ACCOUNTS
of chemical research

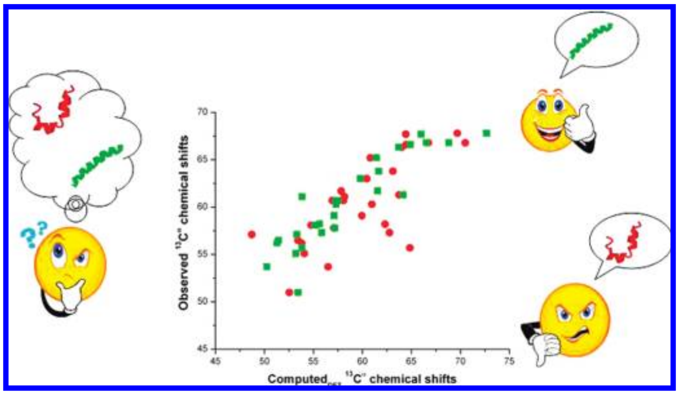
Assessing the Accuracy of Protein Structures by Quantum Mechanical Computations of $^{13}C^{\alpha}$ Chemical Shifts

JORGE A. VILA[†,‡] AND HAROLD A. SCHERAGA[*,†]

[†]Baker Laboratory of Chemistry and Chemical Biology, Cornell University,
Ithaca, New York 14853-1301, [‡]Universidad Nacional de San Luis,
IMASL-CONICET, Ejército de Los Andes, 950-5700 San Luis, Argentina

RECEIVED ON FEBRUARY 28, 2009

CONSPECTUS



Two major techniques have been used to determine the three-dimensional structures of proteins: X-ray diffraction and NMR spectroscopy. In particular, the validation of NMR-derived protein structures is one of the most challenging problems in NMR spectroscopy. Therefore, researchers have proposed a plethora of methods to determine the accuracy and reliability of protein structures. Despite these proposals, there is a growing need for more sophisticated, physics-based structure validation methods. This approach will enable us to (a) characterize the "quality" of the NMR-derived ensemble as a whole by a single parameter, (b) unambiguously identify flaws in the sequence at a residue level, and (c) provide precise information, such as sets of backbone and side-chain torsional angles, that we can use to detect local flaws.

Rather than reviewing all of the existing validation methods, this Account describes the contributions of our research group toward a solution of the long-standing problem of both global and local structure validation of NMR-derived protein structures. We emphasize a recently introduced physics-based methodology that makes use of observed and computed $^{13}C^{\alpha}$ chemical shifts (at the density functional theory (DFT) level of theory) for an accurate validation of protein structures in solution and in crystals. By assessing the ability of computed $^{13}C^{\alpha}$ chemical shifts to reproduce observed $^{13}C^{\alpha}$ chemical shifts of a single structure or ensemble of structures in solution and in crystals, we accomplish a global validation by using the *conformationally averaged* root-mean-square deviation, *ca*-rmsd, as a scoring function. In addition, the method enables us to provide local validation by identifying a set of individual amino acid conformations for which the computed and observed $^{13}C^{\alpha}$ chemical shifts do not agree within a certain error range and may represent a nonreliable fold of the protein model.

Although it is computationally intensive, our validation method has several advantages, which we illustrate through a series of applications. This method makes use of the $^{13}C^{\alpha}$ chemical shifts, not shielding, that are ubiquitous to proteins and can be computed precisely from the φ, ψ, and χ torsional angles. There is no need for *a priori* knowledge of the oligomeric state of the protein, and no knowledge-based information or additional NMR data are required. The primary limitation at this point is the computational cost of such calculations. However, we anticipate that enhancements in the speed of calculating these chemical shifts coupled with the ever-increasing computational power should soon make this a standard method accessible to the general NMR community.

Published on the Web 07/02/2009 www.pubs.acs.org/acr
10.1021/ar900068s CCC: $71.50

1. Introduction

Between the first time that chemical shifts were observed by Arnold et al., in 1951,[1] and the "structural genomics" initiative (that started in 2000) to develop a technology for high-throughput structure determination and expand our understanding of protein structure and function, a vast amount of experimental and theoretical advances in the nuclear magnetic resonance (NMR) field have taken place. Many recent reviews in the field attest to this.[2−8] Despite this formidable progress in NMR spectroscopy, quality assessment remains as a crucial test for NMR-derived protein structures.[9,10] A number of methods have been developed over the years (WHAT IF,[11] PROCHECK,[12] RPF,[13] MolProbity,[14] etc.) because validation of protein structure conformations is essential for both the spectroscopists, since it enables them to focus on aspects of the structure that might contain errors, and the users, because validation of existing models enables them to determine the quality and suitability of the protein models for any specific purpose. Despite the available tools for assessing the accuracy of NMR-derived proteins, there is consensus that more sophisticated structural validation methods are needed;[15,16] i.e., there is a need for a very sensitive, physics-based method to detect whether or not a given structure or region of the structure, at a residue level, is erroneous.

Residual dipolar couplings (RDC) represent a powerful tool with which to identify errors in NMR structures.[6,15−18] Regrettably, as noted by Nabuurs et al.,[16] they are not routinely acquired in most of the structural genomics efforts nor are they available for the great majority of the deposited structures in the BioMagResBank;[19] viz., as noted by Simon et al.,[18] there are 116 RDC data sets compared with 2276 nuclear Overhauser effect (NOE) data sets associated with proteins. On the other hand, most chemical shifts are available from any NMR experiment because the first step in NMR spectroscopy, before the collection and analysis of structural restraints such as those derived from NOE, consists of the acquisition of NMR data that lead to the assignment of the chemical shifts for all nuclei (^{1}H, ^{15}N, and ^{13}C) in the molecule. Among all these nuclei, we focus our attention on only one, namely, $^{13}C^{\alpha}$ chemical shifts, because they are exquisitely sensitive to their local environment and provide a conformational "fingerprint" of each amino acid residue in a protein. The backbone and side-chain conformations of a residue are influenced by interactions with the rest of the protein, but once these conformations are established by these interactions, the $^{13}C^{\alpha}$ shielding of each residue, but not the shielding from other nuclei such as ^{1}H, ^{15}N, or $^{13}C'$, depends mainly on its own backbone[20,21] and side-chain[22−24] conformation, with no significant influence of either the amino acid sequence[21,24,25] or the position of the given residue in the sequence,[26] or the oligomerization state of the protein. These properties, together with the facts that this nucleus is ubiquitous in proteins and that the computation of the $^{13}C^{\alpha}$ shielding at the quantum chemical level of theory can be carried out with coarse-grained parallelization (one residue per processor), make this nucleus an attractive candidate in order to validate protein structures.[26−29]

In this Account, we report our efforts to develop a purely physics-based, structure validation method that enables us (a) to characterize the "quality" of the NMR-derived ensemble, as a whole, by a single parameter; and (b) to unambiguously identify flaws in the sequence, at a residue level.

In section 2, we will focus on the factors affecting the computation, at the quantum mechanical level of theory, of the $^{13}C^{\alpha}$ chemical shifts in proteins, such as the sensitivity of the shielding/deshielding of $^{13}C^{\alpha}$ nuclei to changes in the protonation/deprotonation state of distant ionizable groups,[30] the values of the bond lengths and bond angles adopted to represent the geometry of the molecule,[4] etc. In addition, we will demonstrate that the validation method is strong, rather than weak.[31] Finally, given that our central interest is in the $^{13}C^{\alpha}$ chemical shifts, not their shielding, computation of an accurate value for the shielding of the reference, namely, tetramethylsilane (TMS), is crucial. For this reason, we illustrate how an *effective* shielding[32] value can be computed.

In section 3, we will show, first, how our method can be used, in terms of a new scoring function called the *ca*-rmsd (conformationally averaged root-mean-square deviation), for validation of two highly accurate protein structures solved by both NMR and X-ray methods and, second, how a comparison between observed and computed $^{13}C^{\alpha}$ chemical shifts (at the DFT level of theory) enables us to determine, at the residue level, the existence of local flaws in the sequence. The latter is a very important problem because it is known that ambiguities in the assignment of intra- or intersubunit nuclear Overhauser effects (NOEs) might result in a wrong fold.[16,33] The increasing number of oligomer structures in biology (~65% of the proteins in every genome are expected to be homo-oligomers[33]) may only exacerbate this problem. Hence, the ability of our validation method to detect flaws in the sequence might be of very valuable assistance for determining wrong folds in homodimers, particularly if information regarding the oligomerization state in solution or the structure of homologous monomers is not available.

Finally, in section 4, a discussion of ongoing progress in our validation method, to speed it up without loss of accuracy, and its impact on progress in the field, is presented, together with a limit of the method that exists in current applications.

2. General Background

The foundation of the method and the most relevant approximations adopted to make the computation of the $^{13}C^{\alpha}$ chemical shifts accurate, but feasible, are discussed briefly here.

Computational Approach. At the core of the $^{13}C^{\alpha}$-based validation method are the following most important approximations adopted to compute chemical shifts. First, all the experimentally determined conformations to be validated were *regularized*, i.e., all residues were replaced by the standard ECEPP/3[34] residue geometry, in which bond lengths and bond angles are fixed (rigid-geometry approximation), and second, hydrogen atoms were added, if necessary. This problem is central to all the results reported here because it is known that quantum mechanical calculations are very sensitive to bond lengths and bond angles.[35] In fact, even proteins solved at a high level of accuracy, as by X-ray diffraction, are not expected to provide the best correlation with the observed $^{13}C^{\alpha}$ chemical shifts.[35] Consequently, we explore the dependence of the $^{13}C^{\alpha}$-chemical shift calculations, rather than shielding, on the bond lengths and bond angles.

For this test, we chose the structure of ubiquitin deposited in the Protein Data Bank[36] (PDB) [PDB id 1UBQ]; it possesses nonregular geometry and has been solved by X-ray diffraction at 1.8 Å resolution.[37] We also examined the corresponding structure with regularized geometry, named here as 1UBQ$_{reg}$. Analysis of the differences between computed and observed $^{13}C^{\alpha}$ chemical shifts for the 1UBQ and 1UBQ$_{reg}$ structures, in terms of rmsd, leads to 3.28 and 2.38 ppm, respectively. The value obtained for 1UBQ$_{reg}$ (2.38 ppm) is slightly lower than the previously reported value (2.60)[32] because of improvement in the regularization procedure. Further analysis of the agreement of these structures with the deposited electron density data[37] of 1UBQ, in terms of the R-factor, leads to 19.2% and 23.1% for 1UBQ and 1UBQ$_{reg}$, respectively; the all-heavy-atom rmsd between these two structures is 0.142 Å. The better agreement obtained with 1UBQ$_{reg}$, rather than 1UBQ, in terms of observed and computed $^{13}C^{\alpha}$ chemical shifts, is consistent with the long-time recognition that the bond lengths and bond angles of both X-ray and NMR-derived structures are not as highly accurately defined as in studies of small molecules,[35] with which the ECEPP/3 geometry has been parametrized.[34] Hence, we first *regularized* all the structure for consistent comparison of computed and experimental results.

Second, each amino acid residue **X** in the protein sequence was treated as a terminally blocked tripeptide with the sequence Ac-G**X**G-NMe,[26] with **X** in the conformation of the regularized experimental protein structure, and the $^{13}C^{\alpha}$ isotropic shielding value (σ) for each amino acid residue **X** was computed at the OB98/6-311+G(2d,p) level of theory[32] with the Gaussian 03 package.[38] The remaining residues in each tripeptide were treated at the OB98/3-21G level of theory, i.e., by using the *locally dense basis set* approach.[39]

Third, all ionizable residues were considered neutral during the quantum chemical calculations.[30] This approximation, based on the analysis of 139 conformations of ubiquitin at pH 6.5, indicated that use of neutral, rather than charged, amino acids is a significantly better approximation of the observed $^{13}C^{\alpha}$ chemical shifts in solution for the acidic groups, and a slightly better representation, though significantly less expensive in computational time, for the basic groups.[30]

Fourth, an accurate computation of the reference $^{13}C^{\alpha}$ chemical shifts, not absolute shielding, is of primary interest for protein structure validation because $^{13}C^{\alpha}$ chemical shifts, not the shielding, are the quantities determined with high accuracy in NMR experiments. The most common shielding of the reference used in theoretical applications is that for tetramethylsilane (TMS). Although its computation is a nontrivial problem, because of an assorted number of reasons,[32] it is possible to derive a very accurate solution by using properties of the Normal (or Gaussian) fit of the frequency of the error distribution (between computed and observed $^{13}C^{\alpha}$ chemical shift). With this assumption, an *effective* TMS shielding value can be determined precisely as 184.5 ppm, which must be used in combination with $^{13}C^{\alpha}$ shielding of residues computed at the OB98/6-311+G(2d,p) level of theory.[32]

New Scoring Function: The *ca*-rmsd. For a given protein, the observed $^{13}C^{\alpha}$ chemical shifts represent the contributions from several conformers that coexist in solution. Hence, any scoring function must considerer such dispersion in the conformations of the molecule explicitly in order to be able to reproduce the observed $^{13}C^{\alpha}$ chemical shifts in solution. As a consequence, we hypothesize that the observed chemical shift $^{13}C^{\alpha}{}_{\mathrm{observed},\mu}$ for a given amino acid μ can be interpreted as a conformational average over different internal rotational states represented by a discrete number of different conformations, all of which satisfied NMR restraints such as NOEs, vicinal coupling constants, etc., from which the conformations were derived.[26] Thus, the following quantity can be computed: $^{13}C^{\alpha}{}_{\mathrm{computed},\mu} = \sum_{i=1}^{\Omega} \lambda_i \, ^{13}C^{\alpha}{}_{\mu,i}$, where $^{13}C^{\alpha}{}_{\mu,i}$ is the computed chemical shift for amino acid μ in conformation i out of Ω protein conformations, and λ_i is the Boltzmann weight factor for conformation i, with the condition $\sum_{i=1}^{\Omega} \lambda_i \equiv 1$. With existing computational resources, it is not feasible to determine λ_i at the quantum chemical level, and hence, it is assumed that, under conditions of fast conformational averaging, all Boltzmann weight factors contribute equally and, hence, $\lambda_i \equiv 1/\Omega$. Under this assumption, the computation of the *ca*-rmsd for a protein containing *N* amino acids residues, is straightforward:[26] *ca*-rmsd$^{\alpha} = [(1/N) \sum_{\mu=1}^{N} (^{13}C^{\alpha}_{\text{observed},\mu} - <^{13}C^{\alpha}_{\text{computed},\mu}>)^2]^{1/2}$. Naturally, if $\Omega = 1$, *ca*-rmsd $\equiv$ rmsd, as for any single structure.

Is the $^{13}C^{\alpha}$-based Method a "Strong" Method with Which to Validate X-ray and NMR Structures? A validation method is considered "strong" if it is able to assess how well a structure, or ensemble of structures, predicts experimental data not used in the structure-determination process; otherwise it should be considered "weak", since it is limited to reproducing the observed experimental data used in the determination of the protein models.[31] From this point of view, our use of $^{13}C^{\alpha}$ chemical shifts as a probe for validation is crucial because these experimental data are not used by crystallographers and, hence, our validation method is always "strong", for X-ray-derived structures. However, such a straightforward conclusion cannot be made for NMR-derived structures, because it has been common practice in this field to use information derived from the observed chemical shifts since 1991 when, in a seminal work, Spera and Bax[20] pointed out a clear distinction between the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts in α-helical and β-sheet conformations. However, this Spera and Bax empirical observation provides a set of backbone (ϕ,ψ) dihedral-angle constraints for residues *only* in the regions of regular secondary structure such as α-helix or β-sheet, i.e., to about 40% of the residues in proteins;[40] even more important, no torsional constraints for the side chains are provided, although the influence of the side-chain χ torsional angles on $^{13}C^{\alpha}$ chemical shifts cannot be disregarded.[21,22,24,30] Later, database servers, such as TALOS,[41] provided information about the backbone torsional angles for a larger range (by up to ~75%) of the amino acid residues. Yet, the improvement in the number and accuracy of the backbone torsional angles predicted do not guarantee that the final set of structures will reproduce the observed $^{13}C^{\alpha}$ chemical shifts as accurately as NMR-derived high-resolution proteins solved without using TALOS information. For example, a comparison of the validation results obtained from an ensemble of conformations derived using TALOS information, e.g., for 2JVD, a 48-residue protein (with a *ca*-rmsd per-residue of 0.032 ppm),[29] against validation results obtained from a high-resolution NMR-determined ensemble of conformations obtained *without* using TALOS information, e.g., for 1D3Z, a 76-residue protein (with a *ca*-rmsd per-residue of 0.029 ppm),[32] indicated that the ensemble of conformations of 1D3Z is a better representation of the observed $^{13}C^{\alpha}$ chemical shifts than the ensemble of 2JVD.

Taken as a whole, the concept of "strong" and "weak" is applicable to X-ray structures but is not an issue here, since our validation method deals with reported structures no matter whether an X-ray or NMR technique is used. If chemical-shift derived information was used, as with some structures derived from NMR spectroscopy, our method will also indicate the quality, in terms of the *ca*-rmsd of the final ensemble of conformations, and if such information is misleading, our method will detect it.

3. Global and Local Validation of Proteins Structures

During the past few years, we have applied the $^{13}C^{\alpha}$-based validation method to assess the global quality of an assorted number of proteins in all α-helical,[26,27,29] all β-sheet,[28] and α/β motifs,[26] and spanning a wide range in the number of amino acid residues *N*, namely, in the range $20 \leq N \leq 109$.[26−30] Among all these applications, we selected two highly accurate protein structures solved by both NMR and X-ray methods to illustrate the global validation of proteins and to discuss the question of the legitimacy to choose the X-ray structure as the best set of atomic coordinates, i.e., the "true structure", with which to represent the observed $^{13}C^{\alpha}$ chemical shifts in solution.

Most proteins interact with other proteins, viz., ~80% of ~2 000 yeast proteins were found to be interacting with at least one partner.[42] This might increase the chance of ambiguities in the NOE assignments during protein-structure determination by NMR spectroscopy and, hence, lead to conformational errors. We will also illustrate how the validation of local, rather than global, flaws in the sequence offers an opportunity to spectroscopists for an accurate early detection of the consequences of such possible mis-assignments.

Analysis of the Global Validation of Two Selected Proteins. The selected set of conformations for the analysis were (a) 10 conformers of a 76-residue α/β protein ubiquitin, solved by NMR spectroscopy[43] [PDB id 1D3Z], and the corresponding X-ray structure, solved at 1.8 Å resolution[37] [PDB code 1UBQ]; and (b) 20 conformers of a 48-residue all-α-helical YnzC protein from *Bacillus subtilis* solved by NMR spectroscopy[44] (PDB id 2JVD) and a slightly longer construct of the YnzC protein solved by X-ray diffraction at 2.0 Å resolution[29] (PDB id 3BPH, with three chains in the asymmetric unit) showing identical amino acid residue sequence as the 2JVD structure for the first 46 residues.

Parts a and b of Figure 1 show the results for the validation of these two proteins. In both cases, the *ca*-rmsd (shown as black horizontal line in Figure 1) is a better representation of the observed $^{13}C^{\alpha}$ chemical shift in solution than is a single X-ray structure (green and black bars, yellow and blue bars in parts a and b of Figure 1, respectively). This raises a question as to whether the results reported here are consequences of the "single" model representation of the X-ray data. To answer this question, the room-temperature X-ray structures of ubiquitin (PDB id 1UBQ)[37] and the RNA-binding domain of the nonstructural protein 1 of the influenza A virus (PDB id 1AIL),[45] solved at 1.8 and 1.9 Å resolution, respectively, were used to investigate whether a set of conformations, rather than a single X-ray structure, provides better agreement with *both* the X-ray data and the observed $^{13}C^{\alpha}$ chemical shifts in solution.[46] Among other important findings, our results show that an ensemble of conformations rather than any single structure (shown in parts a and b of Figure 2) sometimes (Figure 2c), but not always (Figure 2d), is a more accurate representation of a protein structure in the crystal; whether or not an ensemble of conformations is a more accurate representation is determined by the dispersion of the coordinates in terms of the all-atom rmsd among the generated models that satisfied the X-ray data.

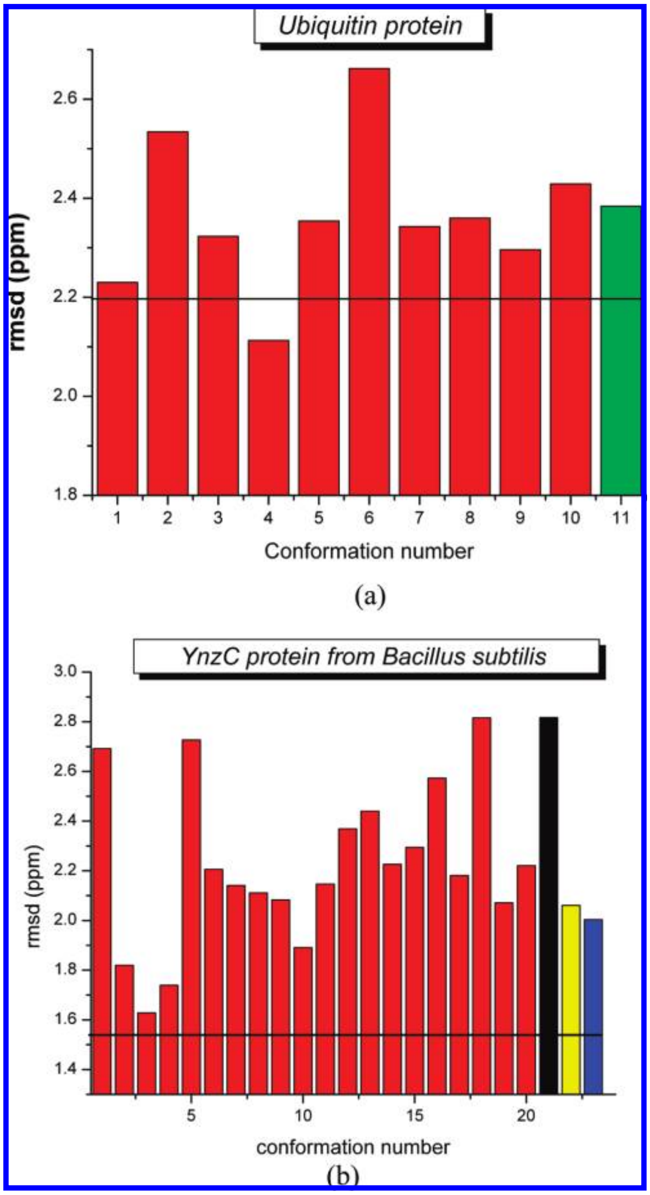


FIGURE 1. (a) Bar diagram of the rmsd between computed and observed $^{13}C^{\alpha}$ chemical shifts for 10 experimental NMR-derived models of ubiquitin (red-filled) [PDB id 1D3Z] and the regularized structure of the X-ray-solved model (2.38 ppm; green-filled) [PDB id 1UBQ]. The black solid horizontal line represents the computed *ca*-rmsd (2.20 ppm) from the 10 NMR conformations; (b) same as (a) for 20 conformations from the NMR-derived models of YnzC (red bars) [PDB id 2JVD] and for each of the three chains in the 2.0 Å crystal structure of YnzC, 3BHP, namely, chains A, B, and C (black, yellow, and blue bars). The amino acid sequence of the YnzC[1−52] (3BHP), YnzC[1−48]$^{-}$ (2JVD) structures are identical *only* for the first 46 residues. Hence, each bar in the figure and the black solid horizontal line representing the computed *ca*-rmsd (1.54 ppm) were computed from the first 46 residues.

Testing the Sensitivity of the Method for Local, Rather than Global, Validation. Despite the enormous progress in techniques and methodologies in both NMR spectroscopy and X-ray diffraction, the existence of errors in the determination of protein structures appears to be common to both techniques.[16,47] Besides the assorted reasons leading to such a problem,[16,47] it is commonly accepted that (global) validation is a necessary, but not sufficient, condition with which to prove that a structure is free of (local) errors. There is, indeed, a need for an accurate validation method at the residue level.[16,46]

As a test of the ability of the $^{13}C^{\alpha}$-based validation method to detect local flaws, we chose to analyze a segment of 27 consecutives residues of a protein structure showing a wrong fold, namely, from the protein dynein light chain 2A (DLC2A, from human) PDB id 1TGQ (now obsolete), and another one showing a correct fold, namely, from protein PDB id 2B95 (that replaced the obsolete 1TGQ in the PDB). Ribbon diagrams of model 1 out of 20 models for proteins 1TGQ and 2B95 are shown in parts a and b of Figure 3, respectively. The difference in the folding between these two structures originated in the oligomeric state assumed during the protein structure determination, namely, as a monomer for 1TGQ and

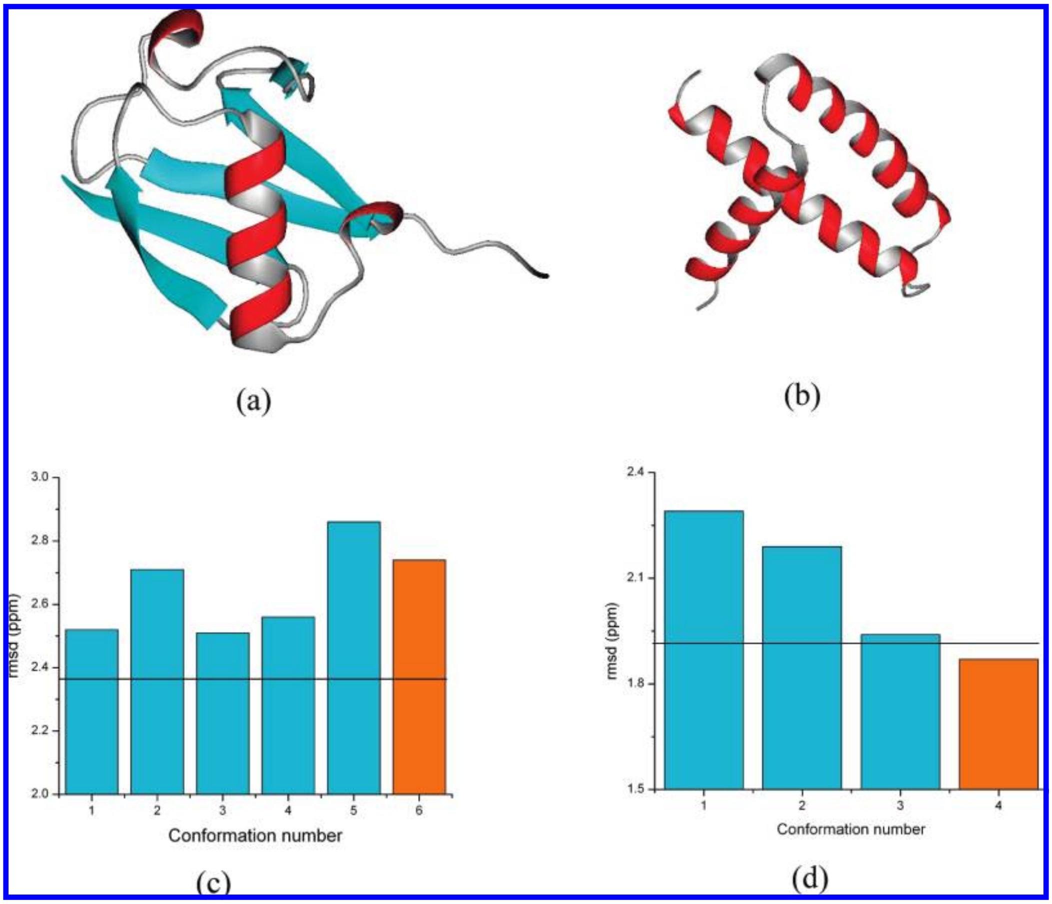


FIGURE 2. Panels (a) and (b) show the ribbon diagram of the protein models of ubiquitin and the RNA-binding domain of the nonstructural protein 1 of the influenza A virus, respectively; these models were obtained[46] after one round of simulated annealing refinement (SAR) starting from the deposited PDB structures of 1UBQ and 1AIL and represented by the orange bars in (c) and (d); these two panels also show the bar diagram of the rmsd between computed and observed $^{13}C^{\alpha}$ chemical shifts, as cyan-filled bars, for the generated ensemble of conformations, generated from the SAR PDB models and, at the same time, showing R and R_{free} factors similar to those of the deposited X-ray structure;[46] the black solid horizontal lines represent the *ca*-rmsd for each ensemble (2.36 and 1.92 ppm for UBQ and AIL, respectively).

a homodimer for 2B95. This was first pointed out by Nabuurs et al.,[16] who carried out a detailed and extensive validation analysis by using several tools such as WHAT IF[11] and PROCHECK[12] for both the protein 1TGQ and the protein DLC2A (from mouse) PDB id 1Y4O (a homologue of 1TGQ since the NMR restraints from 1TGQ were not available). Among other findings, Nabuurs et al.[16] concluded that the use of standard scoring parameters, such as size and number of residual restraint violations, the precision of the structure ensemble, or the fact that most of the residues populate the allowed regions of the Ramachandran map, cannot safely, or unambiguously, assess the accuracy of protein structures. Later, it was shown that structures 1TGQ and 2B95 can be distinguished by comparing how well they fit unassigned NOESY peak list data.[9]

This is an interesting problem for two reasons because it enables us (a) to determine whether our $^{13}C^{\alpha}$-based validation method is able to accurately identify the existence of errors in a segment of 27 residues, from Asp 45 to Asp 71 of protein 1TGQ and the corresponding segment of protein 2B95 (shown in Figure 3), and (b) to illustrate that our validation method is sensitive enough to alert spectroscopists that, even without knowing the correct fold (2B95), or the NMR restraints, the structure of 1TGQ must be revised.

The correlation coefficient R, or *Pearson* coefficient (Press et al., 1992),[48] between observed and computed $^{13}C^{\alpha}$ chemical shifts for the 27 consecutives residues of proteins 1TGQ and 2B95 is 0.74 and 0.90, respectively. Clearly such a significant difference, in terms of R, indicates that careful attention should be paid to the fold of this segment in the protein 1TGQ. Even more important, in the absence of an R value for the correct fold (0.90, for 2B95) the R value obtained from the wrong fold (0.74, for 1TGQ) is low enough to make spectroscopists aware that the conformation of this segment should be carefully revised. In fact, by using the statistical meaning of R^2, it is straightforward to conclude that ∼50% of the observed

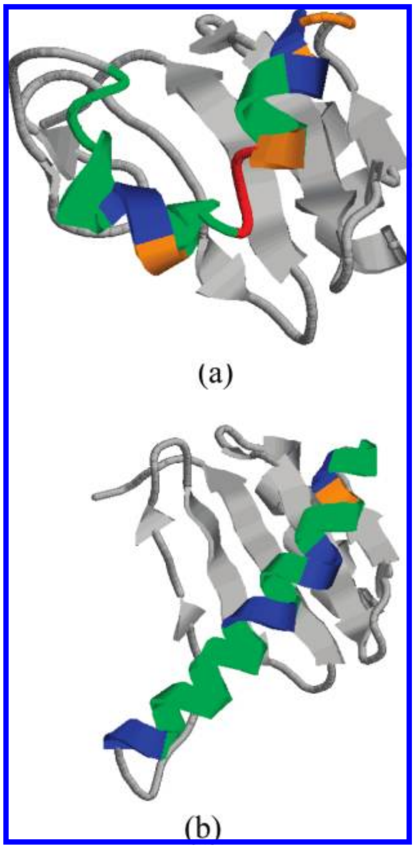


FIGURE 3. (a) Ribbon diagram of model 1, out of 20 models, for the 106-residue protein 1TGQ; (b) same as (a) for model 1 of chain A, out of 20 models, of the 106-residue protein 2B95. In (a) and (b), we highlight in green, blue, orange, and red the 27 residues from residue 45 to 71 of 1TGQ and 2B95, respectively, for which a local validation analysis was carried out. Green color indicates residues with errors within the range of observed standard deviations, i.e., lower than a cutoff value of 2.4 ppm;[47] blue, orange, and red indicate a range of errors greater than the cutoff value, namely, $0 < \Delta\bar{\Delta}_{\mu} \leq 1$ ppm, 1 ppm $< \Delta\bar{\Delta}_{\mu} \leq 2.0$ ppm, and $\Delta\bar{\Delta}_{\mu} > 2.0$ ppm, respectively.

$^{13}C^{\alpha}$ chemical shifts cannot be reproduced by the conformation of this segment in the 1TGQ protein model.

Finally, we tested the ability of our validation method to detect residues in the sequence displaying larger errors than a certain cutoff value between observed and computed $^{13}C^{\alpha}$ chemical shifts. The adopted cutoff value was 2.4 ppm because it is higher than the upper limit of the standard deviation (0.9 ppm $\leq \sigma \leq$ 2.3 ppm) observed by Wang and Jardetzky[49] for $^{13}C^{\alpha}$ chemical shifts (from a database containing more than 6 000 amino acid residues in the α-helix, β-sheet, and statistical-coil conformations). The average error among the nearest-neighbor residues of μ, $\bar{\Delta}_{\mu}$, was adopted as the error for this residue, namely, with $\bar{\Delta}_{\mu} = (1/3)\sum_{x\in\{\mu-1,\mu,\mu+1\}}\Delta_x$, where Δ_x represents the difference between the observed and computed $^{13}C^{\alpha}$ chemical shifts for a given residue in the triplet. Departure of this average error from the cutoff value, $\Delta\bar{\Delta}_{\mu} = \bar{\Delta}_{\mu} - 2.4$ ppm, was used for a colored representation of the error distribution (see Figure 3). Blue, orange, and red colors were used to designate the range of $\Delta\bar{\Delta}_{\mu}$ variations: $0 \leq \Delta\bar{\Delta}_{\mu} \leq 1$ ppm; 1 ppm $< \Delta\bar{\Delta}_{\mu} \leq 2.0$ ppm; and $\Delta\bar{\Delta}_{\mu} > 2.0$ ppm, respectively. Green color indicates that $\Delta\bar{\Delta}_{\mu} < 0$ ppm, and hence, it is free of error since it is within the allowed range of variations. As seen in Figure 3a, the larger errors occur for protein 1TGQ, for Leu 55, Met 56, and His 57 (highlighted in red in Figure 3a). Not surprising, large errors are located in the turnlike region connecting two antiparallel α-helices in protein 1TGQ. On the other hand, all the $\Delta\bar{\Delta}_{\mu}$ values in 2B95 are lower than 1 ppm, except for Thr 49 with 1.4 ppm, and indicated by the orange color in Figure 3b.

4. Concluding Remarks and Perspectives

While computationally intensive, there are four main advantages of this new methodology: (a) it can be used for proteins of *any* class or size; (b) it provides a *strong* methodology with which to validate, at a high-quality level, protein structures as a whole, i.e., by using the *ca*-rmsd; (c) it has potential value to be adopted as a standard routine for determination of local flaws in the sequence without *prior* knowledge of the oligomeric state of the protein in solution, the correct fold of the protein, the NMR restraints, or additional NMR data; and (d) it does not use any knowledge-based information and, hence, it is a purely *physics-based* method.

The most relevant limitation of the method is related to the computational cost. However, recent progress in our laboratory shows that $^{13}C^{\alpha}$ chemical shifts in proteins, computed at the DFT level of theory with a large basis set, can be reproduced accurately (within an average error of ~0.4 ppm) and faster (by ~9 times) by using a small basis set (work in progress). The speed-up of the calculations of the $^{13}C^{\alpha}$ chemical shifts, together with the ever-increasing computational power, will significantly alleviate the computational cost of the method, and hence, it could be adopted as a standard by the NMR community with which to validate a significantly large number of deposited and new protein models.

We are grateful for the contributions of our co-workers whose names appear as coauthors on citations in this Account; in addition we thank Dr. Y.A. Arnautova of the Department of Chemistry, Cornell University, and Dr. P. Serrano of the Department of Molecular Biology, The Scripps Research Institute, for valuable discussions. This research was supported by grants from the National Institutes of Health (GM-14312 and GM-24893) and the National Science Foundation (MCB05-41633). Support was also received from the CONICET, FONCyT-

ANPCyT (PAV 22642/22672), and from the Universidad Nacional de San Luis (P-328501), Argentina. The research was conducted using the resources of Pople, a facility of the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center.

BIOGRAPHICAL INFORMATION

Jorge A. Vila was born in Rivadavia (Mendoza, Argentina) in 1952. He is Professor of the National University of San Luis−Argentina, Researcher of CONICET−Argentina, and Senior Research Associate at Cornell University (U.S.A.).

Harold A. Scheraga was born in Brooklyn, NY, in 1921. He attended the City College of NY, receiving his B.S. in 1941, and he received his Ph.D. at Duke University in 1946. Following postdoctoral work at Harvard Medical School, he joined the faculty of Cornell University in 1947, where he is now Todd Professor of Chemistry, Emeritus.

FOOTNOTES

*Corresponding author: has5@cornell.edu.

REFERENCES

1 Arnold, J. T.; Dharmatti, S. S.; Packard, M. E. Chemical effects on nuclear induction signals from organic compounds. *J. Chem. Phys.* **1951**, *19*, 507.

2 Szilagyi, L. Chemical shifts in proteins come of age. *Prog. Nucl. Magn. Reson. Spectrosc.* **1995**, *27*, 325–443.

3 Helgaker, T.; Jaszuński, M.; Ruud, K. Ab initio methods for the calculation of NMR shielding and indirect spin−spin coupling constant. *Chem. Rev.* **1999**, *99*, 293–352.

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

5 Wüthrich, K. NMR studies of structure and function of biological macromolecules (Nobel Lecture). *J. Biomol. NMR* **2003**, *27*, 13–39.

6 Bax, A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* **2003**, *12*, 1–16.

7 Dyson, H. J.; Wright, P. E. Unfolded proteins and protein folding studied by NMR. *Chem. Rev.* **2004**, *104*, 3607–3622.

8 Sternberg, U.; Witter, R.; Ulrich, A. 3D structure elucidation using NMR chemical shifts. *Annu. Rep. NMR Spectrosc.* **2004**, *52*, 53–104.

9 Bhattacharya, A.; Tejero, R.; Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **2007**, *66*, 778–795.

10 Billeter, M.; Wagner, G.; Wüthrich, K. Solution NMR structure determination of proteins revisited. *J. Biomol. NMR* **2008**, *42*, 155–158.

11 Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8*, 52–56.

12 Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.

13 Huang, Y. J.; Powers, R. Montelione GT Protein NMR Recall, Precision, and F-measure scores (RPF scores): Structure quality assessment measures based on information retrieval statistics. *J. Am. Chem. Soc.* **2005**, *127*, 1665–1674.

14 Davis, I. W.; Fay-Leaver, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, B. W., III; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: All atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007**, *35*, W375–383.

15 Snyder, D. A.; Bhattacharya, A.; Huang, J. Y.; Montelione, G. T. Assessing precision and accuracy of protein structures derived from NMR data. *Proteins* **2005**, *59*, 655–661.

16 Nabuurs, S. B.; Spronk, C. A. E. M.; Vuister, G. W.; Vriend, G. Tradional Biomolecular Structure Determination by NMR Spectroscopy Allows for Major Errors. *PLOS Comp. Biol.* **2006**, *2*, 71–79.

17 Clore, G. M.; Garrett, D. R-factor, free R, and complete cross-validation for dipolar coupling refinement of NMR structures. *J. Am. Chem. Soc.* **1999**, *121*, 9008–9012.

18 Simon, K.; Xu, J.; Kim, C.; Skrynnikov, N. R. Estimating the accuracy of protein structures using residual dipolar couplings. *J. Biomol. NMR* **2005**, *33*, 83–93.

19 Jurgen, F. D.; Aart, J. N.; Wim, V.; Jundomg, L.; Alexandre, M. J. J. B.; Robert, K.; John, L. M.; Eldon, L. U. BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J. Biomol. NMR* **2005**, *32*, 1–12.

20 Spera, S.; Bax, A. Empirical correlation between protein backbone conformation and C^{α} and C^{β} ^{13}C Nuclear Magnetic Resonance chemical shifts. *J. Am. Chem. Soc.* **1991**, *113*, 5490–5492.

21 Pearson, J. G.; Le, H.; Sanders, L. K.; Godbout, N.; Havlin, R. H.; Oldfield, E. J. Predicting chemical shifts in proteins: Structure refinement of valine residues by using *ab initio* and empirical geometry optimizations. *J. Am. Chem. Soc.* **1997**, *119*, 11941–11950.

22 Havlin, R. H.; Le, H.; Laws, D. D.; deDios, A. C.; Oldfield, E. 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.* **1997**, *119*, 11951–11958.

23 Iwadate, M.; Asakura, T.; Williamson, M. P. C^{α} and C^{β} carbon-13 chemical shifs in protein from an empirical database. *J. Biomol. NMR* **1999**, *13*, 199–211.

24 Villegas, M. E.; Vila, J. A.; Scheraga, H. A. Effects of Side-Chain Orientation on the ^{13}C Chemical Shifts of Antiparallel β-sheet Model Peptides. *J. Biomol. NMR* **2007**, *37*, 137–146.

25 Sun, H.; Sanders, L. K.; Oldfield, E. Carbon-13 NMR shielding in the twenty common amino acids: Comparisons with experimental results in proteins. *J. Am. Chem. Soc.* **2002**, *124*, 5486–5495.

26 Vila, J. A.; Villegas, M. E.; Baldoni, H. A.; Scheraga, H. A. Predicting $^{13}C^{\alpha}$ chemical shifts for validation of protein structures. *J. Biomol. NMR* **2007**, *38*, 221–235.

27 Vila, J. A.; Ripoll, D. R.; Scheraga, H. A. Use of $^{13}C^{\alpha}$ chemical shifts in protein structure determination. *J. Phys. Chem. B* **2007**, *111*, 6577–6585.

28 Vila, J. A.; Arnautova, Y. A.; Scheraga, H. A. Use of $^{13}C^{\alpha}$ chemical shifts for accurate determination of β-Sheet structures in solution. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1891–1896.

29 Vila, J. A.; Aramini, J. M.; Rossi, P.; Kuzin, A.; Su, M.; Seetharaman, J.; Xiao, R.; Tong, L.; Montelione, G. T.; Scheraga, H. A. Quantum Chemical $^{13}C^{\alpha}$ Chemical Shift Calculations for Protein NMR Structure Determination, Refinement, and Validation. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14389–14394.

30 Vila, J. A.; Scheraga, H. A. Factors affecting the use of $^{13}C^{\alpha}$ chemical shifts to determine, refine, and validate protein structures. *Proteins* **2008**, *71*, 641–654.

31 Kleywegt, G. J. On vital aid: The why, what and how of validation. *Acta Crystallogr.* **2009**, *D65*, 134–139.

32 Vila, J. A.; Baldoni, H. A.; Scheraga, H. A. Performance of density functional models to reproduce observed $^{13}C^{\alpha}$ chemical shifts of protein solution. *J. Comput. Chem.* **2009**, *30*, 884–892.

33 Wang, X.; Bansal, S.; Jiang, M.; Prestegard, J. H. RDC-assisted modeling of symmetric protein homo-oligomers. *Protein Sci.* **2008**, *17*, 899–907.

34 Némethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to praline-containing peptides. *J. Phys. Chem.* **1992**, *96*, 6472–6484.

35 de Dios, A. C.; Pearson, J. G.; Oldfield, E. Chemical shifts in proteins: An *ab initio* study of carbon-13 nuclear magnetic resonance chemical shielding in glycine alanine and valine residues. *J. Am. Chem. Soc.* **1993**, *115*, 9768–9773.

36 Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

37 Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **1987**, *194*, 531–544.

38 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T., Jr.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O., Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.;

Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision E.01*; Gaussian, Inc.: Wallingford, CT, 2004.

39 Chesnut, D. B.; Moore, K. D. Locally dense basis-sets for chemical-shift calculations. *J. Comput. Chem.* **1989**, *10*, 648–659.

40 Xu, X.-P.; Case, D. A. 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* **2001**, *21*, 321–333.

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

42 Levy, E. D.; Pereira-Leal, J. B.; Chotia, C.; Teichmann, S. A. 3D complex: A structural classification of protein complexes. *PLOS Comp Biol.* **2006**, *2*, 1396–1406.

43 Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.

44 Aramini, J. M.; Sharma, S.; Huang, Y. J.; Swapna, G. V. T.; Ho, C. K.; Shetty, K.; Cunningham, K.; Ma, L.-C.; Zhao, L.; Owens, L. A.; Jiang, M.; Xiao, R.; Liu, J.; Baran, M. C.; Acton, T. B.; Rost, B.; Montelione, G. T. Solution NMR structure of the SOS response protein YnzC from Bacillus subtilis. *Proteins* **2008**, *72*, 526–530.

45 Liu, J.; Lynch, P. A.; Chien, C.-y.; Montelione, G. T.; Krug, R. M.; Berman, H. M. Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. *Nat. Struct. Biol.* **1997**, *4*, 896–899.

46 Arnautova, Y. A.; Vila, J. A.; Martin, O. A.; Scheraga, H. A. What can we learn by computing $^{13}C^{\alpha}$ chemical shifts for X-ray protein models? *Acta Crystallogr.* **2009**, *D65*, 697−703.

47 Kleywegt, G. J. Validation of protein crystal structures. *Acta Crystallogr.* **2000**, *D56*, 249–265.

48 Press, H. W.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. In *Numerical Recipes in FORTRAN 77. The Art of Scientific Computing*, second ed.; Cambridge University Press: New York, 1992; Chapter 14, pp 630−633.

49 Wang, Y. J.; Jardetzky, O. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci.* **2002**, *11*, 852–861.