J Biomol NMR (2007) 38:221–235
DOI 10.1007/s10858-007-9162-x

ARTICLE

Predicting $^{13}C^{\alpha}$ chemical shifts for validation of protein structures

Jorge A. Vila · Myriam E. Villegas · Hector A. Baldoni · Harold A. Scheraga

Received: 5 February 2007 / Accepted: 20 April 2007 / Published online: 9 June 2007

Abstract The $^{13}C^{\alpha}$ chemical shifts for 16,299 residues from 213 conformations of four proteins (experimentally determined by X-ray crystallography and Nuclear Magnetic Resonance methods) were computed by using a combination of approaches that includes, but is not limited to, the use of density functional theory. Initially, a validation test of this methodology was carried out by a detailed examination of the correlation between computed and observed $^{13}C^{\alpha}$ chemical shifts of 10,564 (of the 16,299) residues from 139 conformations of the human protein ubiquitin. The results of this validation test on ubiquitin show agreement with conclusions derived from computation of the chemical shifts at the *ab initio* Hartree–Fock level. Further, application of this methodology to 5,735 residues from 74 conformations of the three remaining proteins that differ in their number of amino acid residues, sequence and three-dimensional structure, together with a new scoring function, namely the *conformationally averaged* root-mean-square-deviation, enables us to: (a) offer a criterion for an accurate assessment of the *quality* of NMR-derived protein conformations; (b) examine whether X-ray or NMR-solved structures are better representations of the observed $^{13}C^{\alpha}$ chemical shifts in solution; (c) provide evidence indicating that the proposed methodology is more accurate than automated predictors for validation of protein structures; (d) shed light as to whether the agreement between computed and observed $^{13}C^{\alpha}$ chemical shifts is influenced by the identity of an amino acid residue or its location in the sequence; and (e) provide evidence confirming the presence of dynamics for proteins in solution, and hence showing that an ensemble of conformations is a better representation of the structure in solution than any single conformation.

Keywords ^{13}C chemical shift prediction · Solution structure · Protein structure validation · X-ray and NMR structures · Ubiquitin

Introduction

Since the observation of numerous carbon resonances of Hen Egg-White Lysozyme (Allerhand et al. 1973), it has been recognized that this NMR spectroscopic technique constitutes a source of important structural information for proteins. Growing interest followed in the characterization of the factors affecting observed ^{13}C chemical shifts, among other nuclei, and how these measured quantities can be used for protein structure determination and/or refinement. The existence of a plethora of papers and reviews regarding the prediction of chemical shifts in biological systems is evidence of the importance of the problem (Oldfield and Allerhand 1975; Howard and Lilley 1978; Malthouse 1985; Chesnut and Moore 1989; de Dios et al. 1993a, b; Case et al. 1994; Laws et al. 1995; Luginbühl et al. 1995; Jameson 1996; Wishart and Nip 1998; Iwadate et al. 1999; Berman et al. 2000; Case 2000; Wishart and Case 2001; Xu and Case 2001; Oldfield 2002; Xu and Case

Electronic Supplementary Material The online version of this article (doi: 10.1007/s10858-007-9162-x) contains supplementary material, which is available to authorized users.

J. A. Vila · H. A. Scheraga (✉)
Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA
e-mail: has5@cornell.edu

J. A. Vila · M. E. Villegas · H. A. Baldoni
Instituto de Matemática Aplicada San Luis, CONICET, Universidad Nacional de San Luis, Ejército de Los Andes 950-5700 San Luis, Argentina

2002; Meiler 2003; Neal et al. 2003; Hunter et al. 2005; Dyson and Wright 2005; Moon and Case 2006). The motivation to predict ^{13}C chemical shifts arises from the fact that they can reveal detailed information about protein structure because their magnitudes are very sensitive to, and depend mainly on, the backbone torsional angles (ϕ, ψ) (Spera and Bax 1991), although the influence of the side-chain torsional angles χ's cannot be disregarded (Havlin et al. 1997; Pearson et al. 1997; Xu and Case 2001; Sun et al. 2002; Villegas et al. 2007). However, computation of the chemical shifts at the ab initio Hartree–Fock or density functional theory (DFT) level for protein structures represents a formidable task and, despite the existence of powerful computational resources, this task cannot be accomplished unless some approximations are adopted. Thus, for example, Xu and Case (2001) made use of a database of peptide chemical shifts, computed at the DFT level, for an automatic prediction of ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts with their SHIFTS program (Xu and Case 2001). The database was constructed from computed chemical shifts of 1335 peptides whose backbone torsional angles are limited to areas of the Ramachandran map around helical and β-sheet conformations. They obtained very good agreement between computed and observed chemical shifts in several applications to proteins. However, the adopted approach limits the application of the methodology to regions of regular secondary structure, i.e., to about 40% of the residues in proteins (Xu and Case 2001).

A different approach to solve this problem at the quantum chemical level has been introduced by Oldfield's group (Havlin et al. 1997; Sun et al. 2002). They have been able to compute the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shielding surfaces for the 20 naturally occurring amino acid residues by using a combination of approaches that includes, but is not limited to, the following: (1) embedding each amino acid residue in an N-formyl-amino acid amide molecule; (2) using the *locally dense* basis set [6-311++G(2d,2p)] approach for the heavy atoms and hydrogen of the backbone C^{α} and C^{β}, but not beyond C^{β} (except for isoleucine side chains for which a more extended basis set was used), with a 3-21G basis set for the other atoms; and (3) choosing the basic side-chain geometries that are the most abundant forms present in proteins. Their results show good agreement with either X-ray or average-solution NMR structures of the protein ubiquitin, and rationalize many interesting behaviors of observed chemical shifts in proteins, such as the observed increase in the isotropic shielding of β-sheet over helical geometries (Spera and Bax 1991). However, there is considerable dependence of ^{13}C chemical shifts on both the backbone torsional angles (ϕ, ψ) and the side-chain torsional angles χ's (Havlin et al. 1997; Pearson et al. 1997). After a detailed comparison involving X-ray and NMR-derived protein conformations of ubiquitin, Xu and Case (2001) noted that modification of side-chain orientation is frequently very useful for improving the prediction of ^{13}C chemical shifts. This is not unexpected since the three torsional angles ϕ, ψ and χ^1 are not independent of each other but involve the bonds connected to a common C^{α} atom (Dunbrack and Karplus 1994; Chakrabarti and Pal 1998).

Recently, Lindorff-Larsen et al. (2005) presented a protocol for the experimental determination of ensembles of protein conformations that contain the native structure and its associated dynamics. Among other important conclusions, these authors noted that side chains, even in the core of the protein, occupy multiple rotameric states (having a liquid-like character) and show considerable variability within each rotamer, e.g., 38 out of 68 non-Ala/Gly residues in the human protein ubiquitin were reported to populate more than one χ^1 rotamer (Lindorff-Larsen et al. 2005). In line with their observation, in recent work (Villegas et al. 2007), we have been able to show that, for most of the naturally occurring amino acid residues in β-sheet strands, proper consideration of the side-chain preferences, in particular but not limited to the χ^1 torsional angle, may be crucial for good agreement between observed and computed $^{13}C^{\alpha}$ chemical shifts. Hence, for protein structure refinement at a high-level of accuracy, or evaluation of the quality of NMR-derived structures, proper consideration of the side-chain positions may also be required.

Therefore, the question arises as to whether it is possible to predict $^{13}C^{\alpha}$ chemical shifts for proteins without either presupposition about the side-chain positions or a priori assignment of the conformation of any amino acid residue to a particular region of the Ramachandran map. An additional question that arises is: can the computed $^{13}C^{\alpha}$ chemical shifts be used to assess the *quality* of protein conformations? To address these questions, use is made here of the following combination of approaches: (1) each of the experimentally determined conformations of a protein was *regularized*; (2) each amino acid **X** in the protein sequence was treated as a terminally blocked tripeptide in the regularized experimentally determined backbone and side-chain conformations; and (3) computation of the $^{13}C^{\alpha}$ chemical shifts for each amino acid residue **X** was carried out by using the *locally dense* basis set approach. It is worth noting that, although the proposed methodology could be used to compute the chemical shifts for all the nuclei of a given amino acid residue, we will focus here only on the computed $^{13}C^{\alpha}$ chemical shifts for all the 20 naturally occurring amino acid residues. In this respect, there is abundant information indicating that the main factors affecting $^{13}C^{\alpha}$ chemical shifts are the backbone (ϕ, ψ) dihedral angles and the side-chain torsional angles

(Spera and Bax 1991; Pearson et al. 1997; Iwadate et al. 1999; Villegas et al. 2007) with no influence of amino acid sequence (Iwadate et al. 1999; Xu and Case 2002).

In this study, we first validated our methodology by comparing the results obtained from the analysis of the computed $^{13}C^{\alpha}$ chemical shifts of ubiquitin with that of an independent study carried out by Sun et al. (2002) at the ab initio Hartree–Fock level. Then, we applied this methodology to compute the $^{13}C^{\alpha}$ chemical shifts for three proteins that differ in their number of amino acid residues, sequence and function, as well as in their three-dimensional structure.

The results obtained from a detailed analysis of the correlation between observed and computed $^{13}C^{\alpha}$ chemical shifts of four proteins, listed in Table 1, enable us to (a) estimate the *quality* of protein structures, i.e., by using the computed $^{13}C^{\alpha}$ chemical shifts to obtain a new scoring function: the *conformationally averaged* root-mean-square-deviation (*ca*-rmsd$^{\alpha}$); (b) examine whether an X-ray or an NMR-solved structure is a better representation (in terms of the *ca*-rmsd$^{\alpha}$) of the observed $^{13}C^{\alpha}$ chemical shifts in solution; (c) analyze the reliability of the ranking of NMR-solved structures in the Protein Data Bank (PDB); and (d) provide some evidence that the current methodology is more accurate for validation of protein structures than automated server predictors, such as SHIFTS and SHIFTX (Xu and Case 2002; Neal et al. 2003).

Methods

Experimental set of structures

The experimental set of structures contains four proteins (experimentally determined by X-ray crystallography and NMR methods listed in Table 1) (Vijay-Kumar et al. 1987; van Nuland et al. 1994; Cornilescu et al. 1998; Napper et al. 1999; Amann et al. 2003; Babini et al. 2004; Lindorff-Larsen et al. 2005), and their coordinates were obtained from the Protein Data Bank (PDB) (Berman et al. 2000). These proteins are identified by a four-symbol PDB code. Details of the four-protein set analyzed, including: (i) protein names and PDB codes; (ii) total number of amino acid residues; (iii) experimental conditions; (iv) number of conformers used; (v) structural class, and any other data relevant to this work, are listed in Table 1.

Table 1 Set of proteins analyzed[a]

Protein name[b] (PDB code)	Experimental conditions[c]	<rmsd$^{\alpha}$>[d] (ppm)	Number of observed ^{13}C chemical shifts[e]
1D3Z (76) [10] (i)	NMR (298, 6.6) [TSP]	2.7 ± 0.2 **[2.5]** **(2.3)**	76 (*6457*) [$\alpha + \beta$]
1UBQ (76) [1] (ii)	X-ray (1.8)	2.7 **[2.7]** **(2.6)**	
1XQQ (76) [128] (iii)	Molecular dynamics	3.0 ± 0.2 **[2.4]** **(2.1)**	
1HDN (85) [30] (iv)	NMR (303, 6.5) [TSP]	7.2 ± 0.2 **[6.9]**	85 (*2371*) [$\alpha + \beta$]
1CM2 (85) [1] (v)	X-ray (1.8)	2.7 **[2.7]**	
1M9O (40) [23] (vi)	NMR (293, 5.8) [TSP]	4.4 ± 0.4 **[3.5]**	38 (*5525*) [n/a]
1TTX (109) [20] (vii)	NMR (298, 6.5) [DSS]	3.4 ± 0.2 **[2.7]**	109 (*6705*) [α]

[a] For which the $^{13}C^{\alpha}$ chemical shifts were computed, as explained in the ''Methods'' section

[b] Four-symbol code used to designate the protein in the Protein Data Bank (Berman et al. 2000); (i), (ii) and (iii) pertain to ubiquitin; (iv) and (v) pertain to the Histidine-containing Phosphocarrier protein HPr; (vi) pertains to tristetraprolin; and (vii) pertains to β-parvalbumin. The total number of residues per protein is in parentheses; the total number of conformers used in this work is in brackets. Hence, the total number of residues is 16,299. References to the authors of the experimental or theoretical characterizations of the structures are in parentheses, namely: (i) Cornilescu et al. (1998); (ii) Vijay-Kumar et al. (1987); (iii) Lindorff-Larsen et al. (2005); (iv) van Nuland et al. (1994); (v) Napper et al. (1999); (vi) Amann, et al. (2003); and (vii) Babini et al. (2004)

[c] Experimental method; temperature in degrees Kelvin and pH (in parentheses); and reference used for the observed $^{13}C^{\alpha}$ chemical shifts [in brackets]; if the structure was solved by X-ray diffraction, the resolution in Å is in parentheses. For details of the molecular dynamics refinement against order parameters used to generate the 128 conformers for 1XQQ see Lindorff-Larsen et al. (2005)

[d] The mean values <rmsd$^{\alpha}$> were computed using Eq. 4, as described in the ''Methods'' section. *Ca*-rmsd$^{\alpha}$ values are reported in boldface and brackets, and in bold face in parentheses, with correction factors of 1.82 ppm and 1.25 ppm, respectively, and were computed using Eq. 2 as described in the ''Methods'' section and as discussed in Results and discussion section ''Analysis of the error distributions of...''. For a single structure, as for X-ray-derived conformations, the reported value was computed by using Eq. 3 as described in the ''Methods'' section

[e] Total number of ^{13}C chemical shifts as listed in the *accession number* under which the data can be found (Biological Magnetic Resonance Data Bank); *accession number* (italicized in parenthesis) from where the observed values for the $^{13}C^{\alpha}$ chemical shifts were taken to compute the correlation coefficient (Press et al. 1992), *R*, the rmsd$^{\alpha}$ and the *ca*-rmsd$^{\alpha}$; the structural class is given [in brackets]; [n/a] is used to denote the non-existence of any regular secondary structure elements

Conversion of the experimental structures from flexible to rigid ECEPP geometry

In order to carry out the present study, all the experimentally determined conformations (a) were *regularized*, i.e., all residues were replaced by the standard ECEPP/3 residues (Némethy et al. 1992) in which bond lengths and bond angles are fixed (rigid geometry approximation), and (b) hydrogen atoms were added, if necessary. The conversion process was carried out by generating the new rigid-geometry conformation from the N-terminus by adding one residue at a time and minimizing the root-mean-square-deviation (rmsd) between all heavy atoms in the generated fragment and the corresponding fragment in the experimental structure (Ripoll et al. 2005). The procedure was iterated until the C-terminal group was added to the chain. The final conformations resulting from this regularization procedure are quite close to the experimental ones in all cases, with rmsd values for all the heavy atoms up to 0.2 Å, as for 1 UBQ.

Quantum-chemical calculations of the $^{13}C^{\alpha}$ chemical shift

It is possible to obtain theoretical shielding values of good quality by using large basis sets located only on the atoms whose shifts are of interest while the rest of the atoms in the molecule are treated with more modest basis sets (Chesnut and Moore 1989). This is called the locally dense basis set approach, and its use enables us to minimize the length of the chemical-shift calculations while maintaining the accuracy of the results (Laws et al. 1995; Pearson et al. 1997; Vila et al. 2003; Vila et al. 2004a, 2004b; Villegas et al. 2007). A recent analysis of the effect of the locally dense approximation to treat the effect of near-neighbor residues in both sequence and space, such as those in consecutive strands of a β-sheet, has been presented by Villegas et al. (2007), who reported that an extraordinary reduction of computational cost without significant loss of accuracy was found.

Based on these observations, in this work we decided to compute the $^{13}C^{\alpha}$ chemical shifts by treating a single residue, i.e., the guest residue X, with a locally dense basis set [6-311+G(2d,p)], while the rest of the molecule is treated with the simpler 3-21G basis set. This notation refers to the basic basis sets of Pople and co-workers (Hehre et al. 1986) as implemented in Gaussian-98 (Frisch et al. 1998). All the calculated isotropic shielding values (σ) were referenced with respect to a tetramethylsilane (TMS) $^{13}C^{\alpha}$ chemical shift scale (δ), as described previously (Vila et al. 2002).

Following the approach of Xu and Case (2001) for the computation of the shielding, all the ionizable groups were assumed to be uncharged. Alternatively, the charges could have been computed by exploring the 2^{ξ} possible ionization states at a given fixed pH for all the ξ ionizable groups (Ripoll et al. 1996, 2004, 2005) of every conformation of the ensemble. Whether this alternative method would lead to a better representation of the experimentally observed $^{13}C^{\alpha}$ chemical-shift values than the procedure used here (based on uncharged side-chains) is a very important issue. However, such analysis is beyond the scope of this work, but is presently undergoing consideration in our laboratory.

Conversion of the computed TMS-referenced values for the $^{13}C^{\alpha}$ chemical shifts to either 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS) or 3-(Trimethylsilyl) propionate sodium salt (TSP) references was carried out by adding 1.7 ppm or 1.82 ppm, respectively, to the computed values (Wishart et al. 1995).

Method used to compute the $^{13}C^{\alpha}$ chemical shifts in proteins

The following approaches were used: (1) each of the experimentally determined conformations was *regularized*; (2) each amino acid **X** in the protein sequence was treated as a terminally blocked tripeptide with the sequence Ac-G**X**G-NMe in the conformation of the regularized experimental protein structure; and (3) computation of the $^{13}C^{\alpha}$ chemical shifts was carried out with a 6-311+G(2d,p) *locally dense* basis set for each amino acid residue **X**, while the remaining residues in the tripeptide were treated with a 3-21G basis set. In step (2), a local minimization for relaxing any possible steric interactions between the side chain of amino acid **X** and the rest of the tripeptide was carried out using the ECEPP/3 force-field. However, the *only* torsional angles allowed to vary during the minimization procedure were those of the N- and C-terminal blocking groups and the neighboring glycine amino acid residues. In other words, all backbone and side-chain torsional angles of the amino acid **X** were identical to those obtained after the regularization procedure. The $^{13}C^{\alpha}$ chemical shifts can be computed, with the current methodology, for both the reduced and oxidized forms of cysteine. However, for the only protein in Table 1 that contains cysteines (in the reduced form) we omitted a comparison between the computed and observed values for the reasons discussed in section ''Zinc-binding domain protein''.

Although it is possible to use computations of ^{13}C chemical shifts to determine whether the peptide group of proline is in the *cis* or *trans* conformation (Schubert et al. 2002), such computations were not carried out because it is not the goal of this paper to carry out a refinement of the structures analyzed. Moreover, no geometry optimization at the *ab initio* level was carried out because there is evi-

dence (Pearson et al. 1997; Vila et al. 2002) that a geometry-optimized structure has only a very small effect on the computed shielding.

The time for calculation of the isotropic shielding values (σ) varied, depending on the amino acid residue type, from ~40 minutes (for Gly) to ~7 hours (for Arg) with an Athlon 2800+ processor. This means that, by using coarse-grained parallelization, all the $^{13}C^{\alpha}$ chemical shifts for a protein with μ amino acid residues could be computed, on average, in about 7 h with a Beowulf class cluster with μ processors.

Use of the Computed $^{13}C^{\alpha}$ chemical shifts from four proteins

The values of the $^{13}C^{\alpha}$ chemical shifts from 139 conformations of ubiquitin (Table 1) were computed here to validate the proposed methodology by comparing the computed and observed chemical shifts with results obtained from another independent study carried out by Sun et al. (2002), as explained in the ''Results and discussion'' section. The computed values of the $^{13}C^{\alpha}$ chemical shifts for 5,735 residues from 74 conformations of three other proteins (Table 1) were used to estimate the *quality* of the protein structures and examine whether an X-ray or an NMR-solved structure is a better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution.

A new scoring function: the conformationally averaged rmsd$^{\alpha}$

There is abundant evidence indicating that the $^{13}C^{\alpha}$ chemical shifts depend on secondary structure (Spera and Bax 1991; Kuszewski et al. 1995; Iwadate et al. 1999) with no influence of amino acid sequence (Iwadate et al. 1999; Xu and Case 2002). Since a protein in solution exists as an ensemble of conformations, we can assume that the observed chemical shifts $^{13}C^{\alpha}_{\text{observed},\mu}$ for a given amino acid μ can be interpreted as a conformational average over different rotational states represented by a discrete number of different conformations all of which satisfied the NMR constraints. Hence, for each observed value, we can compute the following quantity: $^{13}C^{\alpha}_{\text{computed},\mu} = \sum_{i=1}^{\Omega} \lambda_i \, ^{13}C^{\alpha}_{\mu,i}$, where $^{13}C^{\alpha}_{\mu,i}$ is the computed chemical shift for amino acid μ in conformation i out of Ω protein conformations (that must satisfy the observed NMR constraints, such as NOEs, vicinal coupling constants, Residual Dipolar Coupling constants, etc., from which the conformations were derived), and λ_i is the weight factor for conformation i, with the condition $\sum_{i=1}^{\Omega} \lambda_i \equiv 1$. Under conditions of fast conformational averaging, we will assume that $\lambda_i = 1/\Omega$, i.e., all weight factors contribute equally. Hence, for each amino acid μ, we define a function $\Delta^{\alpha}_{\mu} \cong (^{13}C^{\alpha}_{\text{observed},\mu} - <^{13}C^{\alpha}_{\text{computed}}>_{\mu})$ with

$$<^{13}C^{\alpha}_{\text{computed}}>_{\mu} = (1/\Omega) \sum_{i=1}^{\Omega} {}^{13}C^{\alpha}_{\mu,i} \quad (1)$$

where $1 \le \mu \le N$, with N being the number of observed $^{13}C^{\alpha}$ chemical shifts. Then, a new scoring function to assess the quality of an ensemble of conformations, namely the conformationally averaged rmsd$^{\alpha}$ (*ca*-rmsd$^{\alpha}$), can be defined as:

$$ca\text{-rmsd}^{\alpha} = \left[(1/N) \sum_{\mu=1}^{N} (\Delta^{\alpha}_{\mu})^2 \right]^{1/2} \quad (2)$$

For a single structure, as an X-ray derived one, $\Omega = 1$, and hence,

$$ca\text{-rmsd}^{\alpha} \equiv \text{rmsd}^{\alpha} = \left[(1/N) \sum_{\mu=1}^{N} (^{13}C^{\alpha}_{\text{observed},\mu} - {}^{13}C^{\alpha}_{\text{computed},\mu})^2 \right]^{1/2} \quad (3)$$

The reported mean <rmsd> value, for a given set of Ω conformations is computed as:

$$<\text{rmsd}^{\alpha}> = \left[(1/\Omega) \sum_{i=1}^{\Omega} \text{rmsd}^{\alpha}_{i} \right] \quad (4)$$

The supplementary material

Information not considered crucial for the discussion is provided in a supplementary file, which contains two figures. One of the figures is a ribbon diagram of the superposition of 128 NMR-derived models of the protein ubiquitin, and the other one is the frequency of the error distribution computed from 760 (1D3Z) and 9728 (1XQQ) residues of ubiquitin.

Results and discussion

Validation of the methodology: test on ubiquitin

Predictions of the $^{13}C^{\alpha}$ chemical shifts were carried out for 10,564 residues from 139 ubiquitin protein conformations (the first three entries in Table 1). 138 of these are NMR-derived conformations: 128 of these structures (PDB code: 1XQQ) were reported by Lindorff-Larsen et al. (2005) and 10 structures (PDB code: 1D3Z), shown in Fig. 1, were

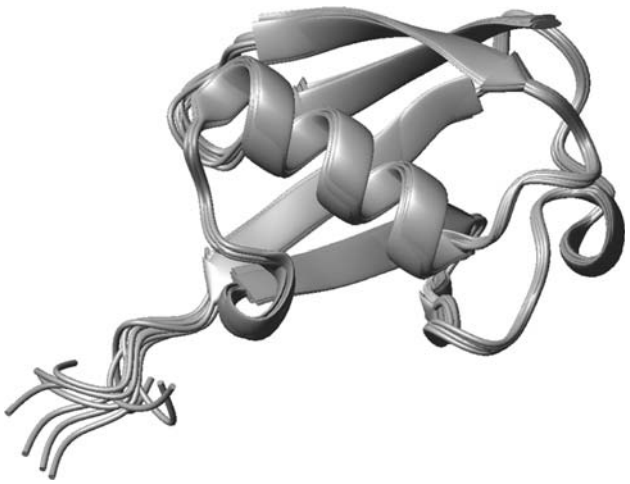

Fig. 1 Ribbon diagram of the superposition of 10 NMR-derived conformations of the protein ubiquitin (PDB code: 1D3Z)

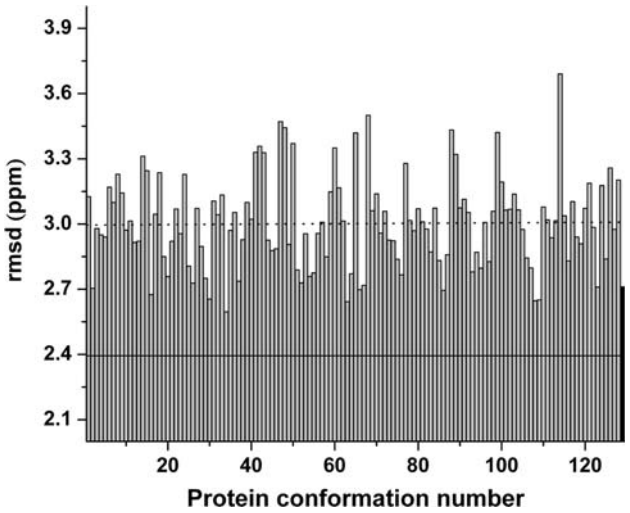

Fig. 2 *Grey filled bars* indicate the rmsd$^{\alpha}$ computed as described in the ''Methods'' section for each of the 128 conformations of ubiquitin (PDB code: 1XQQ). *Black filled vertical bar* indicates the rmsd$^{\alpha}$ computed for the X-ray derived structure of ubiquitin (PDB code: 1UBQ). The *horizontal dotted line* (3.0 ppm) represents the mean-value for the rmsd$^{\alpha}$ computed from 128 conformations of ubiquitin (PDB code: 1XQQ). The *solid horizontal line* (2.4 ppm) indicates the *ca*-rmsd$^{\alpha}$ value computed from 128 conformations of 1XQQ (as explained in section ''Using the *ca*-rmsd$^{\alpha}$ to assess the quality...'')

reported by Cornilescu et al. (1998). The remaining one is an X-ray structure solved at 1.8 Å resolution by Vijay-Kumar et al. (1987) [PDB code: 1UBQ].

A broad dispersion of the rmsd$^{\alpha}$ values among all the 1XQQ conformations is found, i.e., ranging from ~2.6 ppm to ~3.6 ppm, as shown in Fig. 2. Such dispersion has been observed for all four proteins analyzed in this work. Dotted horizontal lines in Figs. 2 and 3 represent the *average* <rmsd$^{\alpha}$> values computed from the 1XQQ (3.0 ± 0.2 ppm) and 1D3Z (2.7 ± 0.2 ppm) set of structures, respectively.

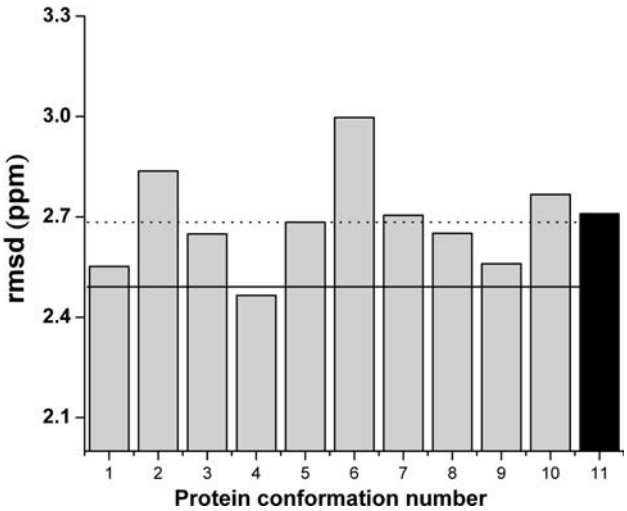

Fig. 3 *Grey filled bars* indicate the rmsd$^{\alpha}$ computed as described in the ''Methods'' section for each of the 10 conformations of ubiquitin (PDB code: 1D3Z). *Black filled vertical bar* indicates the rmsd$^{\alpha}$ computed for the X-ray derived structure of ubiquitin (PDB code: 1UBQ). The *horizontal dotted line* (2.7 ppm) represents the mean-value for the rmsd$^{\alpha}$ computed from 10 conformations of ubiquitin. The *solid horizontal line* (2.5 ppm) indicates the *ca*-rmsd$^{\alpha}$ value computed from 10 conformations of 1D3Z (as explained in section ''Using the *ca*-rmsd$^{\alpha}$ to assess the quality...'')

By using the computed $^{13}C^{\alpha}$ shielding surfaces for all the naturally occurring amino acids residues, Sun et al. (2002) carried out a comparison between computed and observed $^{13}C^{\alpha}$ shielding values for the protein ubiquitin (1D3Z). The best value of the rmsd$^{\alpha}$ deviation reported by Sun et al. (2002), i.e., when using individual amino acid shielding surfaces, is 2.7 ppm whereas, for the X-ray structure (1UBQ), they reported an rmsd$^{\alpha}$ of 3.6 ppm. In our calculations, we found: (a) an *averaged-value* over the 10 NMR-derived conformations of 1D3Z of <rmsd$^{\alpha}$> = 2.7 ± 0.2 ppm, and for the X-ray structure we obtain rmsd$^{\alpha}$ = 2.7 ppm (as shown in Table 1); and (b) a *ca*-rmsd$^{\alpha}$ = 2.5 ppm. Our results for $^{13}C^{\alpha}$ chemical shifts for the protein ubiquitin are in excellent agreement with those obtained by Sun et al. (2002) and, hence, constitute a validation of our methodology. It should be noted that Sun et al. (2002) used a single *average* NMR structure as compared to the ensemble of conformations considered in this work.

Assessment of the quality of derived molecular conformations by using the ca-rmsd$^{\alpha}$

Many *quality* indicators of molecular conformations of proteins have been developed, all based on X-ray structures (Vriend 1990; Morris et al. 1992; Vriend and Sander 1993; Laskowski *et al*. 1993; Pontius et al. 1996; Doreleijers et al. 1998; Wilson et al. 1998; Nabuurs et al. 2004; Melnik et al. 2005; Ban et al. 2006).

Although a discussion of the precision and accuracy of X-ray and NMR-derived conformations (Zhao and Jardetzky 1994; Simon et al. 2005) is beyond the scope of this paper, our interest is to determine the *quality* of the structures as a basis to validate protein conformations. The term *quality* is used here only to establish whether a conformation, or set of conformations, is a ''good representation'' of the observed $^{13}C^{\alpha}$ chemical shifts in solution. It is important to note that such ''good representation'' does not assure that the set of protein conformations would also satisfy other 'quality' indicators such as the 'stereochemical quality' (Laskowski et al. 1993; Vriend 1990), the average numbers of satified NMR constraints, the rmsd$^{\alpha}$ value relative to the mean coordinates, etc. From this point of view, the conclusions derived from our analysis should be considered complementary to other 'quality factor' indicators.

X-ray structures are frequently used as standards for comparison since they appear to be more accurate (Pearson et al. 1995; Celda et al. 1995; Laskowski et al. 1996). However the X-ray structure may not be available or it may not provide, in terms of the $^{13}C^{\alpha}$, a better chemical-shift representation of the structure in solution than that provided by NMR-derived conformations. As an alternative to surmount these problems, we propose to use the values of the *ca*-rmsd$^{\alpha}$-per-residue. Because proteins usually differ, among other things, in the total number of amino acid residues, this normalized *ca*-rmsd$^{\alpha}$ would provide a rapid assessment of the relative quality of different protein conformations.

In the next sections, we will focus on the analysis of the following set of proteins (listed in Table 1): (a) a set of 30 NMR-derived conformations for the histidine-containing phosphocarrier protein (van Nuland et al. 1994) [PDB code 1HDN] and the corresponding X-ray structure solved at 1.8 Å resolution (Napper et al. 1999) [PDB code 1CM2]; and (b) two metal-binding proteins: (i) a set of 23 NMR-derived conformations for the zinc-binding domain from tristetraprolin (Amann et al. 2003) [PDB code 1M9O]; and (ii) a set of 20 NMR-derived conformations for a calcium-binding protein (Human β- Parvalbumin) (Babini et al. 2004) [PDB code 1TTX].

Analysis of a histidine-containing phosphocarrier protein

This protein contains 85 amino acid residues and it belongs to the $\alpha + \beta$ structural class. Figure 4 shows the rmsd$^{\alpha}$ versus the protein conformation number for 30 NMR-derived protein conformations (grey-filled bars) plus the X-ray solved structure (black-filled vertical bar). The X-ray structure differs from the NMR-derived sequences in the identity of the amino acid residue at position 15, i.e., Asp

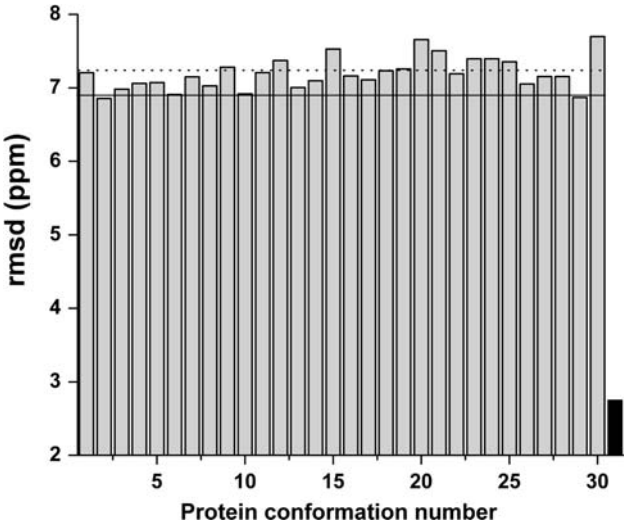

Fig. 4 *Grey filled bars* indicate the rmsd$^{\alpha}$ computed as described in the ''Methods'' section for each of the 30 conformations of a histidine-containing phosphocarrier protein (PDB code: 1HDN). *Black filled vertical bar* indicates the rmsd$^{\alpha}$ computed for the corresponding X-ray structure of the histidine-containing phosphocarrier protein after the mutation of His15 to Asp15 (PDB code: 1CM2). *Horizontal dotted line* (7.2 ppm) represents the mean-value for the rmsd$^{\alpha}$ over the 30 conformations. The *solid horizontal line* (6.9 ppm) indicates the *ca*-rmsd$^{\alpha}$ value computed from 30 conformations of the protein 1HDN by using Eq. 2

for the X-ray conformation and His for the NMR-derived conformations. Regardless of this, it can be seen from Fig. 4 that the X-ray structure is a significantly better representation of the observed $^{13}C^{\alpha}$ chemical shifts (with an <rmsd$^{\alpha}$> of 2.7 ppm) than any of the NMR-derived conformations. The average over all 30 conformations <rmsd$^{\alpha}$> is 7.2 ± 0.2 ppm (shown by a horizontal dotted line in Fig. 4).

Analysis of two metal-binding protein

Zinc-binding domain protein (PDB code 1M9O). This protein was selected because the three-dimensional solution structure of the first domain reveals a novel fold around a central zinc ion. As the authors (Amann et al. 2003) noted, the core structure is disk-like, and the fold is distinct from all previously characterized metal-binding domains (Amann et al. 2003). The only regular secondary structure element is a small one-turn helix from residue 17–20 (see Fig. 5). The structure possesses three Cys residues that are coordinated to the zinc. The observed $^{13}C^{\alpha}$ chemical shifts for the three Cys were omitted from the comparison with the computed values since the *regularized* conformation of the structures does not take the presence of zinc into account. For this reason, the rms deviations reported in Table 1 were computed from 35 out of 38

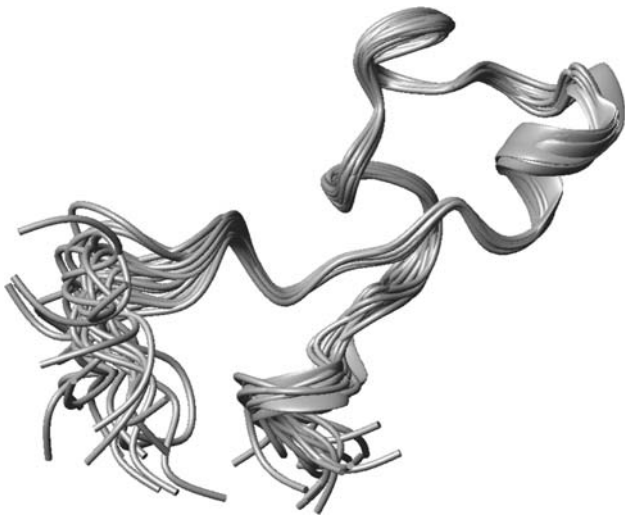

Fig. 5 Ribbon diagram of the superposition of 23 NMR-derived conformations of a metal-binding protein (PDB code: 1M9O)

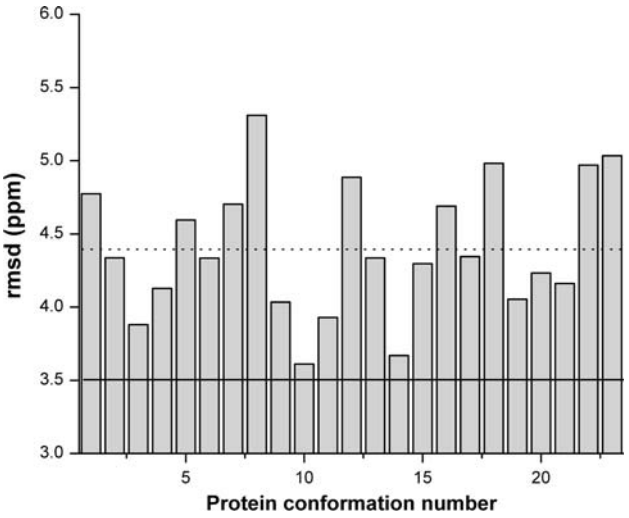

Fig. 6 *Grey filled bars* indicate the rmsd$^{\alpha}$ computed as described in the ''Methods'' section for each of the 23 conformations of a metal-binding protein (PDB code: 1M9O). *Horizontal dotted line* (4.4 ppm) represents the mean-value for the rmsd$^{\alpha}$ over the 23 conformations. The *solid horizontal line* (3.5 ppm) represents the *ca*-rmsd$^{\alpha}$ value computed from 23 conformations of the protein 1M9O by using Eq. 2

observed $^{13}C^{\alpha}$ chemical shifts. Figure 6 shows the distribution of the rmsd$^{\alpha}$ value versus the protein conformation number for 23 NMR-derived protein conformations (grey-filled bars).

The computed values (higher than those obtained for the protein ubiquitin) for the ensemble-averaged rmsd$^{\alpha}$ deviation (as reported in Table 1) imply that it should be possible to improve the structure of 1M9O by using $^{13}C^{\alpha}$ chemical shift information.

Calcium-binding protein (PDB code 1TTX). This is a 109 amino acid residue Ca^{2+} binding protein (Babini et al. 2004) which, in comparison with the zinc-binding protein, possesses a very well-defined, mainly α-helical, secondary structure. The distribution of the rmsd$^{\alpha}$ values as a function of the protein conformation number is shown in Fig. 7.

Common features. There are several common features in the analysis of these two metal-binding proteins. The presence of the bound metal, i.e., zinc or calcium, was not taken into account during the computation of the $^{13}C^{\alpha}$ chemical shifts. We observe a significant dispersion of the rms deviation as a function of conformation number (as shown in Figs. 6, 7). For both proteins, the best representative deposited conformer (in terms of computed chemical shifts) reported by Amann et al. (2003) and Babini et al. (2004), is their conformation No. 1, although in our study the best agreement was obtained for conformations Nos. 10 and 12, for the zinc- and calcium-binding proteins, respectively (see Figs. 6, 7). There is notably better agreement between computed and observed $^{13}C^{\alpha}$ chemical shifts for protein 1TTX than for 1M9O, as reported in Table 1.

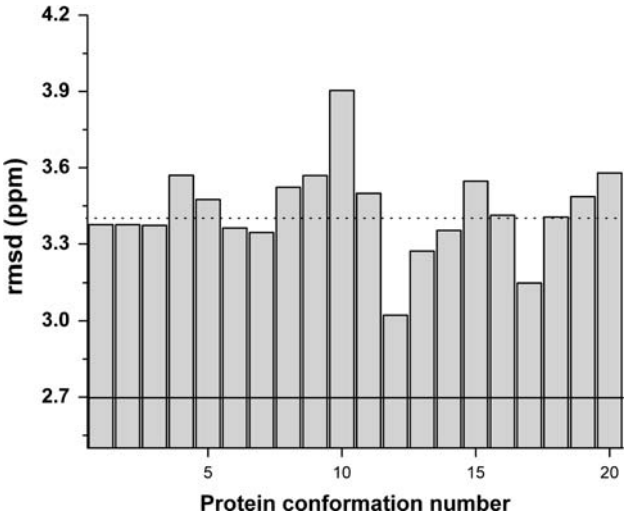

Fig. 7 *Grey filled bars* indicate the rmsd$^{\alpha}$ computed as described in the ''Methods'' section for each of the 20 conformations of a metal-binding protein (PDB code: 1TTX). *Horizontal dotted line* (3.4 ppm) represents the mean-value for the rmsd$^{\alpha}$ over the 20 conformations. The *solid horizontal line* (2.7 ppm) represents the *ca*-rmsd$^{\alpha}$ value computed from 20 conformations of the protein 1TTX by using Eq. 2

Using the *ca*-rmsd$^{\alpha}$ to assess the quality of the NMR-derived conformations

The use of the *ca*-rmsd$^{\alpha}$ value (shown in Table 1 in brackets) as a scoring function to evaluate the quality of

protein conformations has led to the following general conclusions. For the proteins 1XQQ, 1HDN, 1M9O and 1TTX, the results in terms of the *ca*-rmsd$^{\alpha}$ value provide a better, or equal, prediction of the observed $^{13}C^{\alpha}$ chemical shifts than do *any* of the individual conformations (see solid line in Figs. 2, 4, 6, 7). In the case of 1D3Z, a similar conclusion can be derived, except for conformation No. 4, as can be seen from Fig. 3. The $^{13}C^{\alpha}$ chemical shifts of 1XQQ and 1D3Z in solution can be reproduced almost equally well, although the results for the 1XQQ ensemble seem to be slightly better (see values in brackets in Table 1). On the other hand, both the 1XQQ and 1D3Z ensembles are better representations of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the X-ray structure (see the values in brackets in Table 1). In general, the *ca*-rmsd$^{\alpha}$ values are significantly lower when compared with the mean <rmsd$^{\alpha}$> for all proteins.

Since Table 1 contains proteins with different numbers of amino acid residues, in the next two sub-sections we will focus first on the quality of conformations belonging to the same protein, namely ubiquitin, and then we will discuss the relative quality of the conformations belonging to different proteins.

Analysis of the quality of ubiquitin protein conformations

A comparison of the results obtained for the protein ubiquitin shows that there is a larger decrease between the <rmsd$^{\alpha}$> and *ca*-rmsd$^{\alpha}$ values (shown in Table 1) for 1XQQ than for 1D3Z. Our interest centers in understanding the origin of such differences for these two sets of protein conformations. The ensemble of structures of 1D3Z is significantly more compact than that of 1XQQ. This can be seen qualitatively, by a visual inspection of Figs. 1 and S1 (see supplementary material), or quantitatively by comparing the standard deviation (σ) of the error distribution (see ''Analysis of the error distributions of...'' section or Fig. S2a–b). The error distribution derived from 1XQQ (σ = 2.4 ppm) shows a broader dispersion than that from 1D3Z (σ = 1.7 ppm). Although both ensembles satisfied all the experimental NOEs, the backbone residual dipolar coupling, and the side-chain scalar couplings with similar accuracy (Lindorff-Larsen et al. 2005), the ensemble of conformations representing both the native structure and its dynamics, namely 1XQQ, seems to be a slightly better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution, as revealed by the *ca*-rmsd$^{\alpha}$ shown in brackets in column 3 of Table 1, than that derived from a highly refined set of structures, namely 1D3Z, in which the dynamics associated with the backbone and side-chain mobility was not taken into account explicitly (by using information derived from NMR relaxation parameters).

Table 2 Analysis of the $^{13}C^{\alpha}$ error distribution for ubiquitin conformations[a]

PDB code[b]	Parameters for the Gaussian distribution of the errors[c]	
	Mean value[d] (x_o) [ppm]	Standard deviation (σ) [ppm]
1D3Z (760)	0.6 (0.0)	1.7
1UBQ (76)	0.8 (0.1)	1.4
1XQQ (9728)	0.9 (0.3)	2.4

[a] The error between computed and observed $^{13}C^{\alpha}$ chemical shifts, for each residue (μ), was computed as: $\Delta^{\alpha}_{\mu} = (^{13}C^{\alpha}_{observed,\mu} - ^{13}C^{\alpha}_{computed,\mu})$. Then, the frequency of the distribution within a certain interval was binned, and the resulting data were fitted with a Gaussian (or Normal) function (see Fig. S2a–c in supplementary material)

[b] The total number of residues in the ensemble, obtained as the product of the number of residues in the sequence, 76, times the number of conformers that were considered for the error calculations, is shown in parentheses

[c] Parameter obtained from the Gaussian (Normal) distribution of errors, namely the mean (x_o) and the standard deviation (σ), based on the total number of errors between computed and observed $^{13}C^{\alpha}$ chemical shifts

[d] The mean (x_o) value in parentheses was computed assuming a correction factor of 1.25 ppm, rather than 1.82 ppm used here, to convert the TMS-referenced values for the $^{13}C^{\alpha}$ chemical shifts to TSP, as explained in the Results and discussion section ''Analysis of the error distributions of...''

Analysis of the quality among four different sets of proteins

In terms of the *ca*-rmsd$^{\alpha}$-per-residue, computed for each of the protein sets shown in Table 1, the ranking of quality is: 1TTX (0.025) > 1XQQ (0.032) ~ 1D3Z (0.032) ~ 1CM2(0.032) ~ 1UBQ (0.035) >> 1HDN (0.081) > 1M9O (0.10). These results suggest that further refinement of the 1HDN and 1M9O structures is necessary in order to improve the computed *ca*-rmsd$^{\alpha}$-per-residue to the level observed for the other three proteins.

Analysis of the error distributions of ubiquitin protein conformations

An analysis of the error distributions of the computed $^{13}C^{\alpha}$ chemical shifts was carried out for all the conformations of the ubiquitin protein, and is listed in Table 2. The error between computed and observed $^{13}C^{\alpha}$ chemical shifts, for each residue (μ) of each conformation of these proteins, was evaluated as: $\Delta^{\alpha}_{\mu} = (^{13}C^{\alpha}_{observed,\mu} - ^{13}C^{\alpha}_{computed,\mu})$. In all cases, the accumulated error distribution (see Fig. S2a and b for 1D3Z and 1XQQ, respectively, in supplementary material) can be modeled by a Normal (or Gaussian) function with a characteristic mean (x_o) and standard deviation (σ). These values, characterizing the error

distribution of the computed $^{13}C^{\alpha}$ chemical shifts for each of these sets of conformations, are listed in Table 2. All the Normal (or Gaussian) distributions possess a mean value (x_o) close to ~0.7 ppm (which should be compared with the ideal value of $x_o = 0$ ppm). In particular, the σ values listed in Table 2 are similar and within, or show a small departure from, the range of the standard deviation (0.90 ppm $\leq \sigma \leq$ 2.25 ppm) observed by Wang and Jardetzky (2002) for $^{13}C^{\alpha}$ chemical shifts (from a database containing more than 6,000 amino acid residues in α-helix, β-sheet and statistical-coil conformations). An important source of the computed errors from the ubiquitin conformations came from residues exhibiting high mobility, such as those that pertain to loops or highly flexible portions of the molecule, for example, Gly-76 (Δ^{α} = 10.4 ppm, conformation 6) and Asn-60 (Δ^{α} = 9.6 ppm, conformation 2) from 1D3Z; Met-1 (Δ^{α} = 12.8 ppm and 11.6 ppm, conformations 82 and 99, respectively) from 1XQQ, etc.

The existence of large errors ($\Delta^{\alpha} \geq 10$ ppm) for some residues of proteins 1D3Z and 1XQQ (see Fig. S2a and b in the supplementary material) deserves some additional consideration. The origin of such large errors may lie in either the method or in deficiencies in the structure set. However, these are not likely sources because the method was already tested in the ''Validation of the methodology: test on ubiquitin'' section, and all the structures were determined at a high level of resolution. Even more important than the existence of such large errors is their frequency because it will critically affect the computation of the *ca*-rmsd$^{\alpha}$. As can be seen from Fig. S2a and b (supplementary material), such frequencies for 1XQQ are extremely small because of the Gaussian nature of the error distribution. In other words, about 99.7% of the errors lie within 3σ, i.e., with a $\Delta^{\alpha} \leq$ 7.0 ppm).

It is well known that use of different methods and standards for chemical-shift referencing could be an important source of errors (Wishart et al. 1995; Iwadate et al. 1999; Cornilescu et al. 1999). For this reason, it is common practice to apply corrections to the published shifts to bring the data of several proteins into close agreement (Iwadate et al. 1999). More important, protein structures obtained by NMR based on the same standards, such as TSP or TMS or DSS, may require different corrections, in some cases greater than 1 ppm (Iwadate et al. 1999). We avoid including such adjustments because they add empiricism to the predictive method. However, the reader should be aware that such corrections may have significant influence on some parameters that are reported here. Thus, for example, if we use an empirical correction factor of 1.25 ppm, rather than the 1.82 ppm suggested by Wishart et al. (1995), and used here, to convert the computed TMS-referenced values for the $^{13}C^{\alpha}$ chemical shifts to TSP, the systematic shift (~0.7 ppm) observed in the mean value, x_o, for 1D3Z, 1UBQ and 1XQQ will be significantly lower, i.e., the recomputed mean x_o value will be zero (see Fig. S2c in supplementary material), or close to zero, as shown in parentheses in column 2 of Table 2. As a consequence, the recomputed *ca*-rmsd$^{\alpha}$ (values in parenthesis in column 3 of Table 1) will also be lower, i.e., by ~0.2 ppm, than the values reported in brackets in column 3 of Table 1, for 1D3Z, 1UBQ and 1XQQ. Although further calibration, using a larger data set of proteins, would alleviate such problems, it is not computationally feasible at the moment and, hence, a discussion of such effect for the remaining three proteins in Table 1 is beyond the scope of this work.

Distribution of the agreements for the $^{13}C^{\alpha}$ chemical shifts in 1XQQ

A complementary analysis to that of section ''Analysis of the error distributions of...'' is the percentage of agreement between observed and computed $^{13}C^{\alpha}$ chemical shifts, for each amino acid residue in the sequence. Here, agreement means that the computed error (Δ^{α}) between observed and computed $^{13}C^{\alpha}$ chemical shifts is lower than a certain cutoff (2.25 ppm). This selected cutoff corresponds to the highest standard deviation (σ) observed for $^{13}C^{\alpha}$ chemical shifts for α-helix, β-sheet and statistical-coil by Wang and Jardetzky (2002). Among all the protein conformations for ubiquitin (1UBQ, 1D3Z and 1XQQ), we chose to illustrate the analysis on the 128 conformations of ubiquitin from 1XQQ. Thus, the results obtained from 9,728 residues (see Table 2), enabled us to draw two conclusions.

First, the distribution of the agreements does not depend on the amino acid residue type. For example, two alanines in ubiquitin, Ala-28 and Ala-46, have 72% and 25% of agreement, respectively; two out of six glutamines, Gln-31 and Gln-41, have agreements of 83% and 32%, respectively, etc. The fact that some amino acids in regular secondary-structure regions, with little, or no, influence of the side-chain conformations, such as Ala-28 in the α-helical region of the molecule, do not show 100% agreement between observed and computed $^{13}C^{\alpha}$ chemical shifts within a cutoff of 2.25 ppm, deserves some additional consideration. Because of the methodology used here for the calculation of the $^{13}C^{\alpha}$ chemical shifts, the possible existence of small steric clashes or other artifacts related to the calculations other than structural differences among conformations, must be ruled out. For the residue Ala-28, analysis of the dispersion of the backbone (ϕ, ψ) angles [among the 128 conformations of 1XQQ] showed that 41% and 57% of the ϕ and ψ angles, respectively, are, within a cutoff of 10°, in agreement with the canonical α-helical values (–60.0°; –40.0°). A 100% agreement is attained only if the adopted cutoff is 30°. The criterion of 10° was adopted because

most of the α-helices are well represented by canonical values of the dihedral angles, i.e., with $\phi = -62° \pm 8°$ and $\psi = -42° \pm 10°$ (Spera and Bax, 1991), although some exceptions appear for solvent-exposed α-helices which are often bent. The rmsd between observed and predicted $^{13}C^{\alpha}$ chemical shifts for Ala-28, computed from the 128 conformations, is 1.6 ppm and the difference between the maximum and minimum $^{13}C^{\alpha}$ chemical shifts (the range of the predictions) computed among the 128 conformations for Ala-28 is 4.23 ppm. As an additional test on Ala-28, the same analysis was carried out for the 10 conformations of 1D3Z. Significantly lower values were found: (i) for the dispersion of the backbone angles, 100% of the backbone (ϕ, ψ) angles are within a 10° cutoff with respect to the canonical α-helical values; and (ii) for the difference between the maximum and minimum $^{13}C^{\alpha}$ chemical shifts (the range of the predictions), a value of 0.47 ppm, was found. Consistently, a markedly lower rmsd (1.0 ppm) was also obtained. Moreover, 100% of the Ala-28 residues in 1D3Z show an error (Δ^{α}), between observed and computed $^{13}C^{\alpha}$ chemical shifts, lower than the standard deviation ($\sigma = 1.7$ ppm) computed for this set of conformations. This analysis allows us to conclude that the broad range of variations of the backbone (ϕ, ψ) angles of Ala-28, in the set of 128 conformations of 1XQQ, is the origin of the observed dispersion in the distribution of the agreements for this residue.

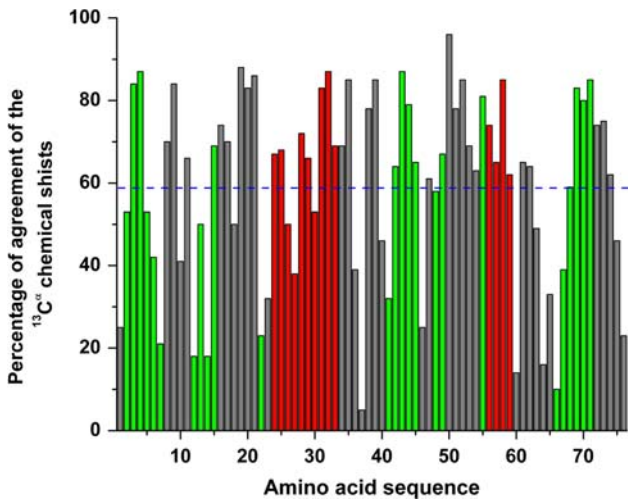

Fig. 8 *Bars* indicates the distribution of the percentage of agreement (for each residue of 1XQQ in the whole ensemble) between computed and observed $^{13}C^{\alpha}$ chemical shifts, within a tolerance of 2.25 ppm, as described in ''Results and discussion'' section ''Distribution of the agreements for the $^{13}C^{\alpha}$ chemical shifts in 1XQQ''. The α-helix portions (from residues 24–33 and 56–59) is shown in red, the extended-strands (from residues 2–7, 12–15, 22, 41–45, 48–49, 55, 66–71) are indicated in *green*, while the rest of the residues are in *light-grey*. The *dashed horizontal line* indicates the average percentage of agreement (~59%) computed over all the residues in the sequence

Second, there is no clear preference, in terms of the agreement, for a particular location of the residue in the sequence, i.e. for α-helix, extended strands or loop regions (as seen in Fig. 8). For example, Val-26 is in an α-helix portion while Val-70 belongs to an extended-strand, and they exhibit agreements of 50% and 80%, respectively. Among all the threonines, Thr-9 pertains to a loop region, whereas, Thr-12, Thr-14, Thr-55 and Thr-66 to extended-strands with very dissimilar percentages of agreement, namely, 84%, 18%, 18%, 81% and 10%, respectively.

Analysis of the average agreement within a given secondary structure element, i.e., without considering the identity of the residues, indicated that there is no significant difference in the average values obtained for the α-helix regions (~67%) or for the extended strand regions (~59%, i.e., as an average over the 7 extended strands in the sequence). Moreover, none of these average values deviate significantly from the one computed for the whole sequence (~59%) (see Fig. 8). These results indicate that the agreement between computed and observed $^{13}C^{\alpha}$ chemical shifts does not show significant dependence on either the identity of the amino acid residue or the location of the residue in the sequence.

Additional evidence for the presence of dynamics: the theoretical minimal-rmsd model for 1XQQ

Examination of the chemical shifts of all the amino acids in the 128 conformations of 1XQQ enable us to identify the amino acid at each position in the sequence whose computed chemical shifts most closely match the observed ones, among all these conformations. This identified set of individual amino acid conformations corresponds to only *one* conformation of the whole chain: the 'theoretical minimal-rmsd model'. Such conformations possess an rmsd equal to 0.28 ppm which is drastically lower, i.e., about 10 times, than any rmsd of any ubiquitin conformation from which it was derived. However, such a conformation is one with multiple atomic clashes, i.e., with a non-bonded energy greater than 10^{12} kcal/mol according to the ECEPP/3 force-field, and with an all heavy-atom rmsd greater than 10 Å when compared with any conformation solved by NMR or X-ray, such as 1XQQ, 1D3Z or 1UBQ. This example illustrates that a single conformation as an 'optimal representation' of the observed $^{13}C^{\alpha}$ chemical shifts in solution does not exist for ubiquitin, and it might not exist for any protein. This constitutes evidence, additional to that showing that the *ca*-rmsd is lower than the <rmsd>, for the presence of dynamics for the structures in solution. In other words, the observed $^{13}C^{\alpha}$ chemical shifts in solution cannot be represented by a single conformation.

This analysis also points out that a necessary condition, to use the $^{13}C^{\alpha}$ chemical shifts for a validation test, is that

the conformation or ensemble of conformations must satisfy *all* the other NMR constraints, as was mentioned in section ''A new scoring function: the conformationally averaged rmsd$^{\alpha}$.'' In other words, validation of a conformation or set of conformations based *only* on the agreement between computed and observed $^{13}C^{\alpha}$ chemical shifts is not sufficient, preventing their use, if additional NMR data such as NOEs are not provided, for protein structure determination (Vila et al. 2007).

Validation of structures: a comparison with results of automated prediction of $^{13}C^{\alpha}$ chemical shifts

Our interest focuses on whether the results from a physics-based method, such as those obtained with the current methodology, are more accurate than those obtained from automated servers, such as SHIFTX (Neal et al. 2003) and SHIFTS (Xu and Case 2001, 2002), for validation of protein conformations. As a test of the sensitivity of the different methods to significant differences in protein structure, we considered conformation 1 of two proteins, i.e., 1D3Z (shown in Fig. 1) and 1M9O (shown in Fig. 5). These two proteins were chosen because they possess large differences in stereochemistry quality. For example, according to the PROCHECK program (Laskowski et al. 1996), ~2% and ~70% of the non-proline non-glycine residues in 1D3Z and 1M9O, respectively, are outside of the most favored regions of the Ramachandran map. The average number of abnormally short interatomic distances per residue, according to the WHAT_IF program (Vriend 1990), is considerably different in 1D3Z and 1M9O (~0.08 and ~1.1, respectively). The correlation coefficient R, or *Pearson* coefficient (Press et al. 1992) between observed $^{13}C^{\alpha}$ chemical shifts and those computed with SHIFTX, SHIFTS and with our methodology for conformation 1 of 1D3Z is: R = 0.98, 0.98 and 0.90, respectively, while the corresponding values for conformation 1 of 1M9O are: R = 0.94, 0.89 and 0.62, respectively. If only the correlation coefficient, R, was used to assess the quality of these two protein conformations, we would conclude that, in the case of SHIFTX, both conformations are of comparable quality; in the case of SHIFTS, one conformation is of better quality and, from our calculations, one conformation shows significantly better quality, in line with its greater stereochemical quality. The SHIFTS predictions can discriminate between structures better than those of SHIFTX, but not as well as the current methodology. This conclusion should not be surprising because SHIFTS makes use of a database of peptide chemical shifts computed at the DFT level (Xu and Case 2001). On the other hand, SHIFTX (Neal et al. 2003) is a hybrid predictive method, which combines an empirical hypersurface approach to calculate chemical shifts from atomic coordinates. In other words, our analysis of 1M9O and 1D3Z conformations indicated that the results from these servers are less sensitive than the current methodology to validate conformations and, therefore, may not be able to provide enough guidance in selecting the most accurate protein structures.

Conclusions

The methodology tested here on a set of 139 conformations of ubiquitin led to results that show better agreement, in terms of *ca*-rmsd$^{\alpha}$, than those from the analysis of Sun et al. (2002). Conceivably, the good agreement in the prediction of the $^{13}C^{\alpha}$ chemical shifts in terms of the *ca*-rmsd$^{\alpha}$ value for both the 1D3Z and the 1XQQ ensembles is a consequence of the fact that it does not seem to be crucial, for accounting for prediction of ^{13}C chemical shifts, as to whether or not the residues belong to α-helix, extended structure, or loop regions of the molecule.

Some additional conclusions regarding the comparison between computed and observed $^{13}C^{\alpha}$ chemical shifts discussed here follow. First, the X-ray-derived structure may be (see Fig. 4) or may not be (see Figs. 2, 3) a better structure than the NMR-derived solution structures with which to compute (represent) the observed $^{13}C^{\alpha}$ chemical shifts. In terms of the *ca*-rmsd$^{\alpha}$ value, the NMR-derived structures of ubiquitin from both 1D3Z (see Fig. 3) and 1XQQ (see Fig. 2) are better representations of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the X-ray structure. Second, there is a random correlation, R, between the rmsd of a single structure and the ranking of the NMR-solved structures deposited in the PDB for all proteins, except 1HDN (see ranking distribution in Figs. 2–4, 6, 7). The computed correlations are: R = 0.02; 0.12; 0.43; 0.14; and –0.04, for 1XQQ, 1D3Z, 1HDN, 1M90 and 1TTX, respectively. As a consequence, if there is a need for selection of a *single* structure from an ensemble of conformers, e.g., as frequently may occur for structure predictions in drug or protein design, analysis of protein–protein interactions, or antibody specificity, etc., some standard criterion should be adopted. We propose here that a fast and straightforward selection of a '*reliable single-model structure*' can be carried out (a) by using the approach presented here to predict the $^{13}C^{\alpha}$ chemical shifts for each of the available models, and (b) select the best-ranked structure based on the *quality* of the models as determined by their rmsd$^{\alpha}$. Furthermore, use of the *ca*-rmsd$^{\alpha}$ value has led to closer agreement with the observed $^{13}C^{\alpha}$ chemical shifts in solution than when individual, or the mean, rmsd$^{\alpha}$ was used. In other words, proteins in solution are conformationally labile, as indicated by both the *ca*-rmsd and the theoretical minimal-rmsd model analyses, and this must be taken into account (as we do here by averaging over all conformations of the set) to

predict the $^{13}C^{\alpha}$ chemical shifts most accurately. This result should not be surprising because we know from Zhao and Jardetzky (1994) that ''*...the ''true'' solution structure is in reality an ensemble of structures...*''

Notably, among all conformations analyzed in this work in terms of the *ca*-rmsd$^{\alpha}$-per-residue, the set of 20 conformations of Protein 1TTX (Babini, et al. 2004) is the best representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution.

Although quantum mechanical methodologies such as the one proposed here are much more CPU demanding than the automatic servers, they are extremely sensitive to the coordinates of all atoms, and hence very useful for judging the quality of different conformations for a given protein. In particular, our analysis has demostrated that the current methodology is more accurate than those utilized in the automated servers for the purpose of protein structure validation, which is a necessary, but not sufficient, condition for the process of protein structure generation, evaluation and refinement. Certainly, it does not minimize the importance and the capabilities of the predictions of servers such us SHIFTX and SHIFTS for many other applications, such as chemical shift assignment, chemical shift validation, reference checking, etc. (Neal et al. 2003). Even more important, the information obtained with the current methodology can be incorporated into any of the existing automatic servers to improve their sensitivity.

Further application of the current methodology includes protein structure determination by a combined use of NOEs and torsional constraints for both the backbone and side chain, derived from the $^{13}C^{\alpha}$ chemical shifts computed at the DFT level. This is feasible because the current methodology enables us to identify and select a set of backbone and side-chain torsional angles for all amino acid residues in the sequence, based on the lowest error between computed and observed chemical shifts (Vila et al. 2007).

Acknowledgments We thank B.T. Amann for providing us with the reference used for the ^{13}C chemical shifts of protein 1M9O, and Yelena Arnautova for helpful suggestions. This research was supported by grants from the National Institutes of Health (GM-14312, TW-6335, and GM-24893), and the National Science Foundation (MCB05-41633). Support was also received from the National Research Council of Argentina (CONICET), FONCyT-ANPCyT (PAE 22642 / 22672), and from the Universidad Nacional de San Luis [UNSL] (P-328501), Argentina. This research was conducted using the resources of: (1) two Beowulf-type clusters located at (a) the Instituto de Matemática Aplicada San Luis (CONICET-UNSL); and (b) the Baker Laboratory of Chemistry and Chemical Biology, Cornell University; and (2) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center.

References

Allerhand A, Childers RF, Oldfield E (1973) Natural-abundance carbon-13 nuclear magnetic resonance studies in 20-mm sample tubes. Observation of numerous single-carbon resonances of Hen Egg-White Lysozyme. Biochem 12:1335–1241

Amann BT, Worthington MT, Berg JMA (2003) A Cys_3His zinc-binding domain from Nup475/Tristetraprolin: a novel fold with a disklike structure. Biochem 42:217–221

Babini E, Bertini I, Capozzi F, Del Bianco C, Hollender D, Kiss T, Luchinat C, Quattrone A (2004) Solution structure of human β-parvalbumin and structural comparison with its paralog α-parvalbumin and with their rat orthologs. Biochem 43:16076–16085

Ban Y-E, Rudolph J, Zhou P, Edelsbrunner H (2006) Evaluating the quality of NMR structures by local density of protons. Proteins 62:852–864

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

Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu)

Case DA (2000) Interpretation of chemical shifts and coupling constants in macromolecules. Curr Opin Struct Biol 10:197–203

Case DA, Dyson HJ, Wright PE (1994) Use of chemical shifts and coupling constant in nuclear magnetic resonance structural studies on peptides and proteins. Methods Enzymol 239:392–416

Celda B, Biamonti C, Arnau MJ, Tejero R, Montelione GT (1995) Combined use of ^{13}C chemical shift and $^{1}H^{\alpha}$–$^{13}C^{\alpha}$ heteronuclear NOE data in monitoring a protein NMR structure refinement. J Biomol NMR 5:161–172

Chakrabarti P, Pal D (1998) Main-chain conformational features at different conformations of the side-chains in proteins. Protein Eng 11:631–647

Chesnut DB, Moore KD (1989) Locally dense basis sets for chemical shift calculations. J Comp Chem 10:648–659

Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a diluite liquid crystalline phase. J Am Chem Soc 120:6836–6837

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

de Dios AC, Pearson JG, Oldfield E (1993a) Chemical shifts in proteins: ab initio study of carbon-13 nuclear magnetic resonance chemical shielding in glycine, alanine and valine residues. J Am Chem Soc 115:9768–9773

de Dios AC, Pearson JG, Oldfield E (1993b) Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. Science 260:1491–1496

Doreleijers JF, Rullmann JAC, Kaptein R (1998) Quality assessment of NMR structures: a statistical survey. J Mol Biol 281:149–164

Dunbrack RL Jr, Karplus M (1994) Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1:334–340

Dyson HJ, Wright PE (2005) Elucidation of the protein folding landscape by NMR. Methods Enzymol 394:299–321

Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE Jr, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Andres JL, Gonzalez C, Head-Gordon M, Replogle ES, Pople JA (1998) Gaussian 98. Revision A.7, Inc., Pittsburgh, PA

Havlin RH, Le H, Laws DD, de Dios AC, Oldfield E (1997) An ab initio quantum chemical investigation of carbon-13 NMR shielding tensors in glycine, alanine, valine, isoleucine, serine, and threonine: comparisons between helical and sheet tensors, and effects of χ_1 on shielding. J Am Chem Soc 119:11951–11958

Hehre WJ, Radom L, Schleyer P, Pople JA (1986) Ab initio molecular orbital theory. Wiley, New York

Howard OW, Lilley DMJ (1978) Carbon-13-NMR of peptides and proteins. Prog Nucl Magn Reson Spectrosc 12:1–40

Hunter C, Packer MJ, Zonta C (2005) From structure to chemical shift and vice-versa. Prog Nucl Magn Reson Spectrosc 47:27–39

Iwadate M, Asakura T, Williamson MP (1999) $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ carbon-13 chemical shifs in protein from an empirical database. J Biomol NMR 13:199–211

Jameson CJ (1996) Understanding NMR chemical shifts. Annu Rev Phys Chem 47:135–169

Kuszewski J, Qin JA, Gronenborn AM, Clore GM (1995) The impact on direct refinement against $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts on protein structure determination by NMR. J Magn Reson Ser B 106:92–96

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283–291

Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton J (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

Laws DD, Le H, de Dios AC, Havlin RH, Oldfield E (1995) A basis size dependence study of Carbon-13 nuclear magnetic resonance spectroscopic shielding in Alanyl and Valyl fragments: toward protein shielding hypersurfaces. J Am Chem Soc 117:9542–9546

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

Luginbühl P, Szyperski T, Wüthrich KJ (1995) Statistical basis for the use of $^{13}C^{\alpha}$ chemical shifts in protein structure determination. Magn Resn B 109:220–233

Malthouse JPG (1985) ^{13}C NMR of enzymes. Prog Nucl Magn Reson Spectrosc 18:1–59

Meiler JJ (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26:25–37

Melnik BS, Garbuzynskiy SO, Lobanov MYu, Galzitskaya OV (2005) The difference between protein structures obtained by X-ray analysis and nuclear magnetic resonance. J Mol Biol 39:113–122

Moon S, Case DA (2006) A comparison of quantum chemical models for calculating NMR shielding parameters in peptides: mixed basis set and ONION methods combined with a complete basis set extrapolation. J Comp Chem 27:825–836

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. Proteins 12:345–364

Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AMJJ, Vuister GW, Vriend G, Spronk CAEM (2004) DRESS: a database of Refined solution NMR structures. Proteins 55:483–486

Napper S, Delbaere LTJ, Waygood BEJ (1999) Histidine-containing protein, HPr, of the *Escherichia coli* phosphoenolpyruvate:sugar phosphotransferase system can accept and donate a phosphoryl group. J Biol Chem 274:21776–21782

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. J Biomol NMR 26:215–240

Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J Phys Chem 96:6472–6484

Oldfield E (2002) Chemical shifts in amino acids, peptides and proteins: from quantum chemistry to drug design. Annu Rev Phys Chem 53:349–378

Oldfield E, Allerhand A (1975) Identification of tryptophan resonances in natural abundance C-13 nuclear magnetic-resonance spectra of protein. Application of partially relaxed fourier-transform spectroscopy. J Am Chem Soc 97:221–224

Pearson JG, Le H, Sanders LK, Godbout N, Havlin RH, Oldfield EJ (1997) Predicting chemical shifts in proteins: structure refinement of valine residues by using ab initio and empirical geometry optimizations. J Am Chem Soc 119:11941–11950

Pearson JG, Wang J-F, Markley JL, Le H, Oldfield E (1995) Protein structure refinement using carbon-13 nuclear magnetic resonance spectroscopic chemical shifts and quantum chemistry. J Am Chem Soc 117:8823–8829

Pontius J, Richelle J, Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264:121–136

Press HW, Teukolsky SA, Vetterling WT, Flannery BP (1992) In: Numerical recipes in Fortran 77. The art of scientific computing, 2nd edn. Cambridge University Press, Ch. 14, pp 630–633

Ripoll DR, Vorobjev YN, Liwo A, Vila JA, Scheraga HA (1996) Coupling between folding and ionization equilibria: effects of pH on the conformational preferences of polypeptides. J Mol Biol 264:770–783

Ripoll DR, Vila JA, Scheraga HA (2005) On the Orientation of the Backbone Dipoles in Native Folds. Proc Natl Acad Sci USA 102:7559–7564

Ripoll DR, Vila JA, Scheraga HA (2004) Folding of the Villin headpiece subdomain from random structures. Análisis of the charge distribution as function of pH. J Mol Biol 339:915–925

Simon K, Xu J, Kim C, Skrynnikov NR (2005) Estimating the accuracy of protein structures using residual dipolar couplings. J Biomol NMR 33:83–93

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and Cα and Cβ ^{13}C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

Schubert M, Laudde D, Oschkinat H, Schmieder P (2002) A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on ^{13}C chemical shift statistic. J Biomol NMR 24:149–154

Sun H, Sanders LK, Oldfield E (2002) Carbon-13 NMR shielding in the twenty common amino acids: comparisons with experimental results in proteins. J Am Chem Soc 124:5486–5495

van Nuland NAJ, Hangyi IW, van Schaik RC, Berendsen HJC, van Gunsteren WF, Scheek RM, Robillard GT (1994) The high-resolution structure of the histidine-containing phosphocarrier protein HPr from *Escherichia coli* determined by restrained molecular dynamics from nuclear magnetic resonance nuclear Overhauser effect data. J Mol Biol 237:544–559

Vijay-Kumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 194:531–544

Vila JA, Baldoni HA, Ripoll DR, Scheraga HA (2003) Unblocked statistical-coil tetrapeptides in aqueous solution: quantum-chemical computation of the carbon-13 NMR chemical shifts. J Biomol NMR 26:113–130

Vila JA, Baldoni HA, Ripoll DR, Ghosh A, Scheraga HA (2004a) Polyproline II helix conformation in a proline-rich enviroment: a theoretical study. Biophys J 86:731–742

Vila JA, Baldoni HA, Ripoll DR, Scheraga HA (2004b) Fast and accurate computation of the ^{13}C chemical shifts for an alanine-rich peptide. Proteins 57:87–98

Vila JA, Ripoll DR, Baldoni HA, Scheraga HA (2002) Unblocked statistical-coil tetrapeptides and pentapeptides in aqueous solution: a theoretical study. J Biomol NMR 24:245–262

Vila JA, Ripoll DR, Scheraga HA (2007) Use of $^{13}C^{\alpha}$ chemical shifts in protein structure determination. J Phys Chem B (in press)

Villegas ME, Vila JA, Scheraga HA (2007) Effects of side-chain orientation on the ^{13}C chemical shifts of antiparallel β-sheet model peptides. J Biomol NMR 37:137–146

Vriend GJ (1990) A molecular modeling and drug design. Mol Graph 8:52–56

Vriend G, Sander C (1993) Quality control of protein models: directional atomic contact analysis. J Appl Crystallogr 26:47–60

Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861

Wilson KS, Dauter Z, Lamzin VS, Walsh M, Wodak S, Richelle J, Pontius J, Vaguine A, Laskowski JM, MacArthur MW, Dodson E, Murshudov G, Oldfield TJ, Kaptein R, Rullmann JAC (1998) Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J Mol Biol 276:417–436

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3–34

Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD (1995) ^{1}H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. J Biomol NMR 6:135–140

Wishart DS, Nip AM (1998) Protein chemical shift analysis: a practical guide. Biochem Cell Biol 76:153–163

Xu X-P, Case DAJ (2001) Automatic prediction of ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333

Xu X-P, Case DA (2002) Probing multiple effects on ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts in peptides using density functional theory. Biopolymers 65:408–423

Zhao D, Jardetzky O (1994) An assessment of the precision and accuracy of protein structures determined by NMR. Dependence on distance erros. J Mol Biol 239:601–607