J Biomol NMR (2014) 58:209–225
DOI 10.1007/s10858-014-9818-2

ARTICLE

Enhancing the quality of protein conformation ensembles with relative populations

Vijay Vammi · Tu-Liang Lin · Guang Song

Received: 28 August 2013 / Accepted: 31 January 2014 / Published online: 12 February 2014

Abstract The function and dynamics of many proteins are best understood not from a single structure but from an ensemble. A high quality ensemble is necessary for accurately delineating protein dynamics. However, conformations in an ensemble are generally given equal weights. Few attempts were made to assign relative populations to the conformations, mainly due to the lack of right experimental data. Here we propose a method for assigning relative populations to ensembles using experimental residue dipolar couplings (RDC) as constraints, and show that relative populations can significantly enhance an ensemble's ability in representing the native states and dynamics. The method works by identifying conformation states within an ensemble and assigning appropriate relative populations to them. Each of these conformation states is represented by a sub-ensemble consisting of a subset of the conformations. Application to the ubiquitin X-ray ensemble clearly identifies two key conformation states, with relative populations in excellent agreement with previous work. We then apply the method to a reprotonated ERNST ensemble that is enhanced with a switched conformation, and show that as a result of population reweighting, not only the reproduction of RDCs is significantly improved, but common conformational features (particularly the dihedral angle distributions of ϕ_{53} and ψ_{52}) also emerge for both the X-ray ensemble and the reprotonated ERNST ensemble.

Keywords Residual dipolar couplings · Ubiquitin · Relative populations · Boltzmann weights · Weighted ensemble · Ensemble quality

V. Vammi (✉) · G. Song (✉)
Department of Computer Science, Bioinformatics and Computational Biology Program, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011, USA
e-mail: vsvammi@iastate.edu

G. Song
e-mail: gsong@iastate.edu

T.-L. Lin
Department of Management Information Systems, National Chiayi University, 580 Sinmin Rd., Chiayi City 600, Taiwan

G. Song
Baker Center for Bioinformatics and Biological Statistics, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011, USA

Introduction

The functions of a protein are closely related to not only its structure but also its dynamics. For more and more proteins, it is becoming increasingly evident that their functional behavior is best understood not through one single structure but through the distribution and dynamic transition among a number of conformation states that form the native-state ensemble (Austin et al. 1975; Boehr et al. 2009; DePristo et al. 2004; Frauenfelder et al. 2001; Furnham et al. 2006; Karplus and McCammon 2002; Levin et al. 2007; Phillips 2009). Such an ensemble representation is consistent with the energy landscape theory and the 'protein folding funnels' (Dill and Chan 1997; Frauenfelder et al. 1991; Miyashita et al. 2003). With the rapidly growing Protein Data Bank (PDB) (Berman et al. 2000), protein structures are becoming increasingly more available and for some well-studied proteins, tens and even hundreds of structures (of one same protein) have been determined. These structures have been shown to capture a representative subset of the native-state ensemble (Best et al. 2006).

On the other hand, the advancement of the experimental techniques and the increasing availability of experimental data has brought also a number of exciting recent works that aim to determine protein conformation ensembles instead of a single structure, using the experimental data as constraints (Clore and Schwieters 2004a, b, 2006; Lange et al. 2008; Lindorff-Larsen et al. 2005; Richter et al. 2007, Fenwick et al. 2011). The extent to which some of these ensembles represent the native states is debatable since the ensemble, which in some cases contains over a hundred conformations, may be under-constrained by the experimental data. As a matter of fact, since the experimental observations and data are macroscopic in nature and represent the ensemble and time averages of microscopic conformations, it may not be possible to verify the validity of each conformation individually in such ensembles. Indeed, the concern of most of these ensembles was mostly about representing the dynamics correctly, less about the validity of each individual conformation.

For all of the above ensemble determination protocols, the conformations within the ensemble were given equal weights, i.e., $1/N_e$, where N_e is the size of the ensemble. While weights were listed out as part of the parameters in some of these methods, weights other than equal weights were not studied. Physically these weights represent relative populations of the conformations and thus their relative contributions to the ensemble.

There are a couple of reasons why few work has been carried out to exploit the potential benefit of including these weights (or relative populations). First, an elegant seminal work by Brunger's group had shown earlier that regular NOE data alone was insufficient to determine the relative populations of a two-conformer ensemble (Bonvin and Brunger 1996). Thus it was not clear if there were enough experimental data to determine the populations meaningfully, even though the authors (Bonvin and Brunger 1996) were hopeful that relative populations could possibly be determined when other sources of experimental data were provided. Secondly, equal-weight conformations themselves can capture the relative population information to some extent, by including in the ensemble multiple copies of one similar conformation. The number of copies thus indirectly encodes the weight. However, it is an insufficient way to represent the populations, as it requires more conformations to be in the ensemble and thus may worsen the potential problem of over-fitting mentioned above.

Our hypothesis

In this work we propose that it is feasible to assign relative populations to ensembles by using experimental RDC data as constraints, and that adding relative populations should enhance an ensemble's ability in representing a protein's native states and its dynamics.

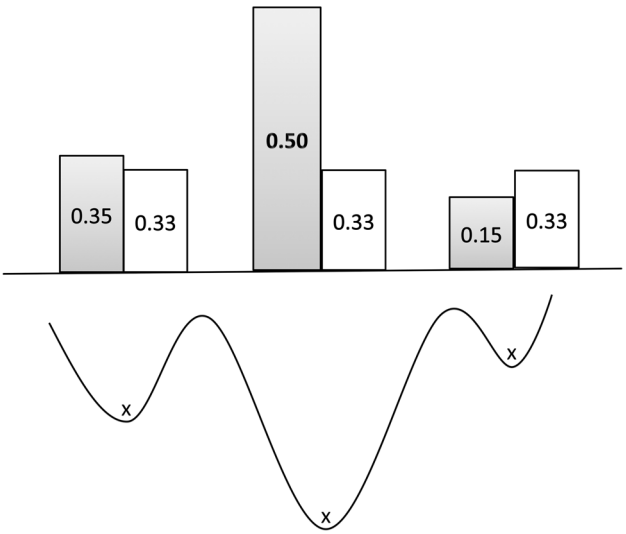

Fig. 1 Pictorial representation of Boltzmann weights versus sampling weights. The *'x' marks* represent conformations on a hypothetical energy landscape. The *white bars* represent sampling weights while the *shaded bars* represent the Boltzmann weights

Assigning relative populations to an ensemble: background and problem definition

For many a protein, the conformation space near its native states can be best represented by a number of inter-connected conformation states, each of which may have a different population, as illustrated in Fig. 1. When an ensemble of conformations are used to represent the conformation space (shown as the cross marks in Fig. 1), its quality in representing the conformation space is determined by three factors:

1. COMPLETENESS: Are all conformation states reached by at least one conformation?
2. COVERAGE: For each of the conformation states that are reached by some conformation(s), what is quality of the coverage? In other words, how well do the finite number of conformations that are in a given conformational state together represent that conformation state?
3. CONTRIBUTION: Is the number of conformations at each conformation state proportional to the ideal Boltzmann weight?

Ideally, we would like to have an ensemble that has an infinite number of conformations that cover all the conformation states according to the Boltzmann distribution. Such an ensemble would have perfect completeness, coverage and contribution. In reality, our ensembles are of

finite sizes, having tens or possibly hundreds of conformations, which are relatively small comparing to the large conformation space. Therefore, we do not have perfect completeness, coverage or contribution.

Another key point to realize is that the matter of completeness and coverage are sampling issues. The conformations in an ensemble could have come from experiments, by structure determination methods such as X-ray crystallography, NMR, etc., or they could have been determined computationally. Whatever the source is, completeness and coverage are sampling issues. They reflect the sampling quality of a given ensemble.

However, how well an ensemble represents the conformation space near the native state and how well it can reproduce experimental data/observations are determined not solely by the ensemble's completeness or coverage. It depends also on the third factor—contribution. Without doubt an ensemble whose conformations are assigned a population (contribution) proportional to their actual Boltzmann weights would represent the conformation space the best, reaching the limit of that ensemble's ability in representing the conformation space. Therefore, an ensemble with a proper assignment of relative contributions given to its conformation states should do better than an ensemble without. As illustrated in Fig. 1, the conformation space of a protein is represented by three conformations, which by default are given an equal weight of 1/3. However, the ensemble can be enhanced if the actual Boltzmann weights (represented by dark shaded blocks) can be determined somehow and assigned to the three conformations. Now, the immediate questions are: are relative contributions even determinable? And if so, how? And what is required to determine them?

In this work, our focus is on this third aspect—contribution. Our hypothesis is that given an ensemble of reasonable quality in completeness and coverage, the relative contributions can be determined by using experimental RDC data as constraints.

We will apply a least-square fitting algorithm to determine the weights. To prevent over-fitting, careful cross-validation is applied. In the following "Materials and methods" section, we present our approach in details.

Materials and methods

Recall that the problem we want to address here is that, given a conformation ensemble and a sufficient amount of experimental RDC data, is it possible to assign meaningful populations to the conformations in the ensemble without incurring over-fitting? To what extent can we assign the populations? There are two extremes. One extreme is to assign each conformation with a population, which is physically unrealistic and generally cannot be achieved. The other is to assign the whole ensemble as a group with a (percentage) population of 1. This is equivalent to equal weights that have been used. Our hypothesis is that sufficient experimental data should allow weight assignment to clusters of conformations, or sub-ensembles, within the ensemble.

In this section, we present our method for assigning relative populations to clusters of conformations within an ensemble. The potential problem of over-fitting that often arises in such a process is carefully addressed. The significance of the assigned relative populations is further examined by cross-validation.

There are four major steps in our method, which are described in order in the following sections. Briefly, the first step, a pre-processing step, merges conformations in the ensemble into small conformation clusters. For ensembles whose sizes are small, this step is skipped. The second step takes the pre-processed ensemble and applies a least squares fitting algorithm to identify a subset of conformations/clusters that best represent the conformation states. Step three takes this subset as a whole and iteratively split it into smaller sets until right before over-fitting starts to occur. Step four attempts to add back to the ensemble some conformations excluded in earlier steps. Lastly, the significance of the relative populations thus assigned is evaluated by cross-validation.

Step I: Pre-processing to reduce the dimensionality of the ensemble

In cases where the ensemble size is large and it has more conformations than the number of experimental RDC data points, clustering (Daura et al. 1999; Shao et al. 2007) is carried out to reduce the dimensionality of the ensemble. Here the dimensionality of an ensemble refers to the structural variety of the ensemble and is set to be the number of clusters in the ensemble. Initially each conformation in the ensemble forms its own cluster. Clustering structurally similar conformations into small clusters thus helps reduce the dimensionality and makes the ensemble manageable for the least square fitting procedure to be applied in the next step.

The distance between a pair of clusters is defined as the average of all the pairwise distances between the conformations in the two clusters. The distance between two conformations is defined by Q_{score}.

$$Q_{score} = \frac{\left(\sum_{\{i!=j\}} \exp\left(-\left(r^{A}_{\{i,j\}} - r^{B}_{\{i,j\}}\right)^2\right)\right)}{N(N-1)} \quad (1)$$

where $r_{i,j}$ is the distance between atoms i and j in a conformation and N is the total number of atoms. Q_{score} value ranges from 0 to 1, 0 being the very dissimilar and 1 being perfectly similar (Eastwood et al. 2001).

Initially each conformation in the ensemble forms its own cluster. The following three steps are iterated. As a result, similar conformations will be bundled together into larger clusters, while the rest remain as singlet clusters.

1. Identify the closest pair of conformations in the ensemble. Merge them into a cluster if their distance is less than a threshold, D_{max}. Otherwise stop the procedure.
2. Grow the cluster formed in step 1 by repeatedly adding to it the next conformation whose average distance to the conformations in the cluster is the smallest and is less than D_{max}, otherwise stop adding.
3. Remove the cluster and go back to step 1.

Step II: Identify representative conformations by least-square fitting to RDCs

Residual dipolar coupling comes from the interaction of two nuclear spins (dipole–dipole) in the presence of the external magnetic field and is defined as (Kontaxis and Bax 2001; Prestegard 1998; Tolman et al. 1995):

$$D_{\{ij\}} = -\frac{\mu h r_i r_j}{(2\pi r)^3}\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle \quad (2)$$

where r_i and r_j are the nuclear magnetogyric ratios of nuclei i and j respectively, h is Plank's constant, μ is permittivity of space, r is the internuclear distance between the two nuclei and Θ is the angle between the internuclear vector and the external magnetic field. The brackets represent the ensemble and time average. Normally, the residual dipolar coupling reduces to zero because of isotropic tumbling. The anisotropic measurement can be obtained by the aid of various types of liquid crystalline media.

For a protein with a number of distinct conformation states, the observed RDC data are best reproduced when the conformations close to these conformation states are present in the ensemble and given proper weighting. The conformations in a given ensemble may not all fall close to a conformation state. Here we use least square fitting to identify which conformations are needed and what relative populations should be given to them in order to best reproduce the experimental RDC data. By doing this, we can pick out key representative conformations from the ensemble. The relative populations assigned to them, however, are subject to the problem of over-fitting, due to the intrinsic nature of least square fitting. However, measures will be taken to identify the onset of over-fitting and prevent it from affecting weight assignment, as addressed in step III.

Appendix 1 describes how RDCs can be back calculated from a single conformation or an ensemble of conformations. In this process of back calculating, singular value decomposition is commonly used to obtain the least square solution for the alignment tensor. Here we apply the same technique iteratively to obtain the least square solution for the relative populations as well. First, equal weights ($1/n$) are given to all clusters (which are determined at step I) and Eq. 13 (see Appendix 1) is used to obtain the optimal Saupe matrix, S. After S is obtained, it is used to determine $w'_k s$ by least squares fitting. The process is iterated until the weights have converged. In the end, each cluster has either positive or zero population, since the weights are derived under the nonnegative constraints (Lawson and Hanson 1995). In the case where there are multiple RDC data sets, different alignment tensors are needed for different media. The optimal weight combination (the relative populations) is obtained by least squares fitting to all the RDC data sets. A detailed description of these iterative least squares fitting algorithms is given in Appendix 2.

The iterative least squares fitting of the conformations in the ensemble to multiple RDC datasets returns a list of clusters/conformations that have non-zero populations. The conformations in these clusters are recognized as representative conformations.

In cases where there are more conformational clusters than the experimental data points, representative clusters are identified through the following procedure.

1. From the pool of all available conformational clusters, randomly select N clusters, where N is the number of experimental data points.
2. Run the least squares fitting algorithm (Appendix 2) to determine cluster weights. Some clusters may have zero weights.
3. Repeat steps 1 and 2 many times and record the cluster weights at each iteration.
4. The top N clusters with the highest average frequencies of having non-zero weights are identified as representative clusters.

The representative clusters form the leaf nodes of a hierarchal clustering tree, built bottom up by merging the closest pair of clusters at each iteration.

Step III: Splitting and the identification of over-fitting

To avoid the potential problem of over-fitting that may take place in the process of assigning relative populations, we take steps to recognize the onset of over-fitting and prevent it from affecting the weight assignment. Recall that there are two extremes in assigning weights. One is to assign each conformation with a population. The other is to assign

the whole ensemble as a group with a population of 1, which is equivalent to having equal weights. In our studies we have found that one may confidently move beyond equal weighting and assign relative (different) populations to sub-ensembles but not to the point that each conformation in the ensemble is given a weight. There exists a limit where one cannot further divide the sub-ensembles into smaller pieces. This limit represents the extent to which relative populations can be assigned and it depends on the quality of the ensemble and the quality and quantity of the experimental data. In reality, the limit is determined through monitoring the onset of over-fitting.

In the following procedure, we iteratively split the ensemble, which is now made up of the representative conformations, into smaller and smaller clusters. The splitting process is the same as the inverse process of hierarchical clustering. At each iteration, only one cluster is spit into two, which corresponds to the merging of the closest pair of clusters in hierarchical clustering. Therefore there are k clusters at the kth iteration. By applying the least squares fitting algorithms as described in Appendix 2, we can assign relative populations (or weights) to these k clusters.

If we have N sets of experimental RDC data that are consistent with one other and contain random measurement noise within them, N sets of weights will be assigned to the k clusters. Now if the weight assignment is correct, we expect that these N sets of weights should strongly correlate with one another. The onset of overfitting is when such correlations start to greatly degrade. That is, it begins to fit to the noise. Since noise is random and uncorrelated in the different experimental data, the weights fitting to noise should also be uncorrelated. This recognition of the onset of over-fitting is even more sensitive when the correlations are computed using only the weights of the two newly birthed clusters at the kth iteration. The idea is that, if the two newly birthed clusters belong to one conformation state and should not have been split, we expect the weights assigned to them by different sets of experimental data should be ambiguous and lack consistency and thus low correlations. On the other hand, if these two clusters belong to difference conformation states and should be split, we expect to see consistent weight assignments from different experimental datasets and thus high correlations.

Replicate experimental data for over-fitting identification

To identify over-fitting as outlined above, all the experimental data is duplicated to create N identical copies and then different random Gaussian noise are added to each of them. These N datasets are thus identical except for the noise in them.

A relatively large N is needed to have a high sensitivity to the onset of over-fitting. N is set to be 20 in this work. The standard deviation (SD) of the random Gaussian noise added to each replica is set to be 80 % of the modeled experimental noise, which are bond-dependent and are set to be 0.26, 0.1, 0.5, 0.1 and 0.1 Hz for NH, CaC, CaHa, CN and CHN datasets respectively as was done in (Clore and Schwieters 2004a).

We use Q-factor to measure how well the weight assignments are correlated with one another. The definition of Q-factor is given in Eq. 3, where it is employed also to measure the similarity between experimental and computed RDC data. The maximum of the Q-factors between any two of the N weight assignments is denoted as MaxQ. A large MaxQ (above a certain threshold) indicates inconsistent weight assignments and thus over-fitting for the two newly birthed clusters. A threshold value of 0.06 is used for MaxQ throughout all the cases investigated below. In summary, the procedure is:

1. Initially all the representative conformations belong to one single cluster.
2. Experimental data is replicated into N sets. $N = 20$.
3. Iteratively split the clusters (the exact inverse process of a hierarchical clustering).
4. Assign sets of weights to clusters based on fitting to the experimental datasets.
5. Check if the weights assigned to the newly birthed clusters are significant (i.e., weight $\geq$0.01). If any weight is found to be insignificant, repeat the process by removing the insignificant cluster.
6. Compute the weight correlations and MaxQ for the two newly birthed clusters.
7. If the minimum of the weight correlations is negative and MaxQ is greater than a predefined threshold, it signifies that over-fitting has occurred. In this case, the two newly birthed clusters are merged back together and the cluster is marked "final", indicating that it can no longer be split. Otherwise, continue and move on to the next iteration. Stop the procedure when there is no cluster left that can be split.

Step IV: Adding back other conformations

By the end of step III, we have partitioned the ensemble into a number of "final" clusters, with N sets of weights assigned to each of them. Now compute the mean weight value and the SD for each cluster. The clusters whose mean weight value is less than its SD are then removed, as they do not consistently have a positive weight.

Each of the remaining clusters is considered as representing an independent conformation state. Since it remains possible that the conformations that were excluded earlier at steps I to III may belong to one of the conformational states that these clusters are representing, adding some of

them back to the clusters thus may possibly improve the quality of the ensemble. The sequence in which conformations are added back is arranged, in increasing order, by the minimum distance between a conformation and any of the clusters. A conformation is added to the cluster to which it is the closest if including it decreases the overall Q-factor.

Estimate the uncertainty in weight assignments

After the conformation states (i.e., the clusters) have been identified and weights assigned to them, it is possible to estimate the uncertainty in the weight assignments, provided that there exist multiple sets of experimental data. This is because least squares fitting can be applied to fit each set of experimental data independently. If there are M sets of experimental data, this will result in M sets of weight assignments, or M weight assignments to each cluster. It is expected that the weight assignments for each cluster are in general not identical, since there is noise in the experimental data and the cluster representation for each conformation state is not perfect. The levels of uncertainty in the weight assignments can be estimated by computing the SD within the weight assignments for each cluster.

Cross-validation

Q-factor is a commonly used measure of the agreement between the experimental and calculated RDCs and is defined as:

$$Q - factor = \frac{\sqrt{\sum (D_{calc} - D_{\exp})^2}}{\sqrt{\sum (D_{\exp})^2}} \quad (3)$$

where D_{calc} is the calculated RDC and D_{exp} is the experimental RDC.

The introduction and assignment of relative populations to an ensemble improves the Q-factors. To assess the significance of such improvement, we leave out CaHa RDC from the experimental data when determining the weights. The CaHa dataset was then used for cross-validation. Lange et al. (2008) used CN vector for cross-validation. Given that the data used in refinement includes CaC, CHN, NH vector orientations, CN RDC might not be the best choice. CaHa vector, on the other hand, is not in the peptide plane and is thus independent of other bond vector orientations, making it a better cross-validation dataset. Cross-validation provides a way to check whether the better fitting gained by assigning relative populations is a fitting to the noise in experimental data or is a fitting to the true data. If the ensemble with relative population assignment does render a better representation of the conformation space, we expect that the fitting to the leaving-out CaHa data should also improve.

All the conformations are stripped off its hydrogen atoms first and then re-protonated using Reduce (Word et al. 1999) before computing Saupe matrices and back-calculating RDC values.

Table 1 RDC datasets used for weighting ubiquitin ensembles, coded according to Lakomek et al. (2008)

Experimental data type	RDC data
NH	A1, A2, A4, A6, A7, A8, A9, A10, A11, A12, A13, A16, A21, A22, A23, A24, A25, A26, A27, A28, A29, A34, A36
NH, CN, CHN, CaC and CaHa	(Ottiger and Bax 1998)(2 sets)

Ubiquitin ensembles and experimental RDC data sets

Ubiquitin has long been used as a model protein to probe protein dynamics and for which abundant experimental RDC datasets are available. A total of 62 RDC data sets, including NH, CN, CHN, CaC, CaHa and side chain methyl, were used to determine EROS ensemble (Lange et al. 2008). Since our procedure requires that the relative populations be determined by fitting to experimental RDC data, it is critical that the data has no significant errors. For this reason we have pruned the above dataset to remove any dataset whose data points are less than 40 and whose Q-factors are significantly higher when back-calculated using structure 1UBQ or 1D3Z (NMR ensemble).

Table 1 lists the experimental datasets used in this work, using the code names given in Lakomek et al. (2008). There exist a few other multi-vector datasets for ubiquitin (Lakomek et al. 2006). However, they are not included here since they display relatively large Q-factor when applied to the NMR structure 1D3Z. For the same reason, NH datasets labeled A3, A5, A30, A31, A32, A33, and A34 (as in Lakomek et al. 2008) are not included either.

Results

In this section, we apply our method to assign relative populations to conformation ensembles of proteins. It is assumed here that the protein that an ensemble represents should have a small number of conformation states, and that some of the conformations in the ensemble, though sparse relative to the large conformation space, fall close to the protein's conformation states. These conformations may come from experimentally determined structures of

the protein. Because of their scarcity, there is no expectation on these conformations that their distribution on the conformation space should be Boltzmann distribution. For such an ensemble, and using experimental RDC data as constraints, we will show to what extent one can meaningfully assign relative populations, or weights, to the ensemble. We aim to answer also, in order to assign meaningful relative populations, what is the minimum requirement on the ensemble. In the end, we apply the method to an ensemble of crystal structures of ubiquitin.

Creating an artificial conformation ensemble and artificial RDC data

To test our method, we first create an artificial energy landscape and a native state ensemble that will be used as a reference (Richter et al. 2007). We create also artificial RDC data based on the ensemble composition. The advantage of using artificial ensembles and RDCs is that we have perfect control of their composition and their noise level.

Creating an artificial native state ensemble

To create an artificial native state ensemble, five distinct conformations of protein ubiquitin are picked from an accelerated MD simulation (Hamelberg et al. 2004). The conformations are chosen such that the minimum RMSD between any two conformations is greater than 2.5 Å. We assume that these five conformations represent the centers of all the (five) possible conformational states of the protein. We then sample more conformations around these centers and use them, together with the centers, to represent the conformation states. This is done using CONCOORD (de Groot et al. 1997). CONCOORD, by default, can produce quite broad distributions of conformations. To ensure that each conformational state is tightly clustered, a damping coefficient of 0.3 is applied when generating the distance restraints from these five conformations. As a result, the average RMSD within any sub-ensemble is close to 1 Å. Thus, the conformations fall into five clearly separated clusters.

Next, we set the Boltzmann weight of each conformation state to be proportional to the number of conformations in its energy well (i.e., the sub-ensemble around each conformation state). The number of conformations sampled in each sub-ensemble and the associated Boltzmann weights are given in Table 2.

Noise conformations

Noise conformations are those that do not contribute to experimental observations. Strictly speaking though, every conformation in the ensemble contributes to the observations to some extent. But those conformations that are away from any of the protein's conformation states have so low a weight that they virtually do not contribute. We consider such conformations as noise conformations as contrast to those that do represent the protein's conformation states.

To create noise conformations, we use CONCOORD to sample around each conformational state without any damping. The average RMSD in this sampling is around 2.5 Å. To guarantee these conformations do represent noise, we remove from them any conformations that can give nearly the same RDC Q-factors as the conformations representing the conformation states.

Generating artificial RDC data

Using all the conformations (1,850 total, see Table 2) of the ensemble, artificial RDC datasets matching the composition of the real experimental RDC data of ubiquitin, are generated. The average A matrix of the ensemble is first calculated. Then for each of the experimental datasets listed in Table 1 the best-fit Saupe matrix is determined using 1D3Z NMR ensemble. An artificial RDC dataset is then created by multiplying the average A matrix with the Saupe matrix. At this point, these RDC datasets are noise-free. We will call them noise-free RDCs.

In reality, experimental data contains noise of about 0.5–1.0 Hz (Clore and Schwieters 2006), we add Guassian noise to the artificially generated RDC data that are originally noise-free. The SDs of the noise are 0.26, 0.1, 0.5, 0.1 and 0.1 Hz for NH, CaC, CaHa, CN and CHN datasets respectively as was done in Clore and Schwieters (2004a). Note that because of the way in which the artificial RDC data are generated, the given conformation ensemble can perfectly reproduce these RDC data prior to the adding of the noise, but not so after. In the rest of this article, unless explicitly noted, artificial RDCs refer to the ones that contain noise.

What is required of the ensemble?

In the section we aim to determine what is the requirement of the ensemble in order to have a meaningful weight assignment. We design four test cases to examine the applicability of the method. The purpose of these four cases is to show that neither under-sampling at each conformation state nor noise conformations hinder weight assignments.

Case I

In this case we assume there is no noise conformations and the ensemble contains only conformations from the five

Table 2 Boltzmann weights of the five conformational states in the artificial ensemble

Conformational state	One	Two	Three	Four	Five	Total
# of Conformations	100	200	350	500	700	1,850
Boltzmann weight	0.054	0.108	0.189	0.27	0.378	1

Table 3 Final weights and cluster compositions for case I

Cluster	Final weight ± SD	Composition	Belongs to	Expected weight
Cluster 1	0.072 ± 0.001	20,0,0,0,0	First state	0.054
Cluster 2	0.097 ± 0.0004	0,8,0,0,0	Second state	0.108
Cluster 3	0.183 ± 0.002	0,0,5,0,0	Third state	0.189
Cluster 4	0.27 ± 0.002	0,0,0,7,0	Fourth state	0.27
Cluster 5	0.376 ± 0.001	0,0,0,0,284	Fifth state	0.378

The convention used for the composition of a cluster is to enumerate in order the number of conformations belonging to each of the five conformational states

conformation states. However, the number of conformations at each state is not proportional to its Boltzmann weight. 21, 60, 6, 7, 290 conformations are randomly selected from conformation state one, two, three, four, and five respectively and mixed together to form an ensemble. Our method is then applied to assign relative populations to this ensemble. Table 3 lists the clusters obtained in the end, along with the composition of the clusters, weights assigned and expected weights of all the clusters.

It is seen from Table 3 that the final weight obtained for each conformation cluster is highly similar to the expected Boltzmann weight and each cluster contains purely conformations that belong to that conformation state. Similar results are obtained when the same experiment is repeated with different replica noise.

Case II

In this case, one of the conformation states (the third) was intentionally not included in the process of generating artificial experimental data. This is done to mimic the scenario where an ensemble contains a cluster of conformations from a state that does not belong to the native ensemble. While the purpose for the first case is to test if the method is able to assign right populations to the conformations contributing to the experimental observations, the purpose for this one is to test whether or not the method is able to assign no weight to conformations that do not contribute.

The new relative Boltzmann weights are given in Table 4. The same conformation ensemble employed in case I, which includes conformations that do not contribute

Table 4 New relative Boltzmann weights after the third cluster is excluded from artificial RDC data generation

Conformational state	One	Two	Four	Five	Total
# of Conformations	100	200	500	700	1,500
Boltzmann weight	0.066	0.133	0.333	0.467	1

Table 5 Final weights and cluster compositions for case II

Cluster	Final weight ± SD	Composition	Belongs to	Expected weight
Cluster 1	0.087 ± 0.0001	9,0,0,0,0	First state	0.066
Cluster 2	0.118 ± 0.004	0,60,0,0	Second state	0.133
Cluster 3	0.322 ± 0.003	0,0,0,7,0	Fourth state	0.333
Cluster 4	0.471 ± 0.001	0,0,0,0,280	Fifth state	0.467

The convention used for the composition of a cluster is the same as Table 3

to the artificial RDC calculations, is used here. After applying our method, the resulting clusters, along with their compositions, assigned weights and the SDs, and expected weights are given in Table 5. From the results it is seen that, as with case I, the weights obtained are highly similar to the expected values. Moreover, each cluster consists purely of conformations belonging to that cluster.

Case III

In the first two cases, the conformations in the ensemble are clearly separated into five distinct clusters. In reality, such distinction is often smeared by the presence of other conformations. These other conformations virtually do not contribute to the experimental observations (the "noise" conformations). However, their presence makes it difficult to identify conformation states, or separate conformations representing a conformation state from those that do not. To mimic this reality, we introduce noise conformations into the ensemble.

The same conformations as used in case I are used here (see Table 2). In addition, an equal number of noise conformations (see above on how they are generated) are added to each cluster so that they represent half of the total conformations in each cluster. As a result, the number of conformations in the ensemble is doubled and becomes 768, of which 384 are noise conformations. Clustering, as

Table 6 Final weights and cluster compositions for case III

Cluster	Final weight ± SD	Composition	Belongs to	Expected weight
Cluster 1	0.069 ± 0.005	8,0,0,0,0	First state	0.054
Cluster 2	0.099 ± 0.001	0,8,0,0,0	Second state	0.108
Cluster 3	0.181 ± 0.003	0,0,5,0,0	Third state	0.189
Cluster 4	0.273 ± 0.005	0,0,0,7,0	Fourth state	0.27
Cluster 5	0.377 ± 0.002	0,0,0,0,281	Fifth state	0.378

The convention used for the composition of a cluster is the same as Table 3

described in step I in the "Materials and methods" section, results in 406 clusters, of which some are singlet clusters. Since the number of clusters is more than the number of unique experimental data points (around 200), "representatives" conformations are identified by following step II (see "Materials and methods").

Table 6 lists the results. There are five clusters, which are composed of 8, 8, 5, 7, and 281 conformations from the five conformational states respectively. All of the 384 noise conformations are successfully filtered out. From Table 6 it is seen that weight assignments for the clusters are highly similar to the expected values.

Case IV

In all of the above cases, we have simulated full coverage of the conformational states by having each of the states represented by at least a few conformations. To assess the impact on the reproduction of the experimental data when one of the conformational states is missing all together, we apply our weighting algorithm again to the ensemble used in case III but this time each of the five clusters used to represent the five conformational states, in turn, is purposely left out. We want to see if the algorithm will produce RDC Q-factors with equal quality, while having substantially different conformational properties than the initial ensemble, by somehow rearranging the weights for the remaining clusters. Table 7 lists the results.

From Table 7, it is seen that the algorithm produces RDC Q-factors with nearly the same quality especially when the missing cluster has a low population, such as cluster one or two. Even with cluster three or four, its missing causes only a small deterioration in Q-factors. In all these cases, most of the contributions of the missing cluster are compensated by the weight adjustment of the remaining clusters or by assigning weight to a new cluster(s) that is formed by some noise conformations. However, when the missing cluster has an especially large population such as that of cluster five, the algorithm cannot recover the RDC Q-factors with nearly the same quality. The results of this test case thus clearly demonstrate the importance of having a full coverage of all the conformational states and that low Q-factors alone are not sufficient to provide full confidence in the completeness or correctness of an ensemble.

X-ray ensemble and experimental data

X-ray structures of the same protein but solved under different conditions are hypothesized to form a native state ensemble of that protein (Best et al. 2006). 68 X-ray structures of ubiquitin with 100 % sequence identity are taken from PDB. After considering the fact that multiple chains exist in some of the structures, a total of 143 different conformations are identified and used to form the ubiquitin conformation ensemble. Table 8 lists all the PDB-ids along with their chain identifiers. To partition this ensemble into proper sub-ensembles and determine their relative populations, we follow the procedure described in the "Materials and methods" section and find that 18 out of 143 crystal structures have a significant weight and are chosen as representative conformations. This new ensemble of 18 crystal structures was then subjected to the splitting procedure and as a result, two more structures are

Table 7 New weight assignments and Q-factors when each of the five clusters, in turn, is purposely left out of the ensemble, as in case IV

	Weights					NH	CaC	CaHa	CHN	CN
	W1	W2	W3	W4	W5					
With none missing	0.07	0.10	0.18	0.27	0.38	0.036	0.051	0.034	0.067	0.04
With state one missing	–	0.11	0.20	0.31	0.37	0.042	0.059	0.041	0.072	0.054
With state two missing	0.10	–	0.26	0.26	0.36	0.047	0.056	0.054	0.074	0.052
With state three missing	0.06	0.19	–	0.31	0.35	0.066	0.067	0.079	0.082	0.08
With state four missing	0.23	0.0	0.26	–	0.27	0.08	0.074	0.084	0.102	0.091
With state five missing	0.0	0.0	0.0	0.40	–	0.202	0.15	0.17	0.166	0.155

CaHa is used for cross-validation. Note that the weights of the remaining four clusters do not add up to 1 in some cases. This happens when noise conformations form a new cluster(s) and are assigned a non-zero weight to compensate for the missing cluster

Table 8 PDB ids as well as chain identifiers of the 143 ubiquitin X-ray conformations used in this work to form the ubiquitin X-ray ensemble

1AAR-A, 1AAR-B, 1CMX-B, 1F9J-A, 1F9J-B, 1NBF-C, 1NBF-D, 1OGW-A, 1P3Q-U, 1P3Q-V, 1S1Q-B, 1S1Q-D, 1TBE-A, 1TBE-B, 1UBI-A, 1UBQ-A, 1UZX-B, 1WR6-E, 1WR6-F, 1WR6-G, 1WR6-H, 1WRD-B, 1XD3-B, 1XD3-D, 1YD8-U, 1YD8-V, 2AYO-B, 2C7M-B, 2C7N-B, 2C7N-D, 2C7N-F, 2C7N-H, 2C7N-J, 2C7N-L, 2D3G-A, 2D3G-B, 2DX5-B, 2FID-A, 2FIF-A, 2FIF-C, 2FIF-E, 2G45-B, 2G45-E, 2GMI-C, 2HD5-B, 2HTH-A, 2IBI-B, 2J7Q-B, 2J7Q-D, 2JF5-A, 2JF5-B, 2O6V-A, 2O6V-C, 2O6V-E, 2O6V-G, 2OOB-B, 2QHO-A, 2QHO-C, 2QHO-E, 2QHO-G, 2WDT-B, 2WDT-D, 2WWZ-A, 2WWZ-B, 2WX0-A, 2WX0-B, 2WX0-E, 2WX0-F, 2WX1-A, 2XEW-A, 2XEW-B, 2XEW-C, 2XEW-D, 2XEW-E, 2XEW-F, 2XEW-G, 2XEW-H, 2XEW-I, 2XEW-J, 2XEW-K, 2XEW-L, 2XK5-A, 2ZCC-C, 2ZNV-C, 3A1Q-A, 3A1Q-D, 3A33-B, 3A9J-B, 3A9K-B, 3ALB-A, 3ALB-B, 3ALB-C, 3ALB-D, 3BY4-B, 3C0R-B, 3C0R-D, 3EEC-A, 3EEC-B, 3EFU-A, 3EHV-B, 3EHV-C, 3H1U-A, 3H1U-B, 3H7P-B, 3H7S-A, 3H7S-B, 3HM3-A, 3HM3-B, 3HM3-C, 3HM3-D, 3I3T-B, 3I3T-D, 3I3T-F, 3I3T-H, 3IFW-B, 3IHP-C, 3IHP-D, 3JSV-B, 3JVZ-X, 3JVZ-Y, 3JW0-X, 3JW0-Y, 3K9P-B, 3KVF-B, 3KW5-B, 3LDZ-E, 3LDZ-F, 3LDZ-G, 3M3J-A, 3M3J-B, 3M3J-C, 3M3J-D, 3M3J-E, 3M3J-F, 3MHS-D, 3NHE-B, 3NOB-B, 3NOB-C, 3NOB-D, 3NOB-E, 3NOB-F, 3NOB-G, 3NOB-H

removed due to their near-zero weights and six clusters are identified. The rest of the 125 structures, one by one, are then tried to be merged into one of the six existing clusters but none gets added.

The resulting conformation clusters along with their weights are given in Table 9. The cluster that contains the unbounded conformation of ubiquitin, 1UBQ, is found to have the largest weight of $\sim$55 %, while the second clusters, consisting exclusively of ubiquitin structures in complex with deubiquitinating enzymes, has the second largest relative population of $\sim$29 %.

While we were working on this manuscript, one work was published in an early edition of PNAS (Piana et al. 2013). The work studied the native equilibrium dynamics of ubiquitin and reported that the protein conformation was exceptionally stable with $\sim$70 % of populated states about 0.5 Å RMSD away from the native state 1UBQ while $\sim$20 % of the populated states showed a conformational switch in Asp52/Gly53/Glu24 residues, referred to as "switched" conformer and the remaining $\sim$10 % had partially frayed alpha helix at the C-terminus (Piana et al. 2013). Our results as shown in Table 9 agree with their findings extremely well. In addition, another recent study of conformational states of ubiquitin found the presence of an alternative conformer in complex with deubiquitinating enzymes. In the authors' own words, "This alternative conformer is likely to have functional significance, because the Asp52/Gly53/Glu24 switched conformer is also found in structures of ubiquitin, ubiquitin aldehyde, or diubiquitin in complex with deubiquitinating enzymes (e.g., PDB entries 2G45, 2HD5, 2IBI, 1NBF, 3I3T, 3IHP, 3NHE, 3MHS, and proximal ubiquitin of 2ZNV, which are all discussed further below). In contrast, the un-switched conformer is seen in essentially all other ubiquitin structures, including the previous structures for monomeric ubiquitin, di- and tetra-ubiquitin, and complexes with other kinds of enzymes" (Huang et al. 2011). Our method not only identifies this special conformation state of ubiquitin (the 2nd cluster in Table 9), but also assigns it an accurate relative population. Several of the PDB entries for ubiquitin in complex with deubiquitinating enzymes are selected and grouped together by our algorithm to form cluster 2, a cluster consisting exclusively of ubiquitin structures in complex with deubiquitinating enzymes.

Table 9 The six conformational clusters and their weights of the weighted X-ray ensemble

Cluster	Final weight ± SD	Composition
Cluster 1	0.55 ± 0.03	1AAR-B, 1UBQ-A, 2C7M-B, 2C7N-H, 2QHO-A, 3EHV-C, 3M3J-A, 3M3J-E
Cluster 2	0.29 ± 0.03	2G45-B, 2G45-E, 2HD5-B
Cluster 3	0.064 ± 0.001	2DX5-B, 3KW5-B
Cluster 4	0.043 ± 0.002	1YD8-V
Cluster 5	0.027 ± 0.004	3HIU-A
Cluster 6	0.026 + 0.001	1TBE-A

The conformations included in each cluster are listed by their PDB ids as well as chain identifiers

The remaining four clusters contain the following structures. Cluster 3 consists of 2DX5 and 3KW5. 2DX5 is a structure of ubiquitin in complex with mouse EAP45-GLUE domain. 3KW5 contains a structure of ubiquitin in complex with ubiquitin carboxy terminal hydrolase L1. Cluster 4 contains 1YD8, a structure of ubiquitin in complex with human GGA3 GAT domain. Cluster 5 contains

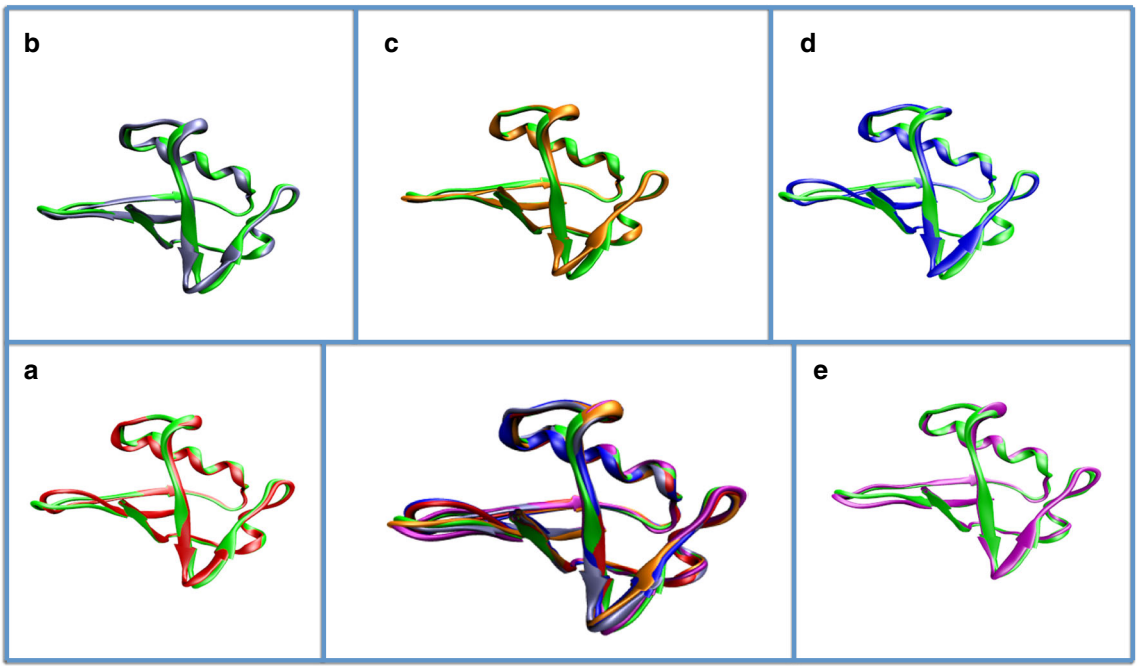

Fig. 2 The final weighted X-ray ensemble that consists of six clusters (see Table 9) and representative conformations for each cluster. *Center*—All the structures overlaid onto one another, 1UBQ-A (cluster 1)—*green*, 2G45-E (cluster 2)—*red*, 2DX5-B (cluster 3)—*ice blue*, 1YD8-V (cluster 4)—*purple*, 3HIU-A (cluster 5)—*orange* and 1TBE-A (cluster 6)—*blue*. **a–e** compare 1UBQ-A with 2G45-E (*red*), 2DX5-B (*ice blue*), 3HIU-A (*orange*), 1TBE-A (*blue*), and 1YD8-V (*purple*) respectively

3H1U, a structure of ubiquitin in complex with cadmium ion. Lastly, cluster 6 contains 1TBE, a structure of ubiquitin in the form of tetraubiquitin. These four clusters all together have a relative population of about 15 %. Figure 2 shows the final structure ensemble (center) as well as individually, a representative conformation from each cluster (panels a–e).

Cross validation

The individual Q-factors obtained for the different bond vectors are shown in Table 10 for the weighted X-ray ensemble along with other recently derived ensembles. By partitioning the ensemble into six sub-ensembles (represented by the clusters) and assigning them relative populations, the Q-factors of all the individual bond vectors are significantly lowered. Remarkably, the cross validation Q-factor, that of CAHA, is also lowered from 0.161 to 0.145 for the weighted X-ray ensemble. This significant improvement in Q-factors further confirms the validity of clustering and relative population assignments discussed above.

In contrast to those of the single structure representation, residue-wise Q-factors of unweighted and weighted ensembles are shown in Fig. 3. It is seen that for most of the residues, the unweighted ensemble has lower Q-factors than the single structure, 1UBQ, while the weighted ensemble further lowers the Q-factors.

Table 10 Q-factors of the different bond vectors of the weighted X-ray ensemble as well as some other ensembles

NH	CaC	CaHa	CN	CHN	Description
0.122	0.097	0.145	0.088	0.186	Weighted X-ray
0.184	0.108	0.161	0.099	0.228	Unweighted X-ray
0.071	0.118	0.069	0.138	0.188	EROS
0.213	0.118	0.128	0.138	0.234	EROS reprotonated
0.066	0.140	0.167	0.096	0.182	ERNST
0.180	0.141	0.177	0.095	0.207	ERNST reprotonated
0.244	0.180	0.236	0.171	0.266	1UBQ
0.114	0.105	0.084	0.120	0.163	1D3Z
0.231	0.175	0.196	0.233	0.281	MUMO

CaHa is used for cross-validation

The Q-factor results of the weighted X-ray ensemble are on the par even with 1D3Z, NMR ensemble that was determined using RDC as one of the restraints and are noticeably better than MUMO (Richter et al. 2007), a ubiquitin ensemble computationally determined using NOE and order parameters as constraints. When compared with EROS and ERNST (Fenwick et al. 2011), the weighted X-ray ensemble falls short especially in the NH and CAHA datasets. However, as was pointed out in (Markwick et al. 2009), the conformations in EROS ensemble may have incorrect geometry. Indeed both reprotonated EROS and

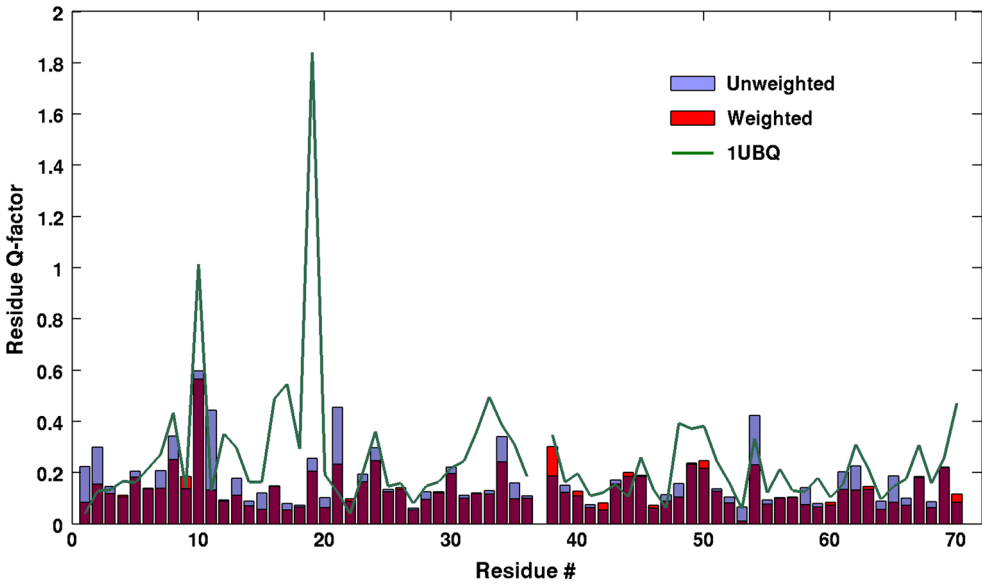

Fig. 3 Residue-wise Q-factors of 1UBQ (the *green line*), the unweighted (*blue bars*) and weighted (*red bars*) X-ray ensemble. The common region between the unweighted and weighted is *colored maroon*

reprotonated ERNST display much higher Q-factor values, see Table 10.

Uncertainty in weight assignments

Uncertainty in weight assignments can be computed when there are multiple datasets (see "Materials and methods"). In the case of ubiquitin, there are 24 NH RDC data sets along with two multi-vector RDC datasets. The whole datasets are partitioned into two subsets such that each subset contains one multi-vector dataset along with an equal proportion of NH RDC datasets. Weights obtained from each subset are compared and their SDs are used for representing the uncertainties in weight assignments (see Table 9).

Effects of weighting on conformational features of ensembles

In addition to improving the reproduction of experimental data, weighting alters conformational properties of the ensemble. One of the interesting features of ubiquitin structure is the presence of a "switched" conformation, which is hypothesized to have a biological function (Huang et al. 2011). The dihedral angles ϕ of residue 53 and ψ of residue 52 play an important role in facilitating the switch. While ϕ_{53} and ψ_{52} of the "switched" conformation are in the range of $\sim 100°$ and $\sim 130°$ respectively, the same two dihedrals are in the range of $\sim -90°$ and $\sim -50°$ respectively for the unswitched conformation such as in 1UBQ. We look into the changes in the population distributions of these dihedral angles before and after reweighting and the results are presented in Fig. 4.

In the first row of Fig. 4 (panels a and b) are shown the differences in the population distributions of dihedral angles ϕ_{53} and ψ_{52} between before and after reweighting the 143-conformation X-ray ensemble. The reweighting significantly alters the dihedral angle distributions of ϕ_{53} and ψ_{52}, shifting much of the populations from being around the unswitched conformation to the switched conformation. The reweighting also reduces the overall ranges of the dihedral angle distributions and makes the two population peaks narrower and sharper. To further demonstrate how strong an effect weighting can have on dihedral angle distributions, all the switched conformations except 2G45 (chain E, a "switched" conformation) are removed from the 143 conformations. The population of the switched conformation in this reduced ensemble before weighting is now less than 1 %. The second row of panels (c and d) of Fig. 4 show the difference in population distributions upon reweighting this ensemble. As is seen, after reweighting the population of the "switched" conformation increases dramatically from less than 1 % to as high as 20 %.

Application to a computationally-determined ensemble

In the recent years many ubiquitin ensembles have been determined computationally. ERNST, standing for ensemble refinement for native proteins using a single alignment tensor, was refined using NOEs and RDCs (Fenwick et al. 2011). ERNST does a very good reproduction of the experimental RDCs as seen from the low Q-factors in Table 11. But as with EROS, there is a significant increase in Q-factors once the ensemble is reprotonated using standard tools. Though the validity of reprotonation is debatable, such a significant increase in Q-factors could be due to the covalently incorrect placement of hydrogen atoms (Markwick et al. 2009). Therefore, we choose to

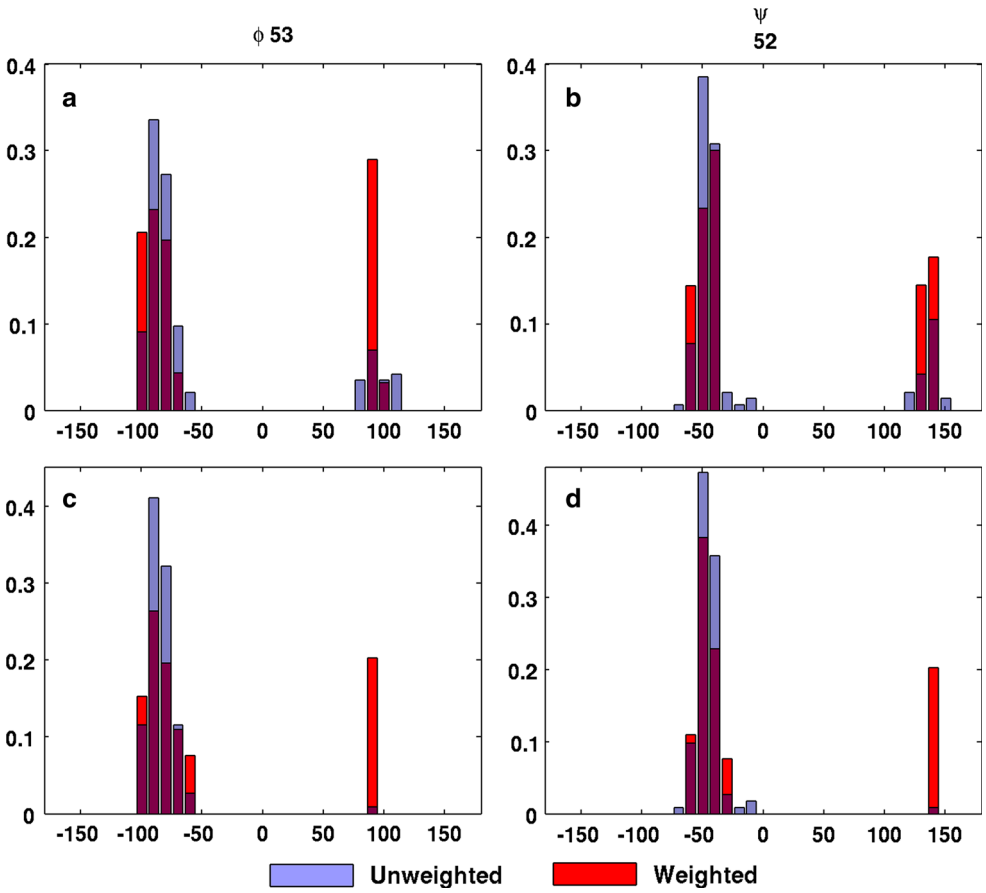

Fig. 4 Effects of weighting on the conformational features of X-ray ensembles. **a**, **b** Population distributions of ϕ_{53} and ψ_{52} dihedral angles before (*blue bars*) and after (*red bars*) weighting of the X-ray ensemble. **c**, **d** Same population distributions but for a modified X-ray ensemble whose "switched" conformations except one are all taken out (see the text). The common region between the unweighted and weighted is *colored maroon*

apply our weighting algorithm to the reprotonated ERNST ensemble instead to avoid introducing into weights errors due to incorrect covalent geometry. The Q-factors obtained after weighting the reprotonated ERNST ensemble are shown in Table 11. From the table it is seen that though weighting lowers the Q-factors, the decreases are mostly quite nominal and the new Q-factors are not as good as those of the weighted X-ray ensemble.

A close look at dihedral angle distributions of ϕ_{53} and ψ_{52} of the ENRST ensemble as we did to the X-ray ensemble in Fig. 4 reveals the reason. Figure 5 (panels a and b) shows that ENRST does not sample the "switched" conformational state at all and all the conformations have ϕ_{53} and ψ_{52} angles similar to 1UBQ. To assess the importance of "switched" conformational state, we add 2G45-E (a representative switched conformation) to ERNST ensemble and then reweight it. Interestingly, the Q-factors now improve significantly (see Table 11) and reach to a level similar to the weighted X-ray ensemble. Moreover, the switched conformation (2G45-E) is assigned to a relative population of 0.30, which is highly similar to the weight of the "switched" conformation in the weighted X-ray ensemble (which is 0.29). This is remarkable since it shows that common conformational features emerge after reweighting even though the two ensembles to which the reweighting scheme has been applied are rather different. In Fig. 5, panels a and b show the population distributions of the ϕ_{53} and ψ_{52} before and after weighting of the reprotonated ERNST ensemble, while panels c and d show the distributions of the same dihedral angles of the same reprotonated ERNST ensemble after a switched conformation is added to it. By comparing between the dihedral angle distributions in Figs. 4 and 5, it is seen that while ERNST (reprotonated) itself does not have similar properties as the weighted X-ray ensemble, the weighted ERNST + 2G45-E

Table 11 Q-factors of the different bond vectors of the ERNST ensembles

NH	CaC	CaHa	CN	CHN	Description
0.066	0.140	0.167	0.096	0.182	ERNST
0.180	0.141	0.177	0.095	0.207	ERNST reprotonated
0.147	0.145	0.178	0.098	0.190	Weighted reprotonated ERNST
0.123	0.112	0.132	0.093	0.172	Weighted (reprotonated ERNST + 2G45-E)

CaHa is used for cross-validation. ERNST reprotonated is the same as ERNST except the hydrogen atoms are replaced using standard geometry. In the last row, the reprotonated ERNST is first enhanced with a switched conformation 2G45-E before the population re-weighting is applied

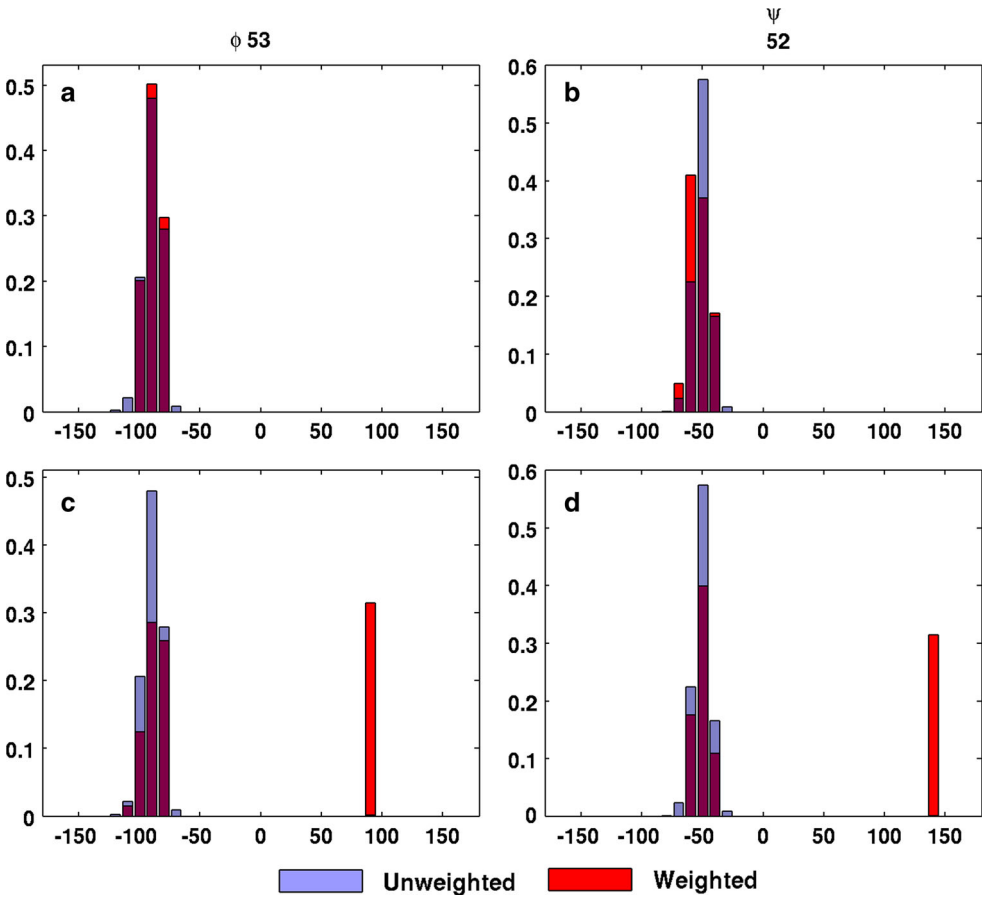

Fig. 5 Dihedral distributions of ϕ_{53} and ψ_{52} in the ERNST ensembles. **a**, **b** Population distributions of ϕ_{53} and ψ_{52} dihedral angles before (*blue bars*) and after (*red bars*) weighting of the ERNST ensemble. **c**, **d** Populations of the same dihedral angles before (*blue bars*) and after (*red bars*) weighting of an enhanced ERNST ensemble (with 2G45-E, a "switched" conformation, added). The common region between the unweighted and weighted is *colored maroon*

(see panels c and d of Fig. 5) shows highly similar conformational properties to the weighted X-ray ensemble (see panels a and b of Fig. 4), especially with respect to the dihedral angle distributions of ϕ_{53} and ψ_{52}.

Discussion and conclusions

Proteins are dynamic molecules and even the native state of a protein is not a single static structure but spread over a broader region of the conformation space. As a result, for many proteins, an ensemble of conformations provides a better depiction of the native states.

In this work we present a method to improve ensembles and their ability to depict the native states. The method works by identifying conformation states within an ensemble and assigning appropriate relative populations, or weights, to them. Each of these conformation states is represented by a sub-ensemble formed by a subset of the conformations.

Our results demonstrate that such weight assignment is feasible and the weights are significant. Since the weights are computed by least squares fitting to the experimental RDC data, one may naturally question the significance of the weights. Are the weights significant and physically meaningful? Or are they merely a result of over-fitting to the noise in the experimental data? To address this concern, we design a sensitive measure to recognize the onset of over-fitting and finish the weight assignment before over-fitting starts to occur. Lastly, the significance of the weights is further examined and verified by cross validation.

The method presented in this work uses experimental RDC data as constraints to assign relative populations to conformations within an ensemble. In order for this method to succeed, what is the requirement on the ensemble and its conformations? Our results indicate the following:

- Undersampling in conformation states, where some conformation states are represented by few conformations, does not hinder weight determination. Experimental structures of the same protein obtained under different conditions or bound states have been suggested to form a native state ensemble of the protein (Best et al. 2006). Such a native state ensemble may cover all the important conformation states of the protein, but not necessarily proportionally, and some of the states may be severely undersampled. As seen from case I, undersampling does not hinder weight assignment and our algorithm can be readily applied to determine the relative populations.

- Noise conformations in an ensemble that do not represent any conformation states can be mostly filtered out. In case II, we create a situation where the ensemble contains a cluster of conformations that do not belong to any conformation state. Case III represents another situation where each conformation state is mixed with a large amount of noise conformations. The presence of noise conformations may make it difficult to identify conformation states, or to separate conformations representing a conformation state from those that do not. However, test results show that our method is able to effectively filter out most of the noise conformations.
- While cases I to III show that given an ensemble with good coverage and completeness, the weighting algorithm is able to identify the clusters and assign them with proper weights and thus lower the Q-factors, case IV indicates the converse is not necessarily true: low Q-factors do not necessarily mean that an ensemble is of good quality. Therefore, cautions must be taken in future ensemble determination and assessment. Measures other than Q-factors are needed to check the quality of computer-generated ensembles. It is not clear what these measures are, but their discovery and identification are going to be critical to the field's progress.

We apply our method to a ubiquitin ensemble of 143 conformations and identify six conformation states. The two most populated conformation states, one of which represents the conformation state near the free state of ubiquitin while the other the "switched" conformer, match closely with conformation states identified by other studies. The relative populations assigned to these two states by our method, agree extremely well with the findings by Shaw's group through long MD simulations(Piana et al. 2013). The validity of such conformation state identification and weight assignments are further confirmed by significant improvement in Q-factors and cross-validation.

We apply our method also on a computationally derived ensemble, ERNST, which was refined against RDCs and NOEs. Even though the reproduction of experimental data, RDCs in this case, worsens after reprotonation, we are able to significantly improve the Q-factors by augmenting the ensemble with a switched conformation and reweighting. In doing so we observe the emergence of common dihedral angle distributions in both the augmented ERNST ensemble and X-ray ensemble.

The method presented in this work can be applied to other proteins to identify conformation states and assign relative populations, provided that sufficient RDC data exist. A good question to ask is how much RDC data is required for weight assignment? And what type of RDC data is required, NH RDCs, multi-vector RDCs, or both? We plan to study this in future work.

The number of conformation states recognized by our method can be used to guide the selection of ensemble size in ensemble determination. Most ensemble determination methods try out different sizes for replica ensembles, usually from 1, 2, 4, 8, up to 16. The method presented here provides an informed estimation of the right size for the ensemble. Since the method requires an ensemble as a starting point, it could be applied alternatively with an existing ensemble determination method until the process converges and a right ensemble size is identified. Our results strongly suggest that relative weights, instead of the default equal-weights, should be considered as parameters in ensemble determination.

Acknowledgments Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged. The authors would also like to thank the two anonymous reviewers for their insightful comments.

Appendix 1: Calculation of RDC's

Given a 3D structure of a protein, the RDC D_{ij} can be expressed using the molecular frame. First, the elements of Saupe matrix is defined as:

$$S_{\mathrm{lm}} = \left\langle \frac{3\cos\beta_l \cos\beta_m - k_{\mathrm{lm}}}{2} \right\rangle \tag{4}$$

where β_l denotes the orientation of the l-th molecular axis with respect to the external magnetic field. The RDC D_{ij} can be reformulated in the molecular frame as:

$$D_{ij} = \frac{-\mu h r_{\mathrm{i}} r_{\mathrm{j}}}{(2\pi r)^3} \left(\alpha_y^2 - \alpha_x^2; \alpha_z^2 - \alpha_x^2; 2\alpha_{x\alpha_y}; 2\alpha_{x\alpha_z}; 2\alpha_{y\alpha_z}\right) \begin{pmatrix} S_{\mathrm{yy}} \\ S_{\mathrm{zz}} \\ S_{\mathrm{xy}} \\ S_{\mathrm{xz}} \\ S_{\mathrm{yz}} \end{pmatrix} \tag{5}$$

where α_{x}, α_{y}, and α_{z} are the cosines of the angles between the bond vector of the two nuclei and the x, y, and z axes of the molecular frame. Let α_{xk}, α_{yk}, and α_{zk} represent the k-th α_{x}, α_{y}, and α_{z}. When all the bond vectors are considered, we have the following formula:

$$D_{\mathrm{exp}} = \left(\frac{-\mu h r_i r_j}{(2\pi r)^3}\right) \begin{pmatrix} \alpha_{\mathrm{y},1}^2 - \alpha_{\mathrm{x},1}^2 & \cdots & 2\alpha_{\mathrm{y},1}\alpha_{\mathrm{z},1} \\ \vdots & \ddots & \vdots \\ \alpha_{\mathrm{y},N}^2 - \alpha_{\mathrm{x,N}}^2 & \cdots & 2\alpha_{\mathrm{y,N}}\alpha_{\mathrm{z,N}} \end{pmatrix} \begin{pmatrix} S_{\mathrm{yy}} \\ S_{\mathrm{zz}} \\ S_{\mathrm{xy}} \\ S_{\mathrm{xz}} \\ S_{\mathrm{yz}} \end{pmatrix} \tag{6}$$

where D_{exp} is the experimental RDCs and N is the total number of data points. Eq. 6 can be rewritten in the following matrix form:

$$D_{\text{exp}} = cAS \tag{7}$$

where c is the constant $\frac{-\mu h r_i r_j}{(2\pi r)^3}$ and A is the N × 5 matrix in Eq. 6 and S is the 5 × 1 vector. Optimal S and thereby D_{calc} (i.e., the calculated RDCs) can be computed by singular value decomposition using Moore–Penrose pseudoinverse of matrix A:

$$S = A^{-1}D_{\text{exp}} \tag{8}$$

$$D_{\text{calc}} = AA^{-1}D_{\text{exp}} \tag{9}$$

Residual dipolar coupling (RDC) calculation from an ensemble

The RDC calculation method for a single structure can be extended to take ensemble averaging into account so that the ensemble D_{calc} can be obtained. First let us consider the assumption that all structures have equal contributions toward the experimental RDC: D_{exp}. When an ensemble with equal weights is considered, we have the following formula:

$$\left(\frac{A_1}{n} + \frac{A_2}{n} + \cdots + \frac{A_k}{n} + \cdots \frac{A_n}{n}\right) S = D_{\text{exp}} \tag{10}$$

where A_{k} is the A matrix obtained from the k-th structure in the ensemble. S can be obtained from the following equation:

$$S = \left(\frac{A_1}{n} + \frac{A_2}{n} + \cdots + \frac{A_k}{n} + \cdots \frac{A_n}{n}\right)^{-1} D_{\text{exp}} \tag{11}$$

Strictly speaking, the Saupe matrix might vary for different conformations of the protein. In this work we assume the same Saupe matrix for all the conformations. This assumption is reasonable especially for proteins that make only small conformation changes, as is the case with ubiquitin.

Now let us consider the case that structures in an ensemble have different populations and thus different amounts of contributions toward the experimental observations D_{exp}. Therefore, weights (representing the relative populations) are given to different structures and the following formula is used to represent the combination:

$$(w_1A_1 + w_2A_2 + \cdots + w_kA_k + \cdots w_nA_n)S = D_{\text{exp}} \tag{12}$$

where n is the total number of structures and w_k and A_{k} are respectively the relative population (or weight) and A matrix of the k-th structure. Thus, S can be obtained from the following formula:

$$S = (w_1A_1 + w_2A_2 + \cdots + w_kA_k + \cdots w_nA_n)^{-1} D_{\text{exp}} \tag{13}$$

Our problem is thus to find the optimal relative populations for the structures in the ensemble so that the experimental RDCs are best reproduced. The solution to this problem is given in Appendix 2.

Appendix 2

The iterative least squares fitting algorithm to a single RDC data set

Iterative Least Squares Fitting ([A_1 A_2 … A_n], D_{exp})
for i = 1 to n do
    new_weights(i) = 1/n
end for
repeat
    old weights = new_weights
    A = old_weights(1)*A_1 + ... + old_weights(n)*A_n
    S = pseudo_inverse(A) * D_{exp}
    AS = [A_1S A_2S .... A_nS]
    new_weights = non_negative_least_squares(AS, D_{exp})
Until old_weights and new_weights converge.
return new_weights

The iterative least squares fitting algorithm to multiple RDC data sets

Iterative Least Squares Fitting Multiple RDCs ([A_1 A_2 … A_n], [D_1, D_2 … D_m])
for i = 1 to n do
    new_weights(i) = 1/n
end for
repeat
    old weights = new_weights
    A = old_weights(1)*A_1 + ... + old_weights(n)*A_n
    for i = 1 to m do
        S(i) = pseudo_inverse(A) * D_i
        AS(i) = [A_1S(i) A_2S(i) .... A_nS(i)]
    end for
    AS_all = $\begin{pmatrix} AS(1) \\ AS(2) \\ \vdots \\ AS(m) \end{pmatrix}$

    D_all = $\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_m \end{pmatrix}$

    new_weights = non_negative_least_squares(AS_all, D_all)
Until old_weights and new_weights converge.
return new_weights

References

Austin RH, Beeson KW, Eisenstein L, Frauenfelder H, Gunsalus IC (1975) Dynamics of ligand binding to myoglobin. Biochemistry 14:5355–5373

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

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

Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 5:789–796

Bonvin AMJJ, Brunger AT (1996) Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? J Biomol NMR 7:72–76

Clore GM, Schwieters CD (2004a) Amplitudes of protein backbone dynamics and correlated motions in a small α/β protein: correspondence of dipolar coupling and heteronuclear relaxation measurements. Biochemistry 43:10678–10691

Clore GM, Schwieters CD (2004b) How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? J Am Chem Soc 126:2923–2938

Clore GM, Schwieters CD (2006) Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small α/β protein: a unified picture of high probability, fast atomic motions in proteins. J Mol Biol 355: 879–886

Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE (1999) Peptide folding: when simulation meets experiment. Angew Chem Int Ed 38:236–240

de Groot BL, van Aalten DM, Scheek RM, Amadei A, Vriend G, Berendsen HJ (1997) Prediction of protein conformational freedom from distance constraints. Proteins 29:240–251

DePristo MA, de Bakker PI, Blundell TL (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 12:831–838

Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4:10–19

Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG (2001) Evaluating protein structure-prediction schemes using energy landscape theory. IBM J Res Dev 45:475–497

Fenwick RB, Esteban-Martín S, Richter B, Lee D, Walter KF, Milovanovic D, Becker S, Lakomek NA, Griesinger C, Salvatella X (2011) Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. J Am Chem Soc 133:10336–10339

Frauenfelder H, Silgar S, Wolynes P (1991) The energy landscapes and motions of proteins. Science 254:1598–1603

Frauenfelder H, McMahon BH, Austin RH, Chu K, Groves JT (2001) The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. Proc Natl Acad Sci USA 98:2370–2374

Furnham N, Blundell TL, DePristo MA, Terwilliger TC (2006) Is one solution good enough? Nat Struct Mol Biol 13:184–185

Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120:11919–11929

Huang KY, Amodeo GA, Tong L, McDermott A (2011) The structure of human ubiquitin in 2-methyl-2,4-pentanediol: a new conformational switch. Protein Sci 20:630–639

Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Mol Biol 9:646–652

Kontaxis G, Bax A (2001) Multiplet component separation for measurement of methyl 13C-1H dipolar couplings in weakly aligned proteins. J Biomol NMR 20:77–82

Lakomek NA, Carlomagno T, Becker S, Griesinger C, Meiler J (2006) A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. J Biomol NMR 34:101–115

Lakomek NA, Walter KF, Fares C, Lange OF, de Groot BL, Grubmuller H, Bruschweiler R, Munk A, Becker S, Meiler J et al (2008) Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. J Biomol NMR 41:139–155

Lange OF, Lakomek NA, Fares C, Schroder GF, Walter KF, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320: 1471–1475

Lawson CL, Hanson RJ (1995) Solving least squares problems. SIAM, Philadelphia

Levin EJ, Kondrashov DA, Wesenberg GE, Phillips GN Jr (2007) Ensemble refinement of protein crystal structures: validation and application. Structure 15:1040–1052

Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433:128–132

Markwick PRL, Bouvignies G, Salmon L, McCammon JA, Nilges M, Blackledge M (2009) Toward a unified representation of protein structural dynamics in solution. J Am Chem Soc 131:16968–16975

Miyashita O, Onuchic JN, Wolynes PG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc Natl Acad Sci 100:12570–12575

Ottiger M, Bax A (1998) Determination of relative N-HN, N-C′, Cα-C′, and Cα-Hα effective bond lengths in a protein by NMR in a dilute liquid crystalline phase. J Am Chem Soc 120(47): 12334–12341

Phillips GN (2009) Describing protein conformational ensembles: beyond static snapshots. F1000 biology reports, vol 1

Piana S, Lindorff-Larsen K, Shaw DE (2013) Atomic-level description of ubiquitin folding. Proc Natl Acad Sci 110:5915–5920

Prestegard J (1998) New techniques in structural NMR—anisotropic interactions. Nat Struct Mol Biol 5:517–522

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

Shao J, Tanner SW, Thompson N, Cheatham TE (2007) Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. J Chem Theory Comput 3:2312–2334

Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. Proc Natl Acad Sci 92:9279–9283

Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285: 1735–1747