J Biomol NMR (2015) 63:341–351
DOI 10.1007/s10858-015-9993-9

ARTICLE

Ensembles of a small number of conformations with relative populations

Vijay Vammi[1] · **Guang Song**[1,2]

Received: 6 February 2015 / Accepted: 14 October 2015 / Published online: 17 October 2015

Abstract In our previous work, we proposed a new way to represent protein native states, using ensembles of a small number of conformations with relative Populations, or ESP in short. Using Ubiquitin as an example, we showed that using a small number of conformations could greatly reduce the potential of overfitting and assigning relative populations to protein ensembles could significantly improve their quality. To demonstrate that ESP indeed is an excellent alternative to represent protein native states, in this work we compare the quality of two ESP ensembles of Ubiquitin with several well-known regular ensembles or average structure representations. Extensive amount of significant experimental data are employed to achieve a thorough assessment. Our results demonstrate that ESP ensembles, though much smaller in size comparing to regular ensembles, perform equally or even better sometimes in all four different types of experimental data used in the assessment, namely, the residual dipolar couplings, residual chemical shift anisotropy, hydrogen exchange rates, and solution scattering profiles. This work further underlines the significance of having relative populations in describing the native states.

Keywords Ubiquitin · NMR · Residual dipolar couplings · Residual chemical shift anisotropy · Hydrogen exchange rates · SAXS · WAXS · Overfitting

Introduction

Proteins are dynamic molecules and often occupy multiple conformational states in their native states. The functional behavior of a protein is thus best understood from the distribution and dynamic transition among these conformational states that form the native state ensemble (Austin et al. 1975; Boehr et al. 2009; DePristo et al. 2004; Frauenfelder et al. 1991, 2001).

Nuclear Magnetic Resonance (NMR) experiments have played a pivotal role in capturing the dynamics of proteins in their native states. Data obtained from NMR experiments have been used as restraints in recovering the underlying structures or ensembles. In that process, two different refinement schemes are routinely followed:

1. *Average structure representation* In this scheme, a single structure is used to explain all the observed experimental data. For Ubiquitin, one of the most studied proteins, a single structure has been shown to be sufficient in reproducing most experimental data (Cornilescu et al. 1998; Maltsev et al. 2014). But it was also pointed out that average structure representations, due to the lack of structural variance, cannot fully capture the underlying dynamics (Clore and Schwieters 2004a, b). This representation becomes less complete when the studied protein occupies multiple

Electronic supplementary material The online version of this article (doi:10.1007/s10858-015-9993-9) contains supplementary material, which is available to authorized users.

✉ Vijay Vammi
vsvammi@iastate.edu

Guang Song
gsong@iastate.edu

[1] Bioinformatics and Computational Biology Program, Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011, USA

[2] Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

distinct sub-states, since the refinement protocol would be over-restrained (under-fitting; Richter et al. 2007).

2. *Ensemble representation* In this representation, an ensemble of conformations is used to explain the experimental data. In the case of Ubiquitin, there has been a number of recent work aimed at determining an ensemble of conformations for the protein, such as MUMO (Richter et al. 2007), EROS (Lange et al. 2008) and ERNST (Fenwick et al. 2011). All of these ensembles are shown to represent the dynamics well but there is little confidence that any given conformation within the ensemble truly belongs to the native state ensemble, since the ensemble might be under-constrained or over-fitted (Phillips 2009; Ángyán and Gáspári 2013).

In this work, we propose a third representation,

3. *Ensembles of a Small number of conformations with relative Populations or ESP in short* In our recent work (Vammi et al. 2014), we showed that the conformation space could be represented by far fewer conformations than the aforementioned ensemble representations and the conformations could be clustered into conformation states and these conformation states could be assigned relative populations, corresponding to their Boltzmann weights. The advantage of using ESP over an average structure is that it overcomes underfitting. The advantage of using ESP over using an ensemble with hundreds of conformations is that it minimizes overfitting. ESP uses a much smaller number of conformations than regular ensembles.

The objective of this work is to establish ESP as a better ensemble representation for describing the native states of a protein. To demonstrate that ESP ensembles are indeed of high quality and minimize overfitting, we resort to a series to significant experimental data that are not used in the determination of these ensembles, and show that ESP ensembles, though having a much smaller number of conformations, are able to reproduce these experimental data equally well or even better sometimes and with less overfitting. Weighted ensembles had been successfully used in modeling unfolded protein conformational ensembles (Choy and Forman-Kay 2001; Fisher et al. 2010) and was considered also in loop modeling (Tripathy et al. 2012), but they are usually not used in determining native state protein ensembles.

Though cross-validation using a subset of the data points that were left out during the ensemble determination stage has been commonly used, unused experimental data of different types present an even better resource for assessing the quality of the ensembles since they are even more unbiased. Since all of the aforementioned ensembles, namely, MUMO, EROS, and ERNST, use NOEs or RDCs as restraints in their construction, experimental data on Residual Chemical Shift Anisotropies (RCSA), amide exchange reactivities, and solution scattering profiles are employed in this study for cross-validation.

Our ensemble representation with relative populations could be thought of as an intermediate scheme between the two refinement schemes aforementioned: *average structure representation* or *ensemble representation*. Both representations have strengths and weaknesses. Average structure representation is the simplest in form but lacks structure variance, while ensemble representation captures the dynamics of the conformation space well but may suffer the problem of over-fitting and there is little confidence that any given conformation within the ensemble truly belongs to the native state ensemble. The advantage of ESP representation is that it has a very limited number of conformation states whose relative populations are rigorously determined (Vammi et al. 2014) without over-fitting. Consequently, there is high confidence on the validity of these conformation states.

Materials and methods

Ensembles of a small number of conformations with relative populations (ESP)

Two ESP ensembles were reported in our previous work (Vammi et al. 2014) and will be used in this work as example ESPs.

(a) *Weighted X-ray ensemble* X-ray conformations resolved in different conditions have been shown to form a native state ensemble (Best et al. 2006). In our previous work (Vammi et al. 2014), 143 such structures of Ubiquitin were collected from PDB (Berman et al. 2000) to form an unweighted X-ray ensemble. After applying the weighting protocol, 16 of these structures were selected to form the weighted X-ray ensemble and six conformational states were identified (Vammi et al. 2014). The weights assigned to the conformational states are in agreement with what was found in the 1 μs equilibrium simulation conducted by Shaw's group (Piana et al. 2013). The conformational state adopted by Ubiquitin when bound to de-ubiquitinating proteins, also called the "switched" conformation (Huang et al. 2011; Sidhu et al. 2011), was given a weight of $\sim$0.30.

(b) *Enhanced ERNST ensemble* Besides the X-ray ensemble, our conformation weighting algorithm was applied to another computationally derived

ensemble, ERNST (Fenwick et al. 2011) to produce an enhanced ERNST ensemble. After introducing a "switched" conformation to the ensemble and then assigning relative populations to the conformations in the ensemble, it was found that the enhanced ERNST ensemble was able to reproduce experimental data in a comparable accuracy to the weighed X-ray ensemble. This enhanced ERNST ensemble contains one X-ray switched conformation and 35 conformations selected from the original ENRST ensemble that has 640 conformations. The composition of the ensemble along with their weights is given in the supplementary information, Table S1.

In this work, these two ESP ensembles are compared with three regular ensembles determined for Ubiquitin: MUMO (Richter et al. 2007; pdb-id: 2NR2), EROS (Lange et al. 2008; pdb-id: 2K39), and ERNST (Fenwick et al. 2011; pdb-id: 2KOX), as well as two NMR structures with pdb-ids 1D3Z (Cornilescu et al. 1998) and 2MJB (Maltsev et al. 2014) and one crystal structure 1UBQ (Vijay-Kumar et al. 1987).

Residual dipolar couplings (RDC)

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

$$D_{\{AB\}} = \sum_{i=x,y,z} -\frac{\mu h \gamma_A \gamma_B}{(2\pi r)^3} \cos^2 \emptyset_i A_{ii} \quad (1)$$

where γ_A and γ_B are the nuclear magnetogyric ratios of nuclei A and B respectively, h is Plank's constant, μ is permittivity of space, r is the internuclear distance between the two nuclei, A_{ii} the principal moment of the alignment tensor and $\varnothing_i$ is the angle between the internuclear vector and ith principal axis of the alignment tensor. The alignment tensor could be determined by fitting a single structure or ensemble to the experimental data. 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. Details regarding back-calculation of RDC's were given in the appendix of our previous work (Vammi et al. 2014).

Experimental RDCs used in this work

The RDCs used to determine the weights for the X-ray ensemble and enhanced ERNST ensemble are given in Vammi et al. (2014), along with the codes assigned to them according to Lakomek et al. (2008). The Q-factors reported in this work use the newly determined RDC dataset in Squalamine and Pf1 media (Maltsev et al. 2014).

Q-factor

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

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

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

Residual chemical shift anisotropy (RCSA)

Along with RDC's, chemical shifts also change upon shifting from an isotropic medium to an anisotropic medium (Cornilescu and Bax 2000; Cornilescu et al. 1998; Liu and Prestegard 2010; Saitô et al. 2010). The change is defined by:

$$\Delta\delta = \sum_{i=x,y,z} \sum_{j=x,y,z} A_{jj} \cos^2 \theta_{ij} \delta_{ii} \quad (3)$$

where δ_{ii} is the principal moment of the chemical shift tensor, A_{jj} the principal moment of the alignment tensor and θ_{ij} is the angle between ith principal axis of the chemical shift tensor and jth principal axis of the alignment tensor. The alignment tensor used in RCSA back-calculations is generally the same as the one computed from RDCs using either a single conformation or an ensemble (Vammi et al. 2014). More information regarding the relation between RDC and RCSA back-calculation of a conformation can be found in Liu and Prestegard (2010).

The experimental dataset of RCSA used in this work were reported in Cornilescu et al. (1998) along with the RDC dataset used for obtaining the alignment tensor. Magnitudes and orientations of the chemical shift tensors reported in Cornilescu and Bax (2000) are used in this work.

Amide hydrogen reactivity

Hydroxide catalyzed amide hydrogen rates were used as a measure to assess conformational distribution of various ensembles (Hernández et al. 2010; LeMaster et al. 2009). The experimental rate constants of amide hydrogen exchange depend not only on the solvent accessibility but also on the chemical environment surrounding the amide hydrogen. Even rarely exposed amide hydrogen could

therefore exhibit a high exchange rate if the chemical environment is conducive for such an exchange. This property makes amide hydrogen reactivity a very sensitive measure of the conformational distribution of the native states.

Poisson Boltzmann electrostatic calculations

The experimental exchange rate constants for all the backbone amide hydrogens of Ubiquitin were reported in the work by LeMaster et al. (2009). In this work, electrostatics calculations needed to predict the exchange rates of conformational ensembles are performed in a similar way to what was described in a previous work (Hernández et al. 2010). Briefly, surface exposure of amide hydrogens in all the conformations belonging to the ensemble is computed using Naccess (Hubbard and Thornton 1993), using default values for the atomic radii and 1 Å for the radius of the probe sphere. For all the amide hydrogens that are not involved in any hydrogen bonding [computed using HBplus (McDonald and Thornton 1994)] and have a surface exposure greater than 0.5 $Å^2$, Poisson-Boltzmann continuum electrostatic computations are done using Delphi (Li et al. 2012). The CHARMM22 atomic charge and radius values (MacKerell et al. 1998) are used in the electrostatic computations. To make the comparisons feasible between different conformations of the ensemble, N-methylacetamide is added to the grid in such a way that the molecule is at least 16 Å away from any atom of the protein. The charge distribution of N-methylacetamide (or its anionic form) is taken from (LeMaster et al. 2009). Serines or threonines are mutated to alanine or α-aminobutyrate respectively before the electrostatic potential is computed.

Gauche side chain χ_1 conformers have remarkably low solvent exposure than their trans counterparts. To account for this, for every conformation, in addition to computing electrostatic potential in the original side chain configuration, a gauche χ_1 rotated side chain configuration also is used (whenever such a rotation was possible; LeMaster et al. 2009). The side chain position with the higher exchange rate is used for further processing.

Solution scattering profile

Small Angle X-ray scattering (SAXS) and wide angle X-ray scattering (WAXS) data encode the information about the shape and size of the bio-molecules in solution (Putnam et al. 2007; Svergun and Koch 2003). The observed intensities from X-ray scattering are sensitive to the overall conformational distribution of the protein and are being regularly used as complementary data to those obtained from NMR or X-ray crystallographic studies (Grishaev et al. 2005; Schwieters and Clore 2007). Though predicting the scattering profiles from either single structure or an ensemble were routinely done using the Crysol software package (Svergun et al. 1995), in this work we use the AXES (Analysis of X-ray scattering data for Ensemble of structures) software (Grishaev et al. 2010). In addition to providing significantly improved predictions, AXES webserver provides an easy way to predict such intensities from ensembles. The predicted intensities of all the ensembles or single structures reported in this paper are computed using a local version of AXES webserver, generously provided by Bax's group. The experimental SAXS/WAXS data used in this work are reported in (Grishaev et al. 2010). The agreement between the predicted and experimental scattering intensities is most commonly denoted by the χ value that is defined as:

$$\chi = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{I_{\exp}(q_i) - I_{calc}(q_i)}{\sigma(q_i)}\right)^2} \quad (4)$$

where I_{exp} and I_{calc} are the experimental and predicted scattering intensities at q_i with a error of σ_i and M is the number of observed scattering intensities.

Results and discussion

Agreement with experimental RDCs

Table 1 lists the Q-factors obtained for different bond vector types using different representations of Ubiquitin. The RDC datasets used for computing these Q-factors consist solely of the newly obtained RDC datasets in Squalamine and Pf1 media (Maltsev et al. 2014). All the representations, except 2MJB, were determined without using these newly determined datasets. This allows the Q-factors reported in Table 1 to serve as a strong cross-validation. It is worth pointing out that the two ESP ensembles, the weighted X-ray ensemble and the enhanced ERNST ensemble, are able to well reproduce the new RDC datasets (in Squalamine and Pf1 media) even though the dataset was not used in determining these two ensembles (Vammi et al. 2014).

The RDC Q-factors obtained for bonds with hydrogen atoms (NH, CaHa, CHN) are highly sensitive to the positions of the hydrogen atoms. Allowing a certain degree of deviation from the ideal covalent geometry can lower the Q-factors significantly. It should be noted that no such optimization of hydrogen atom positions was applied to our weighted X-ray or enhanced ERNST ensemble, while it was to the other representations, whose refinement protocols allowed such deviations from the ideal covalent geometry to better fit experimental RDC data. Nevertheless, the two ESP

Table 1 Q-factors obtained for different bond types by different representations of Ubiquitin

NH	CaC	CaHa	CN	CHN	Description
0.15	0.12	0.17	0.33	0.16	Weighted X-ray
0.18	0.12	0.18	0.34	0.16	Unweighted X-ray
0.12	0.14	0.20	0.31	0.14	ERNST (Fenwick et al. 2011)
0.15	0.15	0.17	0.32	0.16	Enhanced ERNST
0.11	0.13	0.17	0.32	0.21	EROS (Lange et al. 2008)
0.28	0.18	0.21	0.46	0.29	MUMO (Richter et al. 2007)
0.21	0.22	0.23	0.38	0.22	1UBQ (Vijay-Kumar et al. 1987)
0.16 (0.16)	0.13 (0.12)	0.17 (0.16)	0.32 (0.32)	0.19 (0.18)	1D3Z (Cornilescu et al. 1998)
0.08 (0.08)	0.1 (0.1)	0.10 (0.10)	0.30 (0.30)	0.14 (0.14)	2MJB (Maltsev et al. 2014)

The experimental RDCs used for computing these Q-factors consist of the newly obtained Squalamine and pf1 dataset (Maltsev et al. 2014). The Q-factors obtained by using only the first model of 1D3Z and 2MJB are shown in the parenthesis in the respective rows

ensembles have a comparable performance in RDC Q-factors to the other ensembles or average structures. Structure 2MJB gives the best RDC Q-factors, which is not surprising since it utilizes all the RDC data in its refinement process.

ESP ensembles give better agreements with residual chemical shift anisotropies (RCSA)

Table 2 compares the RMSDs between experimental and computed residual chemical shift anisotropies (RCSAs) for carbonyl carbons, nitrogens, and amide hydrogens, using different Ubiquitin ensembles. Since chemical shift anisotropies were not used in determining any of the above structures or ensembles, they can serve as an unbiased dataset for assessing the accuracy of different structures or ensembles. From the table it is seen that weighted X-ray (an ESP ensemble) outperforms its unweighted counterpart in predicting RCSAs: the RMS values of all three atom types are significantly reduced (see Table 2, row 1 and 2). Except for a nominal increase in RMSD for amide hydrogens, enhanced ERNST (another ESP ensemble) also performs better than ERNST itself.

Table 2 RMSDs of residual chemical shift anisotropy (RCSA) as predicted by different representations of Ubiquitin

Carbonyl	Nitrogen	Amide H	Q_{NH}	Description
6.37	16.2	1.57	0.11	Weighted X-ray
6.87	17.3	1.61	0.17	Unweighted X-ray
10.7	16.0	1.53	0.06	ERNST
7.84	15.7	1.61	0.11	Enhanced ERNST
8.63	16.6	1.51	0.07	EROS
13.2	19.63	1.67	0.22	MUMO
13.1	18.6	1.68	0.18	1UBQ
8.59 (8.3)	14.17 (14.04)	1.47 (1.48)	0.10	1D3Z
7.71 (8.43)	15.39 (15.59)	1.50 (1.50)	0.07	2MJB

None of the adjustable parameters in the RCSA was modified while predicting the chemical shifts. Q_{NH} is the RDC Q-factor of the NH dataset that was used in obtaining the alignment tensor. The same alignment tensor was used in the RCSA computations. The RMSDs obtained by using only the first model of 1D3Z and 2MJB are shown in parenthesis in the respective rows

Similar to the sensitivity to hydrogen atom positions in RDC calculations, calculations of the chemical shift tensors of nitrogens and amide hydrogens, and thus their RCSA predictions, depend on the orientations of the amide bond vectors. Comparisons of RCSAs regarding these two atom types should thus be done cautiously and with this in mind. From Table 2, it is seen that both ESP ensembles outperform other representations in carbonyl carbon RCSA. While for nitrogens and amide hydrogens, the performance of ESP ensembles is slightly worse than average structure representations but comparable to other ensemble representations.

In ideal situations, a refinement/weighting using RDC data would implicitly improve the RSCA predictions of the structure/ensemble as an optimization of the bond vector orientation by the RDC data also improves the chemical shift tensor orientation of the involved atoms (chemical shift tensor orientations of N and HN atoms depend upon NH bond vector orientation encoded in NH RDC data while those of Carbonyl atoms depend upon CN bond vector orientation provided by CN RDC data). However, noise in experimental RDC data along with errors in structure/ensemble models preclude such ideal situations. Consequently, RCSAs are considered mostly independent from RDC data and were commonly used in cross-validation for observables determined by RDCs (Cornilescu et al. 1998).

Importance of the "switched" conformation

Along with other differences between ERNST and enhanced ERNST, the "switched" conformation, represented by 2G45-E, was given a population weight of $\sim$0.30 by our weighting protocol (Vammi et al. 2014) in

the enhanced ERNST ensemble. This switched conformation may be related to the conformational state that was reported earlier (Massi et al. 2005), though the latter (Massi et al. 2005) may represent a different conformation state (Huang et al. 2011). The weight assigned to the switched conformation in our case also is higher than theirs (Massi et al. 2005). Comparing ERNST without the "switched" conformation and that with (rows 3 & 4), the latter performs better, confirming the importance of the "switched" conformation.

ESP ensembles reproduce amide exchange rates well

Table 3 summarizes the results of pKa predictions by both single structure representations and ensembles. The number of residues whose predicted pKa values deviate by more than 1 unit is listed for different representations and in the case of single structures, the number of inaccessible amide hydrogens is given on the second column.

Ensembles naturally incorporate backbone flexibility, potentially increasing the number of surface exposed amide hydrogens than an average structure representation. This is evident from Table 3 (column 2) where the number of amide hydrogens that are exposed in ensemble representations but are inaccessible in single structure representations is listed. Figure 1 plots the experimental pKa values in comparison to the predicted pKa values by various ensemble representations of Ubiquitin. A single index, the squared sum of the deviations, is given to every ensemble in the figure to give an overall sense of the quality of the predictions. Only residues exposed significantly in the X-ray, MUMO, EROS and ERNST ensembles and having an experimental pKa value of $\sim$5.0 or higher are shown. [Since a different program was used to compute surface accessibility, our pKa predictions differ from LeMaster and colleagues' computations for some of the residues (Hernández et al. 2010)].

pKa predictions are not possible for residues 24, 31–36, 40–42, 48, 51 and 57–60, even though these residues exhibit high experimental exchange rates. This is because none of the ensembles has any surface exposed amides for these residues, which is needed to reproduce pKa values properly. The observed high experimental exchange rates for these residues could be from sparsely populated conformational states of Ubiquitin, which are not captured by any ensemble representations. Due to this reason, amide exchange rates serve as only a weak cross-validation compared to other experimental validations.

The weighted X-ray ensemble predicts the experimental pKa values quite well, having an overall performance better than all the unweighted ensembles. Likewise, the enhanced ERNST ensemble also predicts the experimental pKa's better than the unweighted ensembles. Comparing with ERNST itself (purple squares), enhanced ERNST (red triangles) performs significantly better on many residues. In addition to having more residues whose predicted pKa values deviate by more than 1 unit from experimental values (see Table 3, column 1), single structure representations suffer also from the lack of solvent exposure for many amide hydrogens (See Table 3, column 2).

In summary, both ESP ensembles (i.e., weighted X-ray and enhanced ERNST) perform well in predicting the experimental pKa values. This further validates that ESP ensembles are of high quality.

Solution scattering profile

Solution scattering profiles are observed scattered intensities of X-rays that are collected as a function of the scattering vector q. Typically a q value of 0 to $\sim$0.3 Å falls

Table 3 Summarized results of pKa predictions by different representations of Ubiquitin

# of residues for which the pKa predictions are off by more than 1 unit	# of residues whose amide hydrogens are not exposed	Description
5	0	Weighted X-ray
7	0	ERNST (Fenwick et al. 2011)
5	0	Enhanced ERNST
10	0	EROS (Lange et al. 2008)
6	0	MUMO (Richter et al. 2007)
9	5	1UBQ (Vijay-Kumar et al. 1987)
9	6	1D3Z (Cornilescu et al. 1998)
7	5	2MJB (Maltsev et al. 2014)

The number of residues whose absolute deviations are greater than 1 are shown in column 1. Additionally, for single structure representations, the number of residues for which no prediction could be made due to buried amide hydrogens (but exposed in ensembles) is given in column 2

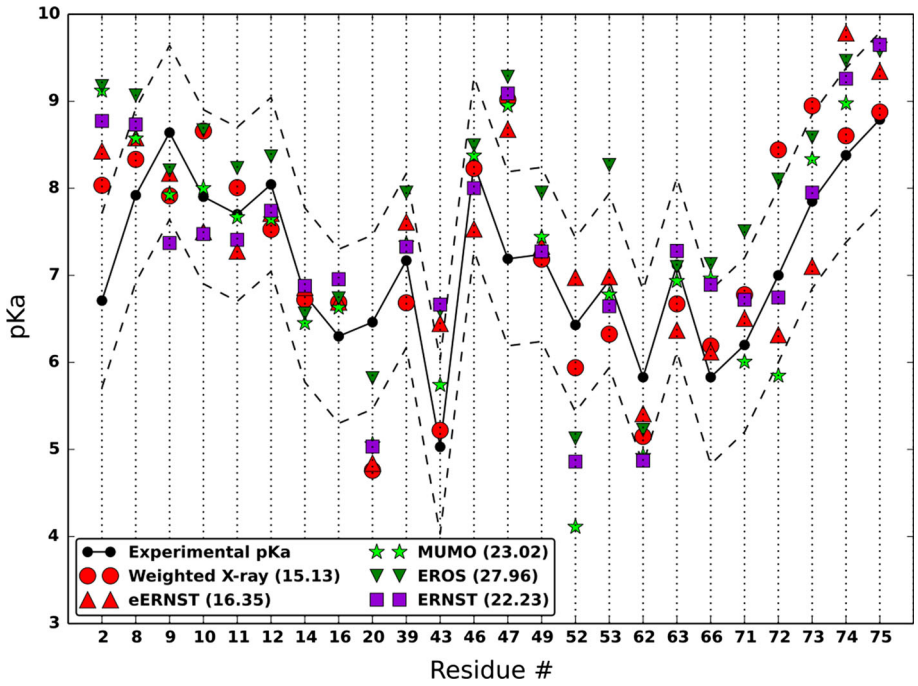

Fig. 1 Experimental pKa values for different residues of Ubiquitin in comparison to the pKa's predicted by different representations of Ubiquitin. Only the hydrogens that are significantly exposed in all the ensembles (X-ray, EROS, ERNST, and MUMO) are shown here. A single index, the squared sum of the deviations, is given to every ensemble to give an overall sense of the quality of the predictions

into the Small Angle X-ray scattering (SAXS) regime while the range for the Wide Angle X-ray scattering (WAXS) regime is ~0.1–2.5 Å. The information encoded in these two regimes along with the results obtained for different structure or ensemble representations of Ubiquitin are presented in the following two sections.

Small angle X-ray scattering (SAXS)

Scattering intensities observed at SAXS encode information about the overall size and shape of the molecule, radius of gyration (R_g) and other low-resolution information (Makowski 2010). Table 4 lists the χ value obtained by different representations of Ubiquitin.

From Table 4 it is seen that, both weighted X-ray and ERNST ensembles have better χ values than their unweighted counterparts. The decreases in χ value confirm that conformations selected to form these two ESP ensembles and the weights assigned to them are meaningful. However, since SAXS data are of low resolution and are not the best data for validating ensembles, this should be taken only as a weak confirmation. Indeed, average structure representations (1D3Z, 1UBQ, or 2MJB) produce an excellent agreement with the experimental data, implying that at low resolution the native states of Ubiquitin appear to be mostly a single conformation.

Figure 2 plots the relative intensities (I_{exp}/I_{calc}) for different representations. While all the representations perform highly similarly at smaller values of q, at higher values of q (>0.14 Å) single structure representations perform the best, followed by the weighted X-ray ensemble.

Table 4 SAXS or WAXS χ values obtained for different representations of Ubiquitin

SAXS χ	WAXS χ	Ensemble
1.24	3.45	Weighted X-ray
1.28	3.65	Unweighted X-ray
1.49	3.75	ERNST
1.37	3.00	Enhanced ERNST
1.27	4.53	EROS
1.36	3.99	MUMO
1.04	4.87	1UBQ
1.17 (0.89)	3.40 (3.59)	1D3Z
0.84 (0.97)	4.98 (4.16)	2MJB

The χ values obtained by using only the first model of 1D3Z and 2MJB are shown in the respective rows. The χ values obtained by using only the first model of 1D3Z or 2MJB are given in parentheses

Wide angle X-ray scattering (WAXS)

Scattering intensities observed at wider angles (higher q) encode information of higher resolution than SAXS but at the cost of potentially bringing in a higher noise level since the intensity of solution scattering also increases. Since data used in this analysis are limited to the range of q values that are less than 1.0 Å, the extent of this noise is limited. WAXS data are often used to validate structural

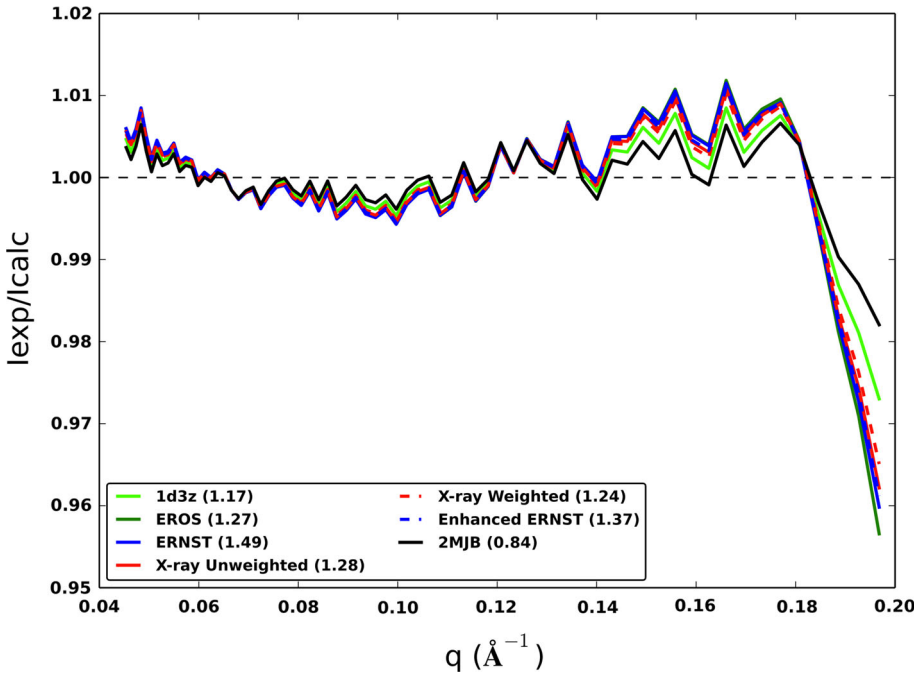

Fig. 2 Relative intensities (I_{exp}/I_{calc}) as a function of q in the SAXS regime for different representations of Ubiquitin. The χ values obtained for different representations also are given

models and to identify structural changes (Makowski 2010). Table 4 lists the WAXS χ values obtained for different representations of Ubiquitin.

Because of its much higher resolution, WAXS data is able to detect conformation state heterogeneity within the native state ensemble. Our first observation based on the WAXS results in Table 4 is that ensemble representations generally do better than the average or single structure representations (1UBQ and 2MJB). 1D3Z is an exception as its WAXS χ value is comparable to those by ensemble representations. It is not clear why 2MJB does worse in this respect than 1D3Z. Perhaps it is because the backbone dynamics captured by the newly determined RDCs (Squalamine and Pf1) with which 2MJB was refined is different from the dynamics represented in other RDC datasets. Secondly, weighted X-ray and enhanced ERNST (the two ESP ensembles) are better than unweighted X-ray ensemble and ERNST ensemble respectively. Thirdly, though weighted X-ray (16 conformations) and weighted ERNST (36 conformations) have significantly fewer conformations than the unweighted X-ray (143 conformations) and ERNST (640 conformations), and the other ensembles such as EROS (116 conformations) and MUMO (144 conformations), these two ESP ensembles clearly outperform the other ensembles in WAXS χ values. This implies that ensemble sizes ought to be fairly limited to avoid overfitting, and that conformations in an ensemble should not be too spread out, and that having too many conformations makes an ensemble highly susceptible to overfitting.

Figure 3a plots the detailed relative intensities (I_{exp}/I_{calc}) computed from different Ubiquitin representations in the WAXS regime. The scattering curves for different representations of Ubiquitin in comparison to the experimental WAXS data is given as in Fig. 3b, which shows that all the representations display a similar trend and peak positions.

To assess the necessity of having ensemble in interpreting the WAXS curves, the spread of the predicted scattering curves for individual members of ensemble in comparison to the ensemble average of the two ESP ensembles is plotted in Fig. 4. The χ values of both ESP ensembles (3.45 for weighted X-ray and 3.00 for eERNST) are lower than the averages of χ values of individual conformations in the ensembles (3.77 for weighted X-ray and 4.11 for eERNST). It can also be seen from the figure that there is sizeable spread among the conformations within the ensemble, though the overall trend is quite similar. Both Figs. 3 (especially panel b) and 4 indicate that the WAXS data provides only a qualitative assessment of these representations (average structures or ensembles) and cannot unequivocally rank their quality.

Table 5 summarizes the results obtained for different experimental data sources used in this work. Only the experimental sources for which significant differences exhibit among the three representations are tabulated. The three representations are qualitatively ranked (Very good > Good > Poor) based on their performance in reproducing these experimental data. From the table, we can see that ESP ensembles outperform regular ensembles (EROS and ERNST) in reproducing all the experimental

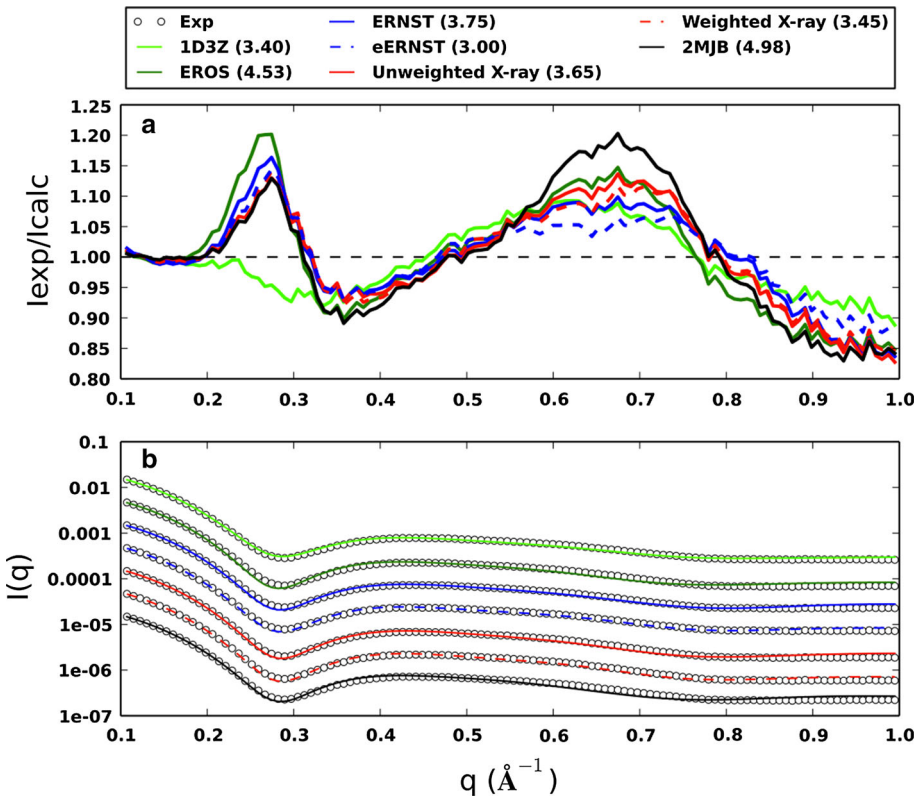

Fig. 3 **a** Relative intensities (I_{exp}/I_{calc}) as a function of q in the WAXS regime for different representations of Ubiquitin. The χ values obtained for different representations also are given. **b** Comparison between experimental WAXS data and predicted curves from different representations of Ubiquitin. Data sets are offset along the y-axis for easier visualization

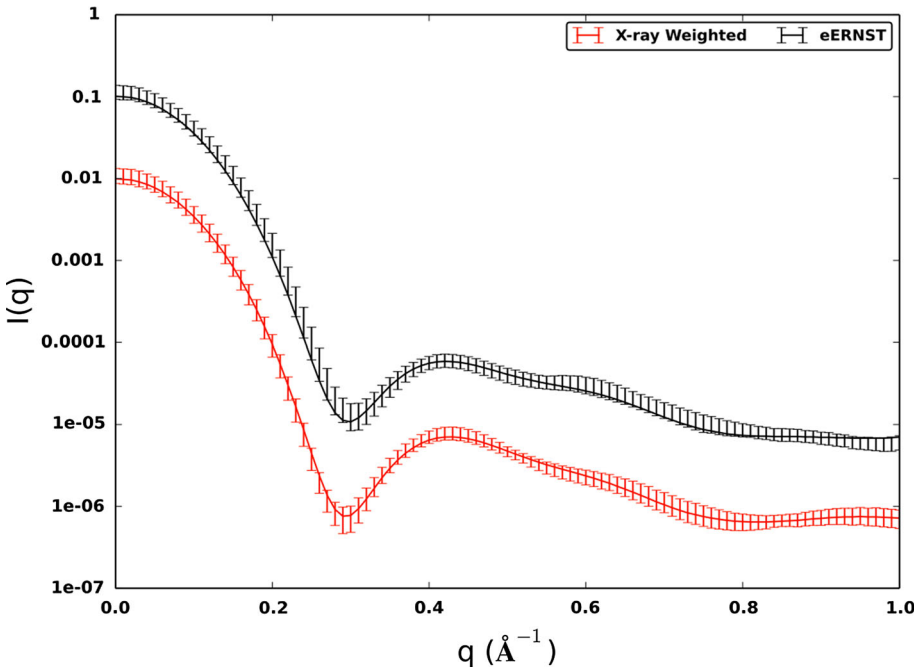

Fig. 4 Spreads (shown as *error bars*) of the predicted scattering curves of the individual members of the ensembles in comparison to the ensemble averages (in *solid lines*) of the two ESP ensembles. Data sets are offset along the y-axis for easier visualization. The χ values of both ESP ensembles (3.45 for weighted X-ray and 3.00 for eERNST) are lower than the averages of χ values of individual conformations in the ensembles (3.77 for weighted X-ray and 4.11 for eERNST)

data except for NH RDCs (it should be noted that regular ensembles have significant deviations from amide planarity that may have contributed to the lower Q-factors in their NH RDCs). Putting these together, it seems that the optimal way to represent the native states of a protein is to use (1) a *small number* of conformations, and (2) with *relative populations*, as in ESP ensembles. This should be the case especially for proteins that have distinct conformation states, while for other proteins single structure representation may be sufficient for most scenarios.

Table 5 A qualitative summary of the cross-validation results using different experimental sources

	NH RDC	pKa	Carbonyl RCSA	Nitrogen RCSA	WAXS	Others
Best regular ensemble (EROS/ERNST)	Very good	Good	Poor	Good	Poor	Similar
Best single structure (1D3Z)	Good	Poor	Good	Very good	Very good	Similar
ESP	Good	Very good	Very good	Good	Very good	Similar

Qualitative ranking (Very good > Good > Poor) is assigned to the three representations by comparing their performance in reproducing RDCs, RCSA, pKa, and solution scattering profiles. Only the experimental data sources for which significant differences exhibit among the three representations are tabulated

Conclusions

In this work, by using Ubiquitin as example and extensive experimental data validations, we demonstrate that it is significant to assign relative populations to conformation ensembles and that ESP ensembles, though having a much smaller number of conformations, are of better quality than regular unweighted ensembles. Specifically, we carry out a thorough cross-validation of two ESP ensembles of Ubiquitin that were determined in an earlier work (Vammi et al. 2014), namely, the weighted X-ray ensemble and the enhanced ERNST ensemble, and show that these two ensembles perform extremely well in all four different types of experimental data: the residual dipolar couplings (RDCs), residual chemical shift anisotropy, hydrogen exchange rates, and solution scattering profile. This is not the case with other ensembles. For example, the MUMO ensemble, which performs well in predicting hydrogen exchange rates, does rather poorly in predicting RDCs. The ERNST or EROS ensemble does well in predicting RDCs but does not perform well in predicting hydrogen exchange rates or the residual chemical shift anisotropies. All these three ensembles (namely MUMO, EROS, and ERNST) do rather poorly in reproducing WAXS χ values. As a result, it is reasonable to conclude that the two ESP ensembles portray the Ubiquitin native states more accurately. Both ensembles reveal that there are six conformation states in Ubiquitin native states, two of which have dominating populations over the others. The conformation state with the largest population contains the unbounded conformation of ubiquitin, 1UBQ, while the one with the second largest population corresponds to the "switched" conformation, consisting exclusively of ubiquitin structures in complex with deubiquitinating enzymes (Vammi et al. 2014).

Qualitatively speaking, the idea of having an ensemble with a small number of conformation states is advantageous. It both captures the dynamical nature of the native state (for which a single average structure is often insufficient to account for) and maintains a strong confidence on the validity of the conformation states. It is the most natural extension of the average structure representation. In contrast, confidence on any individual conformation that it truly belongs to the ensemble is elusive in regular Ubiquitin ensembles since they contain so many conformations and the removal of any single conformation hardly affects the ensemble. Consequently, these ensembles are highly susceptible to over-fitting.

Acknowledgments Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged. The authors would also like to thank Dr. LeMaster for his invaluable help in the initial phase of hydrogen exchange calculations. The authors would also like to extend thanks to Dr. Grishaev for his invaluable advice and help in computing solution scattering profiles.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ángyán AF, Gáspári Z (2013) Ensemble-based interpretations of NMR structural data to describe protein internal dynamics. Molecules 18:10548–10567

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

Choy WY, Forman-Kay J (2001) Calculation of ensembles of structures representing the unfolded state of an SH3 domain. J Mol Biol 308:1011–1032

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
Cornilescu G, Bax A (2000) Measurement of proton, nitrogen, and carbonyl chemical shielding anisotropies in a protein dissolved in a dilute liquid crystalline phase. J Am Chem Soc 122:10143–10154
Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120:6836–6837
DePristo MA, de Bakker PI, Blundell TL (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 12:831–838
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
Fisher CK, Huang A, Stultz CM (2010) Modeling intrinsically disordered proteins with bayesian statistics. J Am Chem Soc 132:14919–14927
Frauenfelder H, Sligar SG, Wolynes PG (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
Grishaev A, Wu J, Trewhella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. J Am Chem Soc 127:16621–16628
Grishaev A, Guo L, Irving T, Bax A (2010) Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. J Am Chem Soc 132:15484–15486
Hernández G, Anderson JS, LeMaster DM (2010) Assessing the native state conformational distribution of ubiquitin by peptide acidity. Biophys Chem 153:70–82
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
Hubbard SJ, Thornton JM (1993) Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London 2
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, 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
LeMaster DM, Anderson JS, Hernández G (2009) Peptide conformer acidity analysis of protein flexibility monitored by hydrogen exchange. Biochemistry 48:9256–9265
Li L, Li C, Sarkar S, Zhang J, Witham S, Zhang Z, Wang L, Smith N, Petukh M, Alexov E (2012) DelPhi: a comprehensive suite for DelPhi software and associated resources. BMC Biophys 5:9
Liu Y, Prestegard J (2010) A device for the measurement of residual chemical shift anisotropy and residual dipolar coupling in soluble and membrane-associated proteins. J Biomol NMR 47:249–258
MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha SA (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616
Makowski L (2010) Characterization of proteins with wide-angle X-ray solution scattering (WAXS). J Struct Funct Genomics 11:9–19
Maltsev AS, Grishaev A, Roche J, Zasloff M, Bax A (2014) Improved cross validation of a static ubiquitin structure derived from high precision residual dipolar couplings measured in a drug-based liquid crystalline phase. J Am Chem Soc 136:3752–3755
Massi F, Grey MJ, Palmer AG (2005) Microsecond timescale backbone conformational dynamics in ubiquitin studied with NMR R1ρ relaxation experiments. Protein Sci 14:735–742
McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793
Phillips GN (2009) Describing protein conformational ensembles: beyond static snapshots. F1000 Biol Rep 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
Putnam CD, Hammel M, Hura GL, Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys 40:191–285
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
Saitô H, Ando I, Ramamoorthy A (2010) Chemical shift tensor–the heart of NMR: insights into biological aspects of proteins. Prog Nucl Magn Reson Spectrosc 57:181
Schwieters CD, Clore GM (2007) A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. Biochemistry 46:1152–1166
Sidhu A, Surolia A, Robertson AD, Sundd M (2011) A hydrogen bond regulates slow motions in ubiquitin by modulating a β-turn flip. J Mol Biol 411:1037–1048
Svergun DI, Koch MHJ (2003) Small-angle scattering studies of biological macromolecules in solution. Rep Prog Phys 66:1735
Svergun D, Barberato C, Koch MHJ (1995) CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28:768–773
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
Tripathy C, Zeng J, Zhou P, Donald BR (2012) Protein loop closure using orientational restraints from NMR data. Proteins: structure. Function, and Bioinformatics 80:433–453
Vammi V, Lin T-L, Song G (2014) Enhancing the quality of protein conformation ensembles with relative populations. J Biomol NMR 58:209–225
Vijay-Kumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 194:531–544