering based similarity and a dimensionality based similarity) are all described in more detail in the results section and we here provide only some of the more technical details. We applied all three methods to a series of protein ensembles that were generated as described in the subsection *Molecular dynamics and structure determination* below.

First, the harmonic ensemble similarity was calculated using Eq. 7. In those calculations the mean positions of the  $C_\alpha$  atoms were estimated as the averages over the ensemble. The covariance matrices were estimated using a recently described shrinkage approach [30,31]. Estimates of the errors of  $D_{HES}$  of the full ensembles A–C were obtained as the standard deviation over 100 bootstrap samples [32] from the ensembles. Also, we validated that the values calculated using the shrinkage estimator gave very similar results to the standard calculations of covariances when the ensembles were large.

Secondly, the clustering based ensemble similarities were calculated using the recently described Affinity Propagation (AP) clustering algorithm [33] and a discretized version of Eq. 2 to estimate the Jensen-Shannon divergence between the ensembles. In the case of the three reference ensembles A–C we used the  $C_\alpha$  RMSD as input to the algorithm. For the calculations used to analyse ensemble averaging in structure determination we calculated the RMSD over all non-hydrogen atoms. We used the negative RMSD between two conformations as a measure of their similarity. The number of clusters in the AP algorithm is determined by the preference for each conformation to act as a cluster centre, and we varied this number to obtain different number of clusters. The results for the reference ensembles A–C were obtained using preferences between  $-100$  and  $-0.1$ , and the results in the study of ensemble refinement were obtained using values between  $-100$  and  $-0.5$ . We also tested the clustering method described for analysing convergence [19], and obtained similar results although the method was not as efficient as AP for separating the B and C ensembles.

The final method for calculating ensemble similarities is based on a dimensionality reduction approach and consists of two steps. First, we use the Stochastic Proximity Embedding (SPE) algorithm [34,35] to represent the high-dimensional conformation space in a low-dimensional subspace. The algorithm attempts to find low dimensional projections that minimize the stress function in Eq. 8 in the results section. By minimizing  $S$  we obtain low dimensional vectors whose pairwise distances are close to the conformational similarities. The basic step in the SPE algorithm involves an update of the positions of the vectors in the subspace. The size of this update is determined by a learning rate which we decreased from 1.0 to 0.001 over 500 steps. At each value of the learning rate we performed  $6 \cdot 10^6$  update steps for the calculations using ensembles A–C (total of 7500 conformations) and  $2 \cdot 10^6$  update steps for the calculations using the reference and NOE-derived ensembles (total of 2307 conformations). We verified that these values of the parameters in SPE were sufficient to find low values of  $S$ , and that the results could be reproduced using multiple independent runs.

In the second step of the calculations we use kernel density methods to estimate the densities in the projected subspaces [36–38]. The results described here were obtained using Gaussian kernels, and in the case of density estimates in dimensions larger

than one we used product kernels. We used a plug-in estimator for the kernel bandwidths [39], but have verified that other kernel types and bandwidth estimates give very similar results.

For comparison with the ensemble RMSD method we calculated the squared value of the eRMSD using its definition [13]:

$$eRMSD(A,B)^2 = (MN)^{-1} \sum_{l,k=1}^{M,N} RMSD(a^l, b^k)^2 \quad (3)$$

where  $a^l$  is the  $l$ th structure in the  $A$  ensemble,  $b^k$  is the  $k$ th structure in the  $B$  ensemble, and  $M$  ( $N$ ) is the number of conformations in the  $A$  ( $B$ ) ensemble.

### Molecular dynamics and structure determination

In this section we describe how the different structural ensembles used in the study were generated. For the first validation part of our study we generated three ensembles (A–C) of the GB1 domain of protein G using MD simulations. The starting structure for the simulations was the first model in the PDB entry 3GB1. We used the polar hydrogen model [40] and a potential of mean force description of the solvent [41]. In addition we applied mass weighted harmonic restraints with energy:

$$E = k \sum m_i (x_i - x_i^{ref})^2 \quad (4)$$

In this equation  $m_i$  is the mass of the  $i$ th atom,  $x_i$  is its position and  $x_i^{ref}$  is the position in the reference conformation. The strength of the restraints is determined by  $k$ , the force constant, chosen to be 0.1, 0.01 and 0.001 in ensemble A, B and C, respectively. The structure was first heated to 300 K over a period of 0.2 ns and then equilibrated for further 2 ns at this temperature before final sampling for 10 ns. Structures were saved every 4 ps giving rise to 2500 structures in each ensemble.

In the second part of our study we analysed to what extent structure determination methods can recover the correct distribution of conformations. We thus generated a reference ensemble of the GB1 domain using an MD simulation with the CHARMM22 topology format and force-field [42] with a modified backbone potential [43] and a generalized Born solvent model [44]. The lengths of bonds involving hydrogen atoms were fixed [45], but no additional restraints were used. The structure was heated to 300 K during 0.2 ns and then equilibrated for 2 ns before sampling for 59 ns. Structures were extracted every 40 ps giving rise to an ensemble consisting of 1475 conformations which we subsequently used to generate synthetic NOE restraints. The native conformation remained stable throughout the simulation. For example, the all-atom (excluding hydrogens) RMSD to the starting conformation is  $1.4 \pm 0.1$  Å (average and standard deviation over the full 59 ns), with no trend of an increasing RMSD during the simulation. In the projections of this ensemble that we describe below, the distribution of conformations appears bimodal. This is caused by a slight structural rearrangement that occurs after 37 ns of the simulation. The rearrangement is localized to residues 10, 11, 40 and 41 which are located close to each other in the structure of Protein G. We note that the corresponding residues in the related GB3 domain have been shown by NMR spectroscopy to display long-timescale motion. The average overall RMSD to the native state is the same before and after the rearrangement.

For the generation of a synthetic NOE dataset we calculated the distances between all pairs of protons in each conformation, and then determined effective distances [46] as  $\langle r^{-3} \rangle^{-1/3}$ . All pairs

where this ‘average’ distance was larger than 5 Å were discarded, leaving a total of 2221 distance restraints (symmetry related methyl and ring-protons were treated using pseudo-atoms [47]). Finally, we classified the pseudo-NOE distance restraints as strong (1.8–2.4 Å), medium (2.4–3.8 Å) and weak (3.8–5.0 Å), and these values were then used as lower and upper bounds in structure determination. While this method for estimating NOEs from MD simulations neglects certain dynamical effects [48] it has been shown to work well in practice [49], in particular as the goal is here to generate a synthetic dataset [50] and not to predict experimental NOEs.

The synthetic data were subsequently used as input to a structure determination protocol. To diminish the bias introduced by using the same force field to generate the data and in the structure determination protocol we used a different solvation model [5,51,52] in these calculations. Structure calculation was performed using biased MD [53] with NOE distance restraints using a simulated annealing protocol described previously [5]. In these ensemble simulations  $N_{rep}$  conformations are simulated in parallel. In practice, if  $N_{cyc}$  cycles of simulated annealing are carried out one obtains an ensemble consisting of  $N_{rep} \cdot N_{cyc}$  conformations. In this formulation  $N_{rep}=1$  corresponds to standard structure determination methods. We carried out  $N_{cyc}=64$ , 64, 32, 32 and 16 simulated annealing cycles in simulations with  $N_{rep}=1, 2, 4, 8$  and 16 molecules, respectively, and thereby obtained five different ensembles of GB1.

All simulations were performed using the CHARMM molecular dynamics program [54].

## Results

As described in the Methods section, the overall strategy for comparing structural ensembles is to model each ensemble as a probability distribution, and then to compare these distributions. The previous sections outlined the methods used to compare the probability distributions, and hence the remaining problem is to estimate the densities from the ensembles. The following three sections describe three different, yet complementary, approaches for estimating probability densities from protein ensembles. In each section we describe the underlying idea and apply the method to compare three test-ensembles that we generated to test the methods. In the final part of the Results section we apply all three methods to a common problem relating to ensemble refinement using NMR data.

### Harmonic ensemble similarity

We model the protein ensemble as generated from a probability density function. The form of such density functions are, however, in general unknown and must either be guessed or inferred from the sample. A commonly used approximation is to assume that the ensemble is derived from a multivariate normal distribution [55]. When the parameters in this distribution are estimated from the ensemble, this is known as the quasi-harmonic approximation because anharmonic contributions may be incorporated indirectly via the estimated parameters [55]. In this framework, the probability of observing a conformation,  $x$ , is given by:

$$p(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \langle x \rangle)^T \Sigma^{-1} (x - \langle x \rangle)\right) \quad (5)$$

Here  $n$  is the number of degrees of freedom,  $\langle x \rangle$  are the mean coordinates,  $\Sigma$  is the covariance matrix that contains information about the fluctuations of the ensemble,  $|\Sigma|$  is its determinant and

$\Sigma^{-1}$  is the inverse of  $\Sigma$ . That is, if we assume that an ensemble  $A$  is drawn from an underlying Gaussian distribution  $p_A(x)$ , we can estimate  $p_A$  by estimating  $\langle x \rangle_A$  and  $\Sigma_A$  from the ensemble.

When  $p_A$  and  $p_B$  are multivariate normal distributions an analytical solution can be found to the Kullback-Leibler divergence between the two [21,56].

$$D_{KL}(p_A, p_B) = \frac{1}{2} \left( (\langle x \rangle_A - \langle x \rangle_B)^T \Sigma_B^{-1} (\langle x \rangle_A - \langle x \rangle_B) + \log \left( \frac{|\Sigma_B^{-1}|}{|\Sigma_A^{-1}|} \right) + \text{tr}(\Sigma_A \Sigma_B^{-1} - I_n) \right) \quad (6)$$

Here  $\langle x \rangle_A$  and  $\langle x \rangle_B$  are the means of the two distributions,  $\Sigma_A$  and  $\Sigma_B$  are the covariance matrices,  $I_n$  is the identity matrix of size  $n$ , and  $\text{tr}()$  denotes the trace of a matrix.

We have been unable to derive a closed-form expression for  $D_{JS}$  between two normal distributions. Instead we use the symmetrized  $\mathcal{J}$ -divergence defined as the average of  $D_{KL}(p_A, p_B)$  and  $D_{KL}(p_B, p_A)$ . In the case of two multivariate normal distributions we term this average the harmonic ensemble similarity ( $D_{HES}$ ):

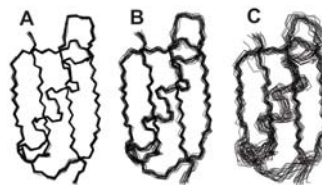
$$D_{HES}(p_A, p_B) = \frac{1}{4} \left( (\langle x \rangle_A - \langle x \rangle_B)^T (\Sigma_A^{-1} + \Sigma_B^{-1}) (\langle x \rangle_A - \langle x \rangle_B) + \text{tr}(\Sigma_A^{-1} \Sigma_B + \Sigma_A \Sigma_B^{-1} - 2I_n) \right) \quad (7)$$

The first term in  $D_{HES}$  is related to the Mahalanobis distance between the two distributions [57], and is zero only when the two means are identical. This term can be interpreted as a generalized RMSD between the two mean conformations, giving different weights to each atom depending on how much it fluctuates. The second term in  $D_{HES}$  depends only on the covariances of the two ensembles, and is zero when  $\Sigma_A = \Sigma_B$ .  $D_{HES}$  is therefore a measure of the similarity between two ensembles that gives weight to both differences in the mean conformation as well as differences in the fluctuations away from this mean. Thus,  $D_{HES}$  will be zero only when both the mean structures and the fluctuations are identical in the two ensembles. The harmonic ensemble similarity score only explicitly includes the first two moments of the distribution of conformations. In this sense, it is similar in spirit to the eRMSD [13] and the covariance-overlap [16], and may not be applicable for ensembles that have more complicated distributions of conformations. A strength of  $D_{HES}$  is, however, that it is a direct measure of the similarity of the distribution functions, and therefore has a clear statistical meaning. Further,  $D_{HES}$  is zero when the two ensembles are identical and non-zero otherwise.

In order to demonstrate the ability of  $D_{HES}$  to quantify the similarity between ensembles, we used MD simulations to generate three sets of conformations, termed A, B and C, of the 56 residue GB1 domain of Protein G. In each of the three simulations we applied mass-weighted harmonic restraints to all atoms, and the three simulations differed by the force-constant used for these restraints as follows A ( $k=0.1$ ), B ( $k=0.01$ ) and C ( $k=0.001$ ). As the force constant is decreased, the amplitude of atomic fluctuations become larger and the ensembles thus become increasingly more 'broad'. We extracted 2500 structures from each MD simulation and all of these structures were used in the calculations described below unless otherwise stated. By construction, the three ensembles were generated so that A and C would both be more similar to ensemble B than to each other. In practice we find that the three ensembles have similar average structures, with the  $C_\alpha$  RMSDs of the mean structures being 0.2 Å (A vs. B), 0.7 Å (A vs. C) and 0.6 Å (B vs. C). On the other hand the ensembles differ in how 'broad' they are. For example the mean pairwise RMSD within ensemble A is only 0.3 Å, whereas it is

0.5 Å and 0.9 Å for B and C, respectively. Representative structures from the three ensembles are shown in Fig. 1A–C. We note that the harmonic restraints do not ensure that the ensembles conform precisely to a multivariate normal distribution because of the additional presence of the MD force field.

In the comparison of the three ensembles we focus on the 56  $C_\alpha$ -atoms giving a total of  $n=3 \cdot 56=168$  degrees of freedom ( $x, y, z$  coordinates for each  $C_\alpha$ -atom). As input to the calculation of  $D_{HES}$  we need to estimate the mean and the covariance matrix of these degrees of freedom from the three ensembles. The mean conformation was estimated as the average over the ensemble. The standard maximum-likelihood method for estimating the covariance matrix,  $\Sigma$ , involves the calculation of each of the matrix elements  $\sigma_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$  individually. However, while this estimate of  $\Sigma$  is (almost) unbiased for an infinitely large dataset, it is known that a more robust overall estimate for  $\Sigma$  can be obtained using a so called shrinkage estimate when the number of data points is small compared to the number of degrees of freedom [58]. The idea behind the shrinkage approach is that an improved (in terms of total mean square error) estimate can be obtained by simultaneously estimating all elements of the covariance matrix. Importantly, the standard (maximum likelihood) estimate of  $\Sigma$  becomes singular when the number of structures is less than then number of degrees of freedom making it complicated to estimate  $\Sigma^{-1}$  (and  $|\Sigma|$ ). As we wish to develop a method that can also be applied to small ensembles such as those determined directly from experiments we therefore used a recently developed shrinkage estimates for the covariance matrices [30,31], and from these we estimated the similarity between all pairs of ensembles using Eq. 7. These calculations show that ensemble A is more similar to ensemble B ( $D_{HES}(A, B) = 93 \pm 1$ ) than to ensemble C ( $D_{HES}(A, C) = 650 \pm 10$ ), and also that ensemble C is more similar to B ( $D_{HES}(B, C) = 118 \pm 2$ ) than to A. We also calculated the two terms in Eq. 7 separately and found that they are of comparable magnitude with the first term making up between 40%–58% of  $D_{HES}$ . These results are in full agreement with the visual inspection of the ensembles in Fig. 1, and are also in line with the way the ensembles were generated. For comparison, we also calculated the interensemble eRMSD score as described previously [13]:  $eRMSD(A, B) = 0.45$  Å,  $eRMSD(A, C) = 0.94$  Å, and  $eRMSD(B, C) = 0.91$  Å. The eRMSD thus gives a similar ordering as the harmonic ensemble similarity score, although ensemble C is found to be only slightly more similar to B than to A. However, while the self-similarity using  $D_{HES}$  is always zero, this is not the case for the eRMSD:  $eRMSD(A, A) = 0.26$  Å,  $eRMSD(B, B) = 0.50$  Å, and  $eRMSD(C, C) = 0.92$  Å. Taken literally, this means that ensemble C is more similar to B than to itself, which makes it more complicated to use the eRMSD to rank the different similarities.



**Figure 1. Representative structures from three ensembles generated using molecular dynamics simulations.** These ensembles of the GB1 domain of protein G were obtained using MD simulations in the presence of mass-weighted harmonic restraints of increasing size. The three ensembles **A**, **B** and **C** were obtained using force constants 0.1, 0.01 and 0.001, respectively. doi:10.1371/journal.pone.0004203.g001

In order to demonstrate that the calculations of  $D_{HES}$  can be carried out using smaller ensembles we repeated the calculations using as little as 25 conformations extracted from each of the full ensembles of 2500 conformations. We denote these subensembles as  $A'$ ,  $B'$  and  $C'$  and present the results as the average and standard deviation over 100 randomly selected subensembles. The resulting similarity scores are  $D_{HES}(A',B')=175\pm 13$ ,  $D_{HES}(A',C')=1065\pm 64$  and  $D_{HES}(B',C')=200\pm 16$ . While there are numerical changes from the results of the full ensembles, the overall trends are identical, including the observation that the distance from B to its two 'neighbouring' ensembles (A and C) is much smaller than the distance between A and C. These observations are noteworthy as the calculations of  $D_{HES}$  here involve the estimation of the  $168\times 168$  covariance matrices from only 25 conformations. In this case the maximum likelihood estimate of  $\Sigma$  would have been singular making it impossible to calculate its inverse and hence  $D_{HES}$ . In contrast, the shrinkage estimates of the covariance matrices are guaranteed to be positive definite and hence invertible.

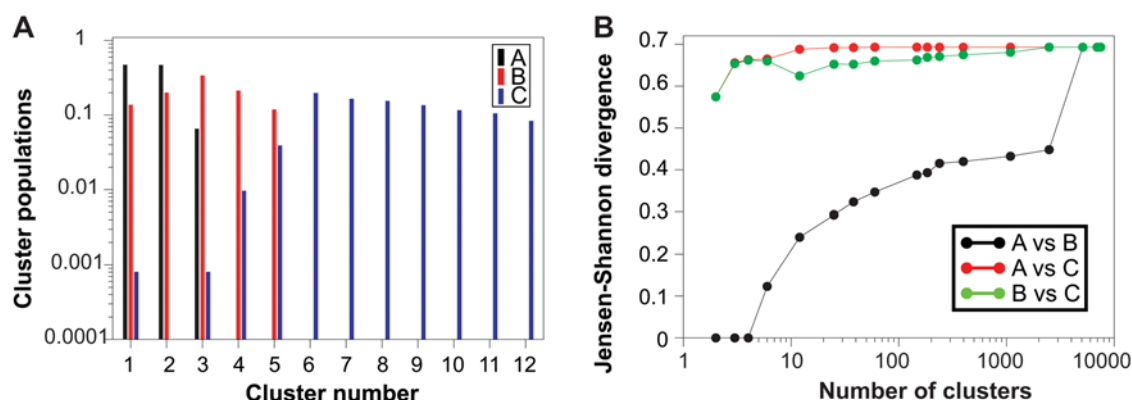
### Ensemble similarity from conformational clustering

Although the calculations of  $D_{HES}$  are very fast and straightforward, there may be two potential problems in its practical application. First is the assumption that the ensembles can be described by a multivariate normal distribution. Although anharmonic contributions are partially included by explicitly estimating the covariances from the ensemble, the analytical solution to  $D_{HES}$  is based on the two ensembles being normally distributed. The second problem in the calculation of  $D_{HES}$  is that in many cases the ensemble size may be considerably smaller than the number of degrees of freedom ( $n$ ). While we have demonstrated above that the  $n\times n$  covariance matrix can be estimated from less than  $n$  conformations using the shrinkage-approach, we have until now only looked at the fluctuations of the  $C_\alpha$  atoms in the ensembles making  $n=3N$  where  $N$  is the number of residues. However, in many cases it is of interest to include fluctuations of side chain atoms in the comparison of the ensembles. Even ignoring hydrogen atoms in the GB1 domain,  $n>1000$  making the estimation of the covariances from small ensembles difficult.

Recently, a method has been proposed to analyse the convergence of MD simulations. The idea is to quantify the co-occurrence of conformations from the first and second half of a

simulation when the two halves are clustered together [19]. We here modify and extend this approach to estimate the similarities between two or more ensembles. To illustrate the method we first calculate the  $C_\alpha$  RMSDs between all pairs of the 7500 structures in ensembles A, B and C. We then use a recently described clustering algorithm termed Affinity Propagation (AP) [33] to divide the conformations into clusters. AP uses the similarities between pairs of structures to decide which conformations should be clustered together. The number of clusters,  $k$ , is chosen based on the 'preference' for each conformation to act as the centre of a cluster. We set the preferences to a common value for all conformations so that all conformations are equally likely to become cluster centres, and vary the preference-value to obtain different levels of clustering. After the structures are clustered we take the population,  $p_i^X$ , of each ensemble ( $X$ ) in each cluster ( $i$ ) as a probability distribution of conformations. This idea is illustrated in Fig. 2A which shows the populations of ensemble A, B and C in each of the 12 clusters obtained using a common cluster-preference value of  $-10$ . At this resolution all structures from ensemble A are found in clusters 1–3, ensemble B populates clusters 1–5, and the broader ensemble C populates 11 of the 12 clusters. Visual inspection of the co-occurrences suggest that the overlap between ensemble A and B is the largest, and further that the overlap between the C and B distributions is larger than between C and A. We then take the cluster populations for an ensemble  $X$ ,  $p^X = \{p_i^X\}$ , as a density estimate over the discrete set of clusters. From these we then quantify the ensemble similarities by calculating  $D_{JS}$  between each pair of ensembles, and find  $D_{JS}(A,B)=0.24$ ,  $D_{JS}(A,C)=0.69$  and  $D_{JS}(B,C)=0.62$  for  $k=12$ .

At high values of the common cluster-preference value each conformation ends up in its own cluster and hence  $k=7500$ . In this limit the three ensembles do not share any clusters making  $D_{JS}$  take on its maximal value. At low cluster-preferences all structures group together into very few clusters making  $D_{JS}$  become smaller. In this way, the cluster preference sets the resolution of the algorithm. We therefore repeated the calculations at a range of cluster preferences and calculated  $D_{JS}$  (Fig. 2B). This plot shows that at most resolutions, ensembles A and B are much more similar than the other two pairs when seen from the perspective of co-occurrence in the clusters. On the other hand ensemble C is clearly different from ensembles A and B. At intermediate resolutions it is clear that ensemble C is more similar to ensemble B than to A.



**Figure 2. Comparison of the three test ensembles using a method that quantifies the co-occurrence of structures during conformational clustering.** **A:** Populations of each of the ensembles A, B and C in each of the 12 clusters that we obtained using a cluster preference of  $-10$ . **B:** Jensen-Shannon divergence between the three ensembles at a series of cluster preferences giving rise to between 2 and 7500 clusters.

doi:10.1371/journal.pone.0004203.g002

## An ensemble similarity score from dimensionality reduction

One advantage of the clustering method described above compared to the harmonic ensemble similarity score is that it does not require that the ensembles are distributed according to a normal distribution. A potential limitation of the approach is the discreteness of the clusters which means that any conformational differences between structures in the same cluster are ignored. Similarly, any similarities between structures in separate clusters are also ignored. We therefore sought an alternative approach which would provide a continuous description of the distribution of conformations. The main obstacle here is the so called ‘curse of dimensionality’ which highlights the problem of estimating densities in high dimensions without having astronomically large samples [59]. For example, a 168-dimensional histogram with only two bins in each dimension would have more than  $10^{50}$  bins.

Dimensionality reduction methods such as principal components analysis (PCA) and multi-dimensional scaling (MDS) have a long tradition in structural biology, and can be used to represent protein conformations in a subspace of dimension  $d$  that is much smaller than the full  $n$ -dimensional configuration space [60,61]. MDS, for example, takes as input the pairwise similarities, typically an RMSD, between all pairs of conformations and provides a  $d$ -dimensional projection that aims to preserve these distances as well as possible [60,62]. Recently, a connection between clustering methods and dimensionality reduction approaches has been found. In particular it has been shown that PCA can be thought of as a continuous solution to the  $K$ -means clustering problem [63]. This observation suggests that dimensionality reduction may be a natural extension to the clustering method described above, and thereby provide a continuous approach to estimate densities of high-dimensional data from small samples.

An important underlying assumption of methods such as PCA and MDS is that the conformational space is linear. In practical terms this means, for example, that the algorithms give equal weight to an RMSD of 1 Å and 10 Å between two conformations when the projection is constructed, whereas it is clear that large RMSD values are not very useful measures of the similarity between conformations [64]. To overcome this problem one of us has previously applied the non-linear projection method Isomap [65] to represent protein topology space in three dimensions [66]. More recently, Isomap was extended and used to derive order parameters to study protein folding [67].

Other non-linear projection methods exist, and we have here chosen Stochastic Proximity Embedding (SPE) [34] because of its advantageous scaling properties with sample size compared to methods such as Isomap. Because of the stochastic nature of the SPE algorithm it should be run multiple times to test the reproducibility of the results. SPE takes as input the structural similarity between all pairs of conformations, and uses an iterative method to obtain a low-dimensional projection in which pairwise distances are approximately preserved locally. We here define the local neighbourhood as pairs of structures for which the  $C_\alpha$  RMSD is less than 1.5 Å, but we have verified that other definitions gave very similar results.

The SPE algorithm represents each conformation as a  $d$ -dimensional vector and finds a collection of vectors so that the Euclidian distance between two vectors in the  $d$ -dimensional projection is close to the RMSD between those two conformations if they are neighbours of each other (i.e. the RMSD is within a specified cut-off). Only for large values of  $d$  can all the restraints be fulfilled perfectly. For smaller  $d$  we calculate the remaining ‘stress’,  $S$ , as a measure of how well the restraints are fulfilled:

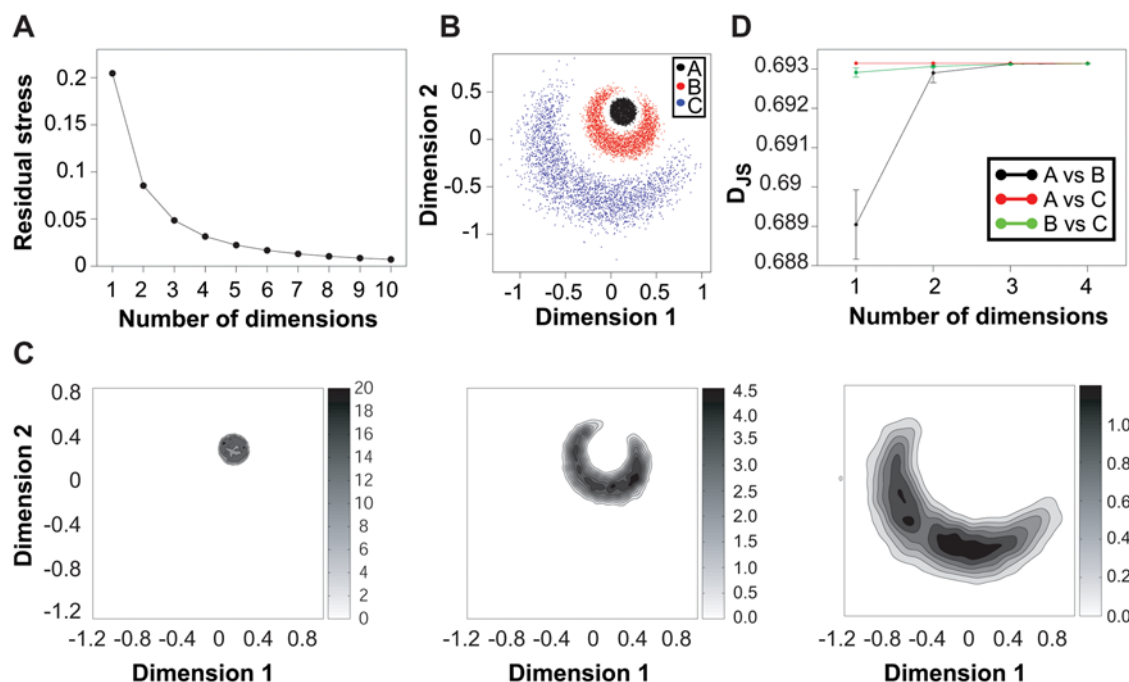
$$S = \frac{\sum (d_{ij} - r_{ij})^2}{\sum r_{ij}} \quad (8)$$

where the sums extend over the neighbours of each conformation [34]. In Eq. 8  $d_{ij}$  is the Euclidian distance between two points in the low dimensional projection and  $r_{ij}$  is the RMSD between the two conformations. In Fig. 3A we show the remaining stress as a function of the dimension of the subspace used in the projection. From this figure it is seen that the conformational ensembles can be well represented in a subspace of dimension considerable smaller than  $n = 168$ . In order to illustrate the low dimensional projections we show in Fig. 3B a scatter-plot of the two-dimensional projection of the 7500 conformations in ensembles A–C. In this plot, each conformation is represented as a single point, and these have been arranged by the SPE algorithm so that structurally similar conformations are located close to each other. As intuitively expected, the conformations in the more ‘narrow’ A-ensemble are located in a small region of the two-dimensional representation of the conformational space. In contrast the broader B- and C-ensembles are spread over larger area of conformational space. It is clear that the SPE algorithm can separate the A, B and C ensembles very efficiently using only the pairwise RMSDs as input. The ‘concentric’ nature of the distributions of conformations is evidence of the fact that the mean structures of the three ensembles are very similar. It is, however, clear that ensembles B and C do in fact not contain conformations that are close to this mean. This is a result of the high dimensional nature of conformational space, which means that the phase space associated with the centre is much smaller than in the outer regions, which therefore contain most of the conformations.

The reduction of the dimensionality of the data makes it feasible to estimate the probability density of each conformational ensemble. Instead of being distributions over the full  $n$ -dimensional conformation space, they will be distributions over the generalized coordinates in the  $d$ -dimensional subspace. As we do not want to restrict ourselves to particular distributions of conformations we use non-parametric density estimation to quantify the similarities between the ensembles [20]. In particular, we use kernel density estimation (KDE) [36,38] to obtain density estimates from the  $d$ -dimensional samples. In Fig. 3C we show the density estimates obtained for ensembles A–C from the two dimensional projections in Fig. 3B. As also observed in the scatter-plot, the density estimates show that the A-ensemble corresponds to a very sharply peaked probability distribution in line with the stronger harmonic restraints applied during the MD simulation. From the estimated densities we then calculate the Jensen-Shannon divergences between the ensembles. The plot in Fig. 3D shows  $D_{JS}$ , averaged over 5 independent SPE projections, for the comparisons between the A, B and C ensembles as a function of the dimensionality of the projection. As is evident from both the scatter plot and the density estimates there is very little overlap between the three densities, and  $D_{JS}$  is therefore close to its maximal value. However, the calculations show that ensembles A and B are more similar than the other two pairs of ensembles, and that ensemble C is more similar to B than to A. This is the same ordering of the three ensembles as found in the clustering analysis, although the actual values of  $D_{JS}$  differ in the two calculations.

## Ensemble averaging in structure determination

After validating the methods using the ensembles described above, we then applied the ensemble similarity calculations to an important problem in structural biology. We wanted to examine to what extent it is possible to recover the correct distribution of



**Figure 3. Comparison of the three test ensembles using a method which involves dimensionality reduction and kernel density estimation.** **A:** Average residual stress according to Eq. 8 over 5 independent SPE projections of the three test ensembles. The standard deviation is smaller than the symbols shown. **B:** Example of a two-dimensional projection of the three ensembles. Each point represents an individual conformation, and the distance between each point is locally approximately the same as the RMSD between those two conformations. The two axes represent the two dimensions in the subspace of the SPE projection. **C:** Contour plots of the two-dimensional kernel estimates of the densities corresponding to the points in panel B. The grey bars next to the plots indicate the scale of the probability densities. **D:** Average and standard deviation of the Jensen-Shannon divergence between the three ensembles calculated using the kernel density estimates. The results are shown for different values of the dimensionality of the projections. doi:10.1371/journal.pone.0004203.g003

conformations based only on a set of experimental data, and whether different structure determination approaches differ in their ability to recover the distributions. In particular, we revisit the problem of whether ensemble (multi-conformer) refinement using only NOE distance restraints can be used to recover the conformational distribution of a protein [18]. Again we used the GB1 domain as a test case, and performed a 59 ns reference MD simulation to generate a synthetic pseudo-experimental NOE dataset. The resulting distances were used as restraints in the structure determination of GB1, and the goal is then to examine which of the determined ensembles is most similar to the reference ensemble.

Most standard NOE-based structure determination protocols apply the available distance restraints on to a single conformation. That is, an optimization algorithm (typically simulated annealing) is used to find individual conformations that satisfy as many of the restraints as possible. Through multiple rounds of simulated annealing one then obtains an ensemble of conformations. It is important to note that in this approach it is attempted that all conformations should individually satisfy the restraints. However, these conformations may still differ because of the stochastic nature of the optimization algorithm and because the distance restraints typically do not uniquely identify a single conformation [4,9,68].

It is clear, however, that in reality — as well as with the synthetic data generated here — the distance restraints reflect a distribution of conformations, and that no single conformation needs to agree fully with the restraints. In order to reflect this ensemble view of the experimental data, it is possible to use the

data in so-called ensemble refinement protocols [2,3,17]. In these simulations, the experimental restraints are applied to multiple copies, or replicas, of the protein at any given time and the optimization protocol then aims to find ensembles that satisfy the restraints when the distances are back-calculated as an average over the ensemble. In general, if  $N_{rep}$  molecules are simulated in parallel none of the individual conformations need to satisfy the restraints individually as long as the  $N_{rep}$  molecules do so as an ensemble.

A practical problem with ensemble simulations is that as  $N_{rep}$  is increased, the number of degrees of freedom in the system also increases because one now attempts to determine  $N_{rep}$  conformations simultaneously. As the total number of distance restraints remains constant, this means that ensemble simulations are more likely to be underdetermined and hence prone to overfitting [7,17]. It has been suggested that  $N_{rep} \approx 2$  provides an optimal compromise between over- and under-restraining [17]. These conclusions were mainly based on cross-validation as an indirect method for examining whether two ensembles are similar or not, although local similarity scores have also been considered [7]. As  $N_{rep}=1$  remains the standard method used in protein structure determination we decided to explore this issue further.

We used the global ensemble similarity scores described above to quantify whether the ensembles obtained from ensemble simulations ( $N_{rep}$  between 2 and 16) are more similar to the ‘true’ reference than those obtained from standard ‘single-conformer refinement’ ( $N_{rep}=1$ ). We carried out multiple simulated annealing cycles in simulations with  $N_{rep}=1, 2, 4, 8$  and 16 molecules and thereby obtained five different ensembles of GB1. Ten

conformations from each ensemble are shown in Fig. 4. All ensembles are in agreement with the ‘experimental’ NOE restraints as a whole, but only in the  $N_{rep}=1$  ensemble are all the individual conformations in good agreement with the NOEs. It is noteworthy that the ensembles are visually very similar, and it is not easy to judge which of the five ‘NOE-derived’ ensembles is most similar to the reference MD ensemble.

We first calculated the harmonic ensemble similarity between the reference ensemble and the ensembles determined using  $N_{rep}$  between 1 and 16. Because of the relatively small ensemble sizes we used only the positions of the  $C_\alpha$  atoms in these calculations. Averaging over 25 randomly selected samples of 64 conformations from each of the NOE-derived ensembles we find that  $D_{HES} = 260 \pm 20$ ,  $157 \pm 3$ ,  $175 \pm 5$ ,  $212 \pm 13$  and  $233 \pm 15$  between the MD-reference ensemble and the ensembles determined using  $N_{rep} = 1, 2, 4, 8$  and  $16$ , respectively. These calculations suggest that the ensemble determined using  $N_{rep} = 2$  is most similar to the reference ensembles, at least when examining the fluctuations of the  $C_\alpha$  atoms.

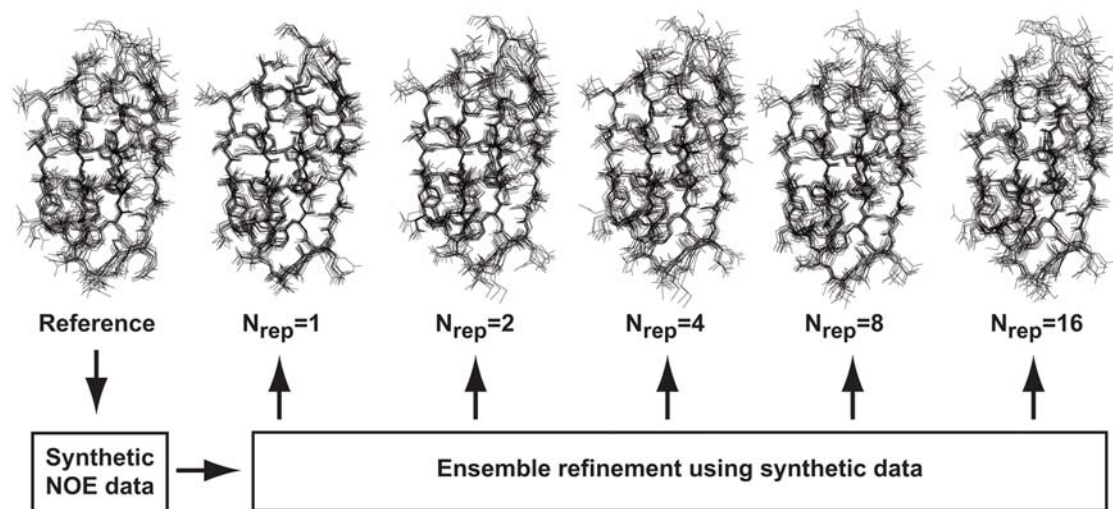
To obtain a more detailed view that includes both atoms in the polypeptide backbone chain and in the side chains we proceeded to calculate the pairwise all-atom (excluding hydrogens) RMSD between all 2307 conformations (reference ensemble and five ensembles determined using NOE restraints). These distances were then used as input to the calculations of ensemble similarities using both the clustering and the projection approach.

We first applied the clustering method to the six ensembles. Figure 5A shows the results of AP clustering at a relatively low value of cluster preferences. In this case the 2307 structures were divided in to eight clusters, and the plot clearly shows that the six ensembles populate the clusters to very different extents. For example, the  $N_{rep} = 1$  ensemble has all of its structures in cluster 1. In contrast only 2.4% of the reference conformations are found in this cluster.

We repeated the calculations using a range of preference values and calculated  $D_{JS}$  between the reference MD ensemble and each of the five ensembles determined using  $N_{rep}$  between 1 and 16. In Fig. 5B we show  $D_{JS}$  as a function of  $N_{rep}$  at preference values giving rise to between 5 and 149 clusters. While the numerical

values are different at the different levels of resolution, the general trends are very clear. In particular, the ensemble determined using  $N_{rep} = 2$  is more similar to the reference ensemble than any of the other ensembles.

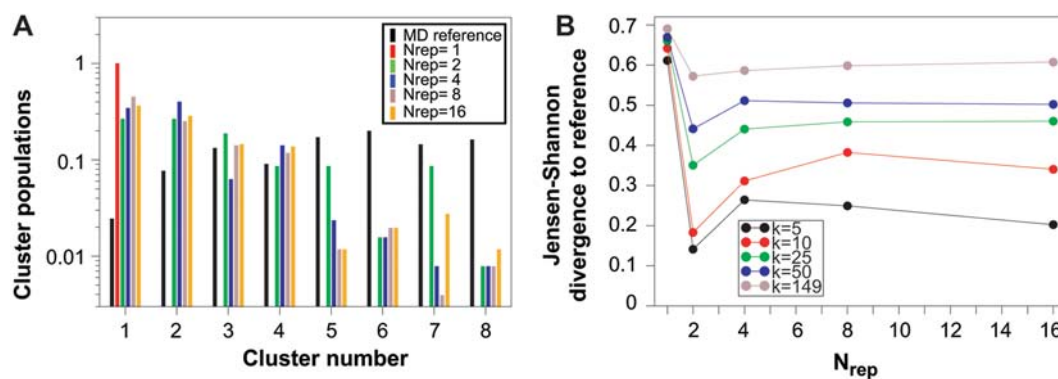
To analyse further the distributions of conformations in the six ensembles we used SPE to project the 2307 conformations in to low dimensional subspaces. In these calculations we defined neighbouring structures as those that had an all-atom (excluding hydrogens) RMSD less than  $1.75 \text{ \AA}$ , but validated that the conclusions below are the same using other definitions of neighbours (not shown). In Fig. 6A we show the remaining stress, averaged over 10 runs, as a function of the dimension of the projection. As with the harmonic ensembles above, the conformations can be well represented in a low dimensional projection. This is noteworthy because we here use the RMSD between all 436 non-hydrogen atoms corresponding to a much higher dimension of the conformation space. As an example of the distributions we show in Fig. 6B the kernel estimates of the six densities in two dimensions. The first of the plots show the density of the reference ensemble, and its bimodal character is evident. This is caused by a small and localized conformational readjustment for residues 10, 11, 40 and 41 that occurs after 36 ns of the reference simulation. The densities obtained from the conformations determined from the NOE restraints clearly show that the ensembles broaden out as  $N_{rep}$  is increased. In particular the ensemble determined using the standard  $N_{rep} = 1$  approach is significantly more narrow than the other ensembles. Visual inspection of these densities again suggest that  $N_{rep} = 2$  provides an optimal compromise between over- and under-restraining, and is able to recover the bi-modal nature of the conformational ensemble. To quantify these observations, we show in Fig. 6C the Jensen-Shannon divergence between the reference ensemble and the five ensembles for projections in dimensions 1–5. This plot clearly shows that at all dimensions the ensemble determined using  $N_{rep} = 2$  is more similar to the reference ensemble than any of the other four ensembles. Thus, all three ensemble similarity methods show that the ensemble determined using  $N_{rep} = 2$  provides the optimal compromise between over- and under-restraining. In contrast, the similarities calculated using the eRMSD method are



**Figure 4. Six ensembles of the GB1 domain of protein G.** The reference ensemble was obtained using molecular dynamics simulations, and was used to generate a set of synthetic pseudo-experimental distance restraints. These restraints were subsequently used in either single-conformer refinement ( $N_{rep} = 1$ ) or ensemble refinement using ensemble sizes  $N_{rep} = 2, 4, 8$  and  $16$ . All non-hydrogen atoms are shown in ten structures from each ensemble.

doi:10.1371/journal.pone.0004203.g004





**Figure 5. Examination of how well a reference ensemble can be recovered using ensemble simulations.** The results shown here were obtained using the clustering method described in the text. **A:** Populations of each of the ensembles (MD-reference and ensembles obtained using NOE restraints) in each of the 8 clusters found using the affinity propagation clustering algorithm with a cluster preference of  $-20$ . **B:** Jensen-Shannon divergence between the reference ensemble and the ensembles obtained using NOE restraints applied to different ensemble sizes ( $N_{rep}$ ). The results are shown for five representative values of the total number of clusters. doi:10.1371/journal.pone.0004203.g005

inconclusive. We find that there is a steady increase in the eRMSD as  $N_{rep}$  increases, with the similarity between the reference and the ensembles determined using  $N_{rep} = 1, 2, 4, 8, 16$  being 1.28, 1.36, 1.42, 1.51, and 1.53, respectively. However, as the eRMSD between the reference ensemble and itself is 1.37 it is not clear to us how these numbers should be interpreted.

## Discussion

We have presented three different algorithms to compare two or more ensembles of protein conformations. We first tested the methods using three ensembles, A–C, obtained by MD simulations with additional harmonic restraints. We then proceeded to use the methods to examine how well an ensemble refinement method is able to recover an ensemble of conformations using (synthetic) experimental data.

The first measure we present is the harmonic ensemble similarity. It is based on the assumption of a multivariate normal distribution whose mean and covariance are estimated from the ensembles at hand. Two ensembles are then compared using an analytical expression of the Kullback-Leibler divergence between two normal distributions. The approach is therefore related in spirit to the allosteric potential recently introduced [69], in which the fluctuations are estimated using normal mode analysis. Using this method to compare the three test ensembles we find that both ensembles A and C are significantly more similar to ensemble B than to each other, in full agreement with the way the ensembles were generated. An important advantage of the harmonic ensemble similarity is that it is easy and fast to calculate and has a well defined meaning. Also, because of its analytical nature it is easy to understand the contributions from differences in the mean structure and from changes in fluctuations.

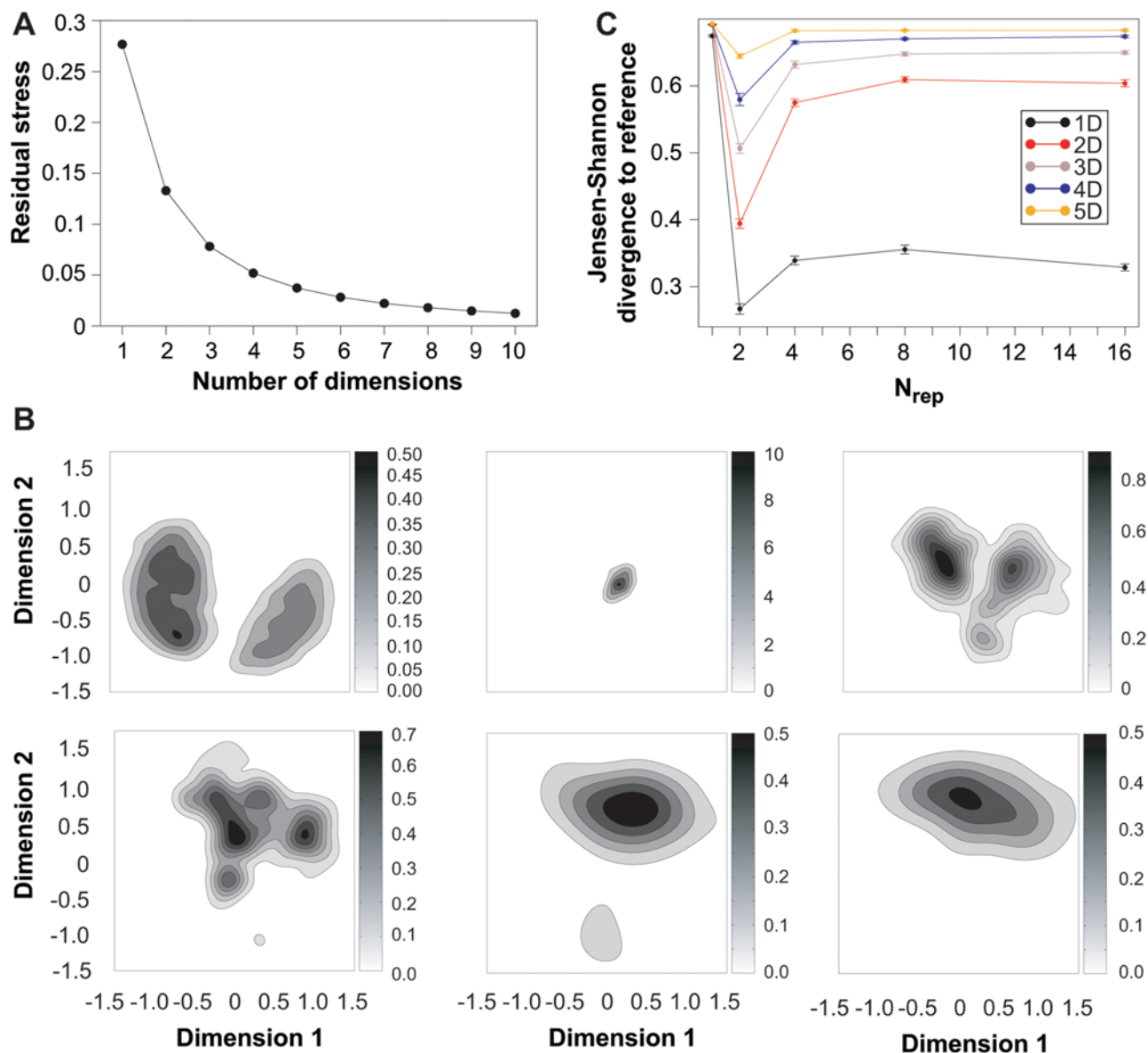
The harmonic ensemble similarity method is based on the assumption of a normally distributed ensemble, and it may not always be clear how to interpret  $D_{HES}$  when this assumption is not fulfilled. We note here that the harmonic assumption also forms the basis of methods such as the normal model analysis [55] which is known to describe reasonably well both the amplitude and directionality of native state fluctuations [70]. Also, if only the mean and covariances are known, the multivariate normal distribution is the least biased choice of a density function. Finally, in both of the cases studied here (harmonic ensembles and ensemble refinement), the results obtained using  $D_{HES}$  were in full

agreement with those obtained using the two other methods described.

A potential limitation in the calculation of  $D_{HES}$  is the uncertainties associated with estimating covariances from sparse data. The shrinkage approach provides a systematic method for estimating the covariance matrix and hence is useful in the common situation when the number of structures is of comparable magnitude to the number of degrees of freedom. In addition, the approach has the added advantage that it produces a regularized and positive definite estimate. Nevertheless, the shrinkage approach can not fully remove the uncertainties associated with estimating covariance matrices from sparse data. Finally we note that the shrinkage method may also be used in others areas of structural biology where estimation of covariance matrices are important. For example, it may be used in the estimation of conformational entropies [55,71,72].

In an attempt to evaluate some of the potential practical problems with the harmonic ensemble similarity we devised and tested a similarity score based on clustering the conformations in the ensembles. In particular we used the co-occurrence in conformational clustering as an estimate for the similarity of two ensembles. In the original implementation of this idea the differences in cluster populations were used as a measure of the similarity of two ensembles [19], but this value may be difficult to interpret [73]. We here substitute this measure by the Jensen-Shannon divergence as a direct measure of the similarity of the two distributions. Also, we use the AP algorithm as a very efficient and fast method for clustering the conformations.

The idea of analysing co-occurrence in clusters was originally designed to examine the convergence of molecular simulations, and hence mainly to answer the question of whether two sub-ensembles are likely to originate from the same distribution [19,73]. However, our goal is different in that we aim to derive a score that will quantify the divergence between two ensembles. As noted [19], the method is well suited to determine that two ensembles are different, but it is not easy to verify that two ensembles are similar. The results in Fig. 2B show this observation very clearly. At low numbers of clusters the algorithm cannot distinguish ensembles A and B, and hence the distance between these two ensembles is very low despite the fact that they differ considerably in their variability. Nevertheless, one can use the method to quantify the similarity of two ensembles *at the chosen clustering level*. The limitations of the clustering method arise



**Figure 6. Examination of how well a reference ensemble can be recovered using ensemble simulations.** The results shown here were obtained using the projection method described in the text. **A:** Average residual stress according to Eq. 8 over 10 independent SPE projections of the six ensembles (MD-reference ensemble and five ensembles obtained from NOE restraints). The standard deviation is smaller than the symbols shown. **B:** Example of the two-dimensional kernel estimates of the densities. The grey bars next to the plots indicate the scale of the probability densities. **C:** Average and standard deviation of the Jensen-Shannon divergence ( $D_{JS}$ ) between the reference ensemble and the ensembles obtained using NOE restraints applied to different ensemble sizes ( $N_{rep}$ ). The results are shown for different values of the dimensionality of the projections and are the averages over 10 independent runs of the SPE algorithm. doi:10.1371/journal.pone.0004203.g006

because of the discrete nature of the clusters over which the ensembles are compared, so that conformations in the same cluster are considered equivalent irrespectively of how similar they are.

In order to overcome these potential problems we devised an approach to estimate the density function over a set of continuous variables. In this way we hoped to solve the problems associated with the discreteness of the clusters. Because of the high dimensionality of conformational space it is very difficult to estimate the densities directly [59]. As our main goal is not to estimate the probability densities of the ensembles, but rather to use the density estimates to compare two ensembles, we therefore turned to a dimensionality reduction approach in order to find ‘order parameters’ that would automatically describe the conformations

in a subspace of reduced dimensionality. In particular, it has previously been found that the effective dimensionality of conformational space may be significantly smaller than the number of degrees of freedom [60,61,74], and we therefore expected that we could capture the conformational distributions well using dimensionality reduction. The results on native state ensembles of the GB1 domain described here show that the ensembles can indeed be well represented in low dimensional subspaces. As with the clustering method there is a compromise between having sufficient resolution, here meaning choosing the dimensionality, and being able to estimate densities. For the test ensembles A–C we find that the method can rank the similarities, and that the results are in accordance with how the ensembles were generated.

Finally we applied the similarity measures to reexamine the problem of ensemble averaging in structure determination. This is an important problem as it is well known that a wrong interpretation of the dynamical contribution to distance restraints may cause errors in structure determination [75]. While ensemble refinement in principle can be used to include dynamical effects, one limitation is that such calculations are more prone to being underrestrained than standard single-molecule refinement is.

An important and general method to compare different structure determination protocols is to generate synthetic data and use this as input to the different protocols. The idea is then to compare the ensembles obtained from structure determination with that used to generate the synthetic data. Until now, such comparison have mainly been carried using validation with independently determined data, cross-validation, measures of the local structural similarity [7,8,17], or using reaction coordinates chosen by hand [76,77]. However, these approaches are only indirect methods for quantifying to what extent one can recover the correct distribution of conformations, and do not take full advantage of the fact that all structural details of the reference ensemble are known.

We here use the ensemble similarity scores as a direct method for quantifying how well the reference ensemble can be recovered. The results show clearly that not taking averaging in to account, i.e. having  $N_{rep} = 1$ , gives rise to an ensemble that is more narrow than the reference ensemble. Increasing values of  $N_{rep}$  gives broader ensembles, and all three similarity scores show that with the dataset used here,  $N_{rep} = 2$  provides the optimal compromise between over- and under-restraining. We emphasize that this result is not necessarily general and applies only to the dataset for GB1 that we used. Also, the dataset that we have generated may be unrealistically large, and different results may therefore be obtained using fewer NOEs and in the presence of systematic errors arising for example from wrong assignments. Finally, the projection of the reference ensemble revealed that the distribution of conformations is bimodal because of a small structural change that occurred during the MD simulation. While this change is very small and localized, it may provide an additional reason for why  $N_{rep} = 1$  is insufficient to recover the correct distribution. Similarly large conformational fluctuations are, however, likely to occur in real proteins and we note that the residues that display slow motion in our simulation correspond to residues that have been found experimentally to have long-timescale motion in a related protein. Other authors [7,17] have also suggested  $N_{rep} = 2$  as being optimal for NOE data, at least when there is sufficient experimental data, although larger values may be needed for other data types [7,78,79]. Further, we stress that the methods we have presented are completely general and can therefore be used to examine these questions in more detail.

As standard structure determination protocols correspond to  $N_{rep} = 1$ , most structures in the Protein Data Bank have been determined in this way. It is therefore of relevance to analyse how well structures determined in this way represent the underlying ensemble. In particular it is relevant to analyse how well such structures represent a ‘typical’ structure from the correct ensemble. The studies described here have not been aimed at examining this question, and may be biased by the bi-modal nature of the reference ensemble used to generate the restraints. However, for this ensemble two lines of evidence point towards the possibility that structure determination using  $N_{rep} = 1$  may introduce a bias. First, examining Fig. 5 it is clear that the first cluster contains all of the conformations in the ensemble generated using  $N_{rep} = 1$ , whereas this cluster has the lowest population for the reference cluster. Secondly, the two-dimensional density estimate of the  $N_{rep} = 1$

ensemble in Fig. 6 is sharply peaked in a region where there is very little density in the reference ensemble. Together, these observations suggest that, in the case studied here, the  $N_{rep} = 1$  is not only too ‘precise’ (ensemble is too narrow), but may also be somewhat ‘inaccurate’ (centre of distribution does not coincide with that of the reference ensemble). These issues should be examined in more detail, but suggest for example that optimizing force fields against structures from the PDB may introduce biases in the estimated parameters. A related problem is that highlighted by the projections in Fig. 3. While the three ensembles here have similar average structures, only the narrow A-ensemble has structures that are very similar to the average. Thus, even if the average conformation can be determined accurately, it may not be sufficiently representative of the underlying high-dimensional probability distribution to be used in for example structure based drug-design.

The three ensemble comparison methods that we present each have different strengths. The harmonic ensemble similarity is simple to calculate and is based on an analytical relationship (Eq. 7). The equation clearly highlights the fact that for two ensembles to be similar both their ‘average’ structure as well as the fluctuations away from this should be similar. Importantly, the computational complexity in the calculation of the harmonic ensemble similarity is only linear in the number of conformations. In contrast, in the current implementations of the two other similarity scores the algorithms require all pairwise RMSDs, the calculations of which may become prohibitively expensive for very large ensembles. For these reasons we suggest that the harmonic ensemble similarity should be used as a starting point for comparing protein ensembles. In cases where significant deviations from the a normal distribution are expected (or found) we suggest to supplement the calculations of  $D_{HES}$  with the clustering and projection based similarity scores. These two methods were designed in different ways to deal with the high-dimensional nature of the data, and hence provide complementary views of the ensemble similarities.

We hope that the methods we present will be of use in many areas of structural biology. For example, the idea of using reference ensembles to validate structure determination protocols has recently been applied to non-native states of proteins [76,77] as well as revisiting ensemble refinement against Xray diffraction data [8]. Also, until now no community wide standards exist for assessing the convergence of simulations [80]. We suggest that measures such as those presented here could be used to examine the convergence of biomolecular simulations in cases where multiple states may be present. The algorithms can also be used to compare different simulation protocols [81] or molecular force fields [14]. Finally, we hope that the ideas presented here will be used to quantify biologically important changes in conformational distributions that may occur during for example ligand binding [82] and enzyme catalysis [83], and that may form the molecular basis for allosteric effects [84].

## Acknowledgments

We thank Alexander Ihler for sharing his code for kernel density estimation, Korbinian Strimmer for sharing his code for covariance estimation, Rajarshi Guha for SPE code, and Brendan Frey for sharing his code for Affinity Propagation. Guillermo Sapiro is thanked for sharing results prior to publication. Niels Richard Hansen and Thomas Hamelryck are thanked for discussions and comments, and the Danish Centre for Scientific Computing is thanked for computing resources.

## Author Contributions

Conceived and designed the experiments: KLL JFB. Performed the experiments: KLL. Analyzed the data: KLL JFB. Contributed reagents/materials/analysis tools: KLL JFB. Wrote the paper: KLL JFB.

## References

1. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nature Struct Biol* 9: 646–652.
2. Scheek RM, Torda AE, Kemmink J, van Gunsteren WF (1991) Structure determination by NMR: The modeling of NMR parameters as ensemble averages. Hoch JC, Redfield C, Poulsen FM, eds. New York, USA: Plenum Press. pp 209–217.
3. Kuriyan J, et al. (1991) Exploration of disorder in protein structures by X-ray restrained molecular dynamics. *Proteins* 10: 340–358.
4. DePristo MA, de Bakker PI, Blundell TL (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* 12: 831–838.
5. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433: 128–132.
6. Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M (2006) Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci USA* 103: 10901–10906.
7. Richter B, Gsponer J, Varnai P, Salvatella X, Vendruscolo M (2007) The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J Biomol NMR* 37: 117–135.
8. Levin EJ, Kondrashov DA, Wesenberg GE, Phillips Jr GN (2007) Ensemble refinement of protein crystal structures: Validation and application. *Structure* 15: 1040–1052.
9. Rieping W, Habeck H, Nilges M (2005) Inferential structure determination. *Science* 309: 303–306.
10. Furnham N, Blundell TL, DePristo MA, Terwilliger TC (2006) Is one solution good enough? *Nature Struct Mol Biol* 13: 184–185.
11. Goldsmith-Fischman S, Honig B (2003) Structural genomics: Computational methods for structure analysis. *Prot Sci* 12: 1813–1821.
12. Sierk ML, Pearson WR (2004) Sensitivity and selectivity in protein structure comparison. *Prot Sci* 13: 773–785.
13. Brüschweiler R (2003) Efficient RMSD measures for the comparison of two molecular ensembles. *Proteins* 50: 26–34.
14. Rueda M, et al. (2007) A consensus view of protein dynamics. *Proc Natl Acad Sci USA* 104: 796–801.
15. Hess B (2000) Similarities between principal components of protein dynamics and random diffusion. *Phys Rev E* 62: 8438–8448.
16. Hess B (2002) Convergence of sampling in protein simulations. *Phys Rev E* 65: 031910.
17. Bonvin AMJJ, Brünger AT (1995) Conformational variability of solution nuclear magnetic resonance structures. *J Mol Biol* 250: 80–93.
18. Bonvin AMJJ, Brünger AT (1996) Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? *J Biomol NMR* 7: 72–76.
19. Lyman E, Zuckerman DM (2006) Ensemble-based convergence analysis of biomolecular trajectories. *Biophys J* 91: 164–172.
20. Krzanowski WJ (2003) Non-parametric estimation of distance between groups. *J Appl Stat* 30: 743–750.
21. Zhou S, Chellappa R (2006) From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Trans Pattern Anal Mach Intell* 28: 917–929.
22. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Statist* 22: 79–86.
23. Cover TM, Thomas JA (1991) *Elements of information theory*. New York: Wiley.
24. Qian H (2001) Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Phys Rev E* 63: 042103.
25. Wall ME (2006) Ligand binding, protein fluctuations, and allosteric free energy. Clemente-Gallardo J, Moreno Y, Lorenzo JFS, Velazquez-Campoy A, eds. New York, USA: American Institute of Physics. pp 16–33.
26. Jeffreys H (1946) An invariant form for the prior probability in estimation problems. *Proc R Soc Lond A* 186: 453–461.
27. Sibson R (1969) Information radius. *Z Wahrscheinlichkeitstheorie verw Geb* 14: 149–160.
28. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 37: 145–151.
29. Österreicher F, Vajda I (2003) A new class of metric divergences on probability spaces and its applicability in statistics. *Ann Inst Statist Math* 55: 639–653.
30. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* 4: Article 32.
31. Opgen-Rhein R, Strimmer K (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat Appl Genet Mol Biol* 6: Article 9.
32. Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1: 54–75.
33. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. *Science* 315: 972–976.
34. Agrafiotis DK, Xu H (2002) A self-organizing principle for learning nonlinear manifolds. *Proc Natl Acad Sci USA* 99: 15869–15872.
35. Guha R (2007) C-implementation of Stochastic Proximity Embedding.
36. Parzen E (1962) On estimation of probability density function and mode. *Ann Math Statist* 33: 1065–1076.
37. Ihler A (2003) Kernel Density Estimation Toolbox for Matlab.
38. Gray AG, Moore AW (2003) Very fast multivariate kernel density estimation via computational geometry. San Francisco: American Statistical Association.
39. Hall P, Sheather SJ, Jones MC, Marron JS (1991) On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* 78: 263–269.
40. Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular dynamics system. *J Chem Phys* 105: 1902–1921.
41. Lazaridis T, Karplus M (1999) Effective energy function for protein dynamics and thermodynamics. *Proteins* 35: 133–152.
42. MacKerell Jr AD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102: 3586–3616.
43. MacKerell Jr AD, Feig M, Brooks III CL (2004) Improved treatment of the protein backbone in empirical force fields. *J Am Chem Soc* 126: 698–699.
44. Im W, Lee MS, Brooks III CL (2003) Generalized born model with a simple smoothing function. *J Comp Chem* 24: 1691–1702.
45. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J Comput Phys* 23: 327–341.
46. Tropp J (1980) Dipolar relaxation and nuclear Overhauser effects in nonrigid molecules: The effect of fluctuating internuclear distances. *J Chem Phys* 72: 6035–6043.
47. Wüthrich K (1986) *NMR of Proteins and Nucleic Acids*. New York: Wiley.
48. Brüschweiler R, et al. (1992) Influence of rapid intramolecular motion on NMR cross-relaxation rates. A molecular dynamics study of antamanide in solution. *J Am Chem Soc* 114: 2289–2302.
49. Schneider T, Brünger AT, Nilges M (1999) Influence of internal dynamics on accuracy of protein NMR structures: Derivation of realistic model distance data from a long molecular dynamics trajectory. *J Mol Biol* 285: 727–740.
50. Post CB (1992) Internal motional averaging and three-dimensional structure determination by nuclear magnetic resonance. *J Mol Biol* 224: 1087–1101.
51. Jorgensen WJ, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79: 926–935.
52. Beglov D, Roux B (1995) Dominant solvation effects from the primary shell of hydration — approximation for molecular-dynamics simulations. *Biopolymers* 35: 171–178.
53. Paci E, Karplus M (1999) Forced unfolding of fibronectin type 3 modules: An analysis by biased molecular dynamics simulations. *J Mol Biol* 288: 441–459.
54. Brooks BR, et al. (1983) CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J Comp Chem* 4: 187–217.
55. Brooks BR, Janežič D, Karplus M (1995) Harmonic analysis of large systems. I. Methodology. *J Comp Chem* 16: 1522–1542.
56. Kullback S (1952) An application of information theory to multivariate analysis. *Ann Math Statist* 23: 88–102.
57. Mahalanobis P (1936) On the generalised distance in statistics. *Proc Natl Inst Sci India* 12: 49–55.
58. Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal* 88: 365–411.
59. Rother D, Sapiro G, Pande V (2007) Statistical characterization of protein ensembles. *IEEE/ACM Trans Comput Biol Bioinform*. (In press).
60. Levitt M (1983) Molecular dynamics of native protein. II. Analysis and nature of motion. *J Mol Biol* 168: 621–657.
61. Kitao A, Go N (1999) Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol* 9: 164–169.
62. Hou J, Sims GE, Zhang C, Kim SH (2003) A global representation of the protein fold space. *Proc Natl Acad Sci USA* 100: 2386–2390.
63. Ding C, He X (2004) K-means clustering via principal component analysis. Machine learning, Proceedings of the twenty-first international conference (ICML 2004), Banff, Alberta, Canada.
64. Carugo O (2007) Statistical validation of the root-mean-square-distance, a measure of protein structural proximity. *Prot Eng Des Sel* 20: 33–37.
65. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290: 2319–2323.
66. Lindorff-Larsen K, Rogen P, Paci E, Vendruscolo M, Dobson CM (2005) Protein folding and the organization of the protein topology universe. *Trends Biochem Sci* 30: 13–19.
67. Das P, Moll M, Stamati H, Kavrakli LE, Clementi C (2006) Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA* 103: 9885–9890.
68. Zagrovic B, van Gunsteren WF (2006) Comparing atomistic simulation data with the NMR experiment: How much can NOEs actually tell us? *Proteins* 63: 210–218.
69. Ming D, Wall ME (2005) Quantifying allosteric effects in proteins. *Proteins* 59: 697–707.
70. Kondrashov DA, Van Wynsbeghe AW, Bannen RM, Cui Q, Phillips GN (2007) Protein structural variation in computational models and crystallographic data. *Structure* 15: 169–177.
71. Schlitter J (1993) Estimation of absolute and relative entropies of macromolecules using the covariance-matrix. *Chem Phys Lett* 215: 617–621.

72. Andricioaei I, Karplus M (2001) On the calculation of entropy from covariance matrices of the atomic fluctuations. *J Chem Phys* 115: 6289–6292.
73. Grossfield A, Feller SE, Pitman MC (2007) Convergence of molecular dynamics simulations of membrane proteins. *Proteins* 67: 31–40.
74. Amadei A, Linssen AB, Berendsen HJ (1993) Essential dynamics of proteins. *Proteins* 17: 412–425.
75. Chalaux FR, O'Donoghue SI, Nilges M (1999) Molecular dynamics and accuracy of NMR structures: Effects of error bounds and data removal. *Proteins* 34: 453–463.
76. Latzer J, Eastwood MP, Wolynes PG (2006) Simulation studies of the fidelity of biomolecular structure ensemble recreation. *J Chem Phys* 125: 214905.
77. Allen LR, Paci E (2007) Transition states for protein folding using molecular dynamics and experimental restraints. *J of Phys: Condensed Matter* 19: 285211.
78. Burling FT, Weis WI, Flaherty KM, Brünger AT (1996) Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science* 271: 72–77.
79. Clore GM, Schwieters CD (2006) Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small  $\alpha/\beta$  protein: A unified picture of high probability, fast atomic motions in proteins. *J Mol Biol* 355: 879–886.
80. Murdock SE, et al. (2006) Quality assurance for biomolecular simulations. *J Chem Theory Comp* 2: 1477–1481.
81. Beck DA, Armen RS, Daggett V (2005) Cutoff size need not strongly influence molecular dynamics results for solvated polypeptides. *Biochemistry* 44: 609–616.
82. Zidek L, Novotny MV, Stone MJ (1999) Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nature Struct Biol* 6: 1118–1121.
83. Boehr DD, McElheny D, Dyson HJ, Wright PE (2006) The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313: 1638–1642.
84. Kern D, Zuiderweg ER (2003) The role of dynamics in allosteric regulation. *Curr Opin Struct Biol* 13: 748–757.