Quantum-mechanics-derived $^{13}C^{\alpha}$ chemical shift server (*Che*Shift) for protein structure validation

Jorge A. Vila[a,b], Yelena A. Arnautova[a,1], Osvaldo A. Martin[b], and Harold A. Scheraga[a,2]

[a]Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca NY, 14853-1301; and [b]Universidad Nacional de San Luis, Instituto de Matemática Aplicada de San Luis-Consejo Nacional de Investigaciones Cientificas y Técnicas, Ejército de Los Andes 950-5700 San Luis, Argentina

Contributed by Harold A. Scheraga, August 7, 2009 (sent for review July 16, 2009)

A server (*Che*Shift) has been developed to predict $^{13}C^{\alpha}$ chemical shifts of protein structures. It is based on the generation of 696,916 conformations as a function of the ϕ, ψ, ω, $\chi1$ and $\chi2$ torsional angles for *all* 20 naturally occurring amino acids. Their $^{13}C^{\alpha}$ chemical shifts were computed at the DFT level of theory with a small basis set and extrapolated, with an empirically-determined linear regression formula, to reproduce the values obtained with a larger basis set. Analysis of the accuracy and sensitivity of the *Che*Shift predictions, in terms of both the correlation coefficient *R* and the conformational-averaged rmsd between the observed and predicted $^{13}C^{\alpha}$ chemical shifts, was carried out for 3 sets of conformations: (*i*) 36 x-ray-derived protein structures solved at 2.3 Å or better resolution, for which sets of $^{13}C^{\alpha}$ chemical shifts were available; (*ii*) 15 pairs of x-ray and NMR-derived sets of protein conformations; and (*iii*) a set of decoys for 3 proteins showing an rmsd with respect to the x-ray structure from which they were derived of up to 3 Å. Comparative analysis carried out with 4 popular servers, namely SHIFTS, SHIFTX, SPARTA, and PROSHIFT, for these 3 sets of conformations demonstrated that *Che*Shift is the most sensitive server with which to detect subtle differences between protein models and, hence, to validate protein structures determined by either x-ray or NMR methods, if the observed $^{13}C^{\alpha}$ chemical shifts are available. *Che*Shift is available as a web server.

chemical shifts prediction | DFT calculations | validation server

Accurate and fast validation of protein structures constitutes a long-standing problem in NMR spectroscopy (1–3). Investigators have proposed a plethora of methods to determine the accuracy and reliability of protein structures in recent years (4–8). Despite this progress, there is a growing need for more sophisticated, physics-based and fast structure-validation methods (1, 2, 7). With these goals in mind, we recently proposed a new, physics-based solution of this important problem (9), viz., a methodology that makes use of observed and computed $^{13}C^{\alpha}$ chemical shifts (at the DFT level of theory) for an accurate validation of protein structures in solution (9) and in a crystal (10). Assessment of the ability of computed $^{13}C^{\alpha}$ chemical shifts to reproduce observed values for a single or an ensemble of structures in solution and in a crystal was accomplished by using the conformationally-averaged root-mean-square-deviation (*ca*-rmsd) as a scoring function (9). While computationally intensive, this methodology has several advantages: (*i*) it makes use of the $^{13}C^{\alpha}$ chemical shifts, not shielding, that are ubiquitous to proteins; (*ii*) it can be computed accurately from the φ, ψ, and χ torsional angles; (*iii*) there is no need for a priori knowledge of the oligomeric state of the protein; and (*iv*) no knowledge-based information or additional NMR data are required.

However, the primary and the most serious limitation of the method is the computational cost of such calculations, which prevents it from being adopted by spectroscopists and crystallographers as a standard validation routine (9). For this reason, we investigate here the dependence of the accuracy and speed of DFT calculations of the $^{13}C^{\alpha}$ chemical shifts in proteins on the size of the basis set used. The results of this analysis indicate that the $^{13}C^{\alpha}$ 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 approximately 0.4 ppm) and approximately 9 times faster by using a small basis set. As a straightforward application of these findings, a server of the $^{13}C^{\alpha}$ chemical shifts (*Che*Shift) for *all* 20 naturally occurring amino acid residues as a function of the ϕ, ψ, ω, $\chi1$, and $\chi2$ torsional angles was built. This server can be used to validate protein structures of any class or size at a high-quality level and, like the purely physics-based method (9) from which it was derived, it does not use any knowledge-based information. However, the *Che*Shift server also provides accurate $^{13}C^{\alpha}$ chemical shift predictions for each amino acid residue in the sequence in a few seconds, on a single processor. These are the main advantages of this new quantum-mechanics-derived *Che*Shift server over our previous approach (9).

There are several servers that provide fast and accurate predictions of $^{13}C^{\alpha}$ chemical shifts, namely SHIFTS (11, 12), SHIFTX (13), PROSHIFT (14), and SPARTA (15) A brief description of these servers, follows. SHIFTS (11, 12) is a DFT-computed server of chemical shifts for residues in α-helical or β-sheet conformations, plus a coil-database derived as a single average of the sheet and helix data (optimized by comparison with experimental data); SHIFTX[13] is a hybrid predictive approach that employs precalculated empirically-derived chemical shift hypersurfaces in combination with classical or semiclassical equations (ring current, electric field, hydrogen bond, solvent effects, etc.). SHIFTX used 2 databases of 37 protein structures as input to generate the empirical constants, torsional angles and lookup tables; PROSHIFT[14] is a neural-network-trained server, derived by using experimental 3D structures of proteins as input parameter; and SPARTA (15) is a server containing observed chemical shifts for 200 proteins for which a high resolution (≤2.4 Å) x-ray structure is available. The relative importance of the weighting factors for the ϕ,ψ, and $\chi1$ torsional angles and sequence similarity was optimized empirically.

The existence of these servers raises the question as to whether a new server, such as *Che*Shift, is necessary. What new information can we learn from its predictions? Even more important, if there are substantial differences among predictions from these servers and *Che*Shift, what is the origin of such differences? Comparison of the $^{13}C^{\alpha}$ chemical shifts, computed for a large number of proteins using different servers, with the corresponding experimental data are very important because it can shed light on the strengths and weaknesses of each server. It will also enable us to determine which servers are sensitive enough to detect subtle differences between conformations and whether it

Author contributions: J.A.V., Y.A.A., and H.A.S. designed research; J.A.V., Y.A.A., O.A.M., and H.A.S. performed research; J.A.V., Y.A.A., O.A.M., and H.A.S. analyzed data; and J.A.V., Y.A.A., and H.A.S. wrote the paper.

The authors declare no conflict of interest.

[1]Present address: Molsoft L.L.C., 3366 North Torrey Pines Court, Suite 300, La Jolla, CA, 92037.

[2]To whom correspondence should be addressed. E-mail: has5@cornell.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0908833106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0908833106

is able to indicate to spectroscopists or crystallographers whether an ensemble, rather than a single conformation, is a better representation of the observed ^{13}C$^{\alpha}$ chemical shifts in solution (9).

Several attempts have also been made recently to use ^{13}C$^{\alpha}$ chemical shift data to facilitate protein structure determination and refinement (16, 17) and to derive initial protein models for molecular replacement in x-ray crystallography (18). The ability of a given server to guide protein structure refinement can be assessed by its discriminative power when applied to protein decoys generated from a given, native conformation.

To compare the performance of *Che*Shift with that of other existing servers, the following 3 sets of proteins are analyzed in this work: 36 x-ray-derived protein structures solved at 2.3 Å resolution or better for which sets of ^{13}C$^{\alpha}$ chemical shifts were also available (*Section II*, Table S3, *SI Appendix*); 15 pairs of x-ray and NMR-derived protein conformations (*Section II*, Table S4, *SI Appendix*); and decoys from the ROSETTA@HOME set (19), namely 1AIL (20), 1RNB (21), and 1UBI (22) from the Protein Data Bank (PDB) (23), showing an rmsd of up to 3 Å from the corresponding x-ray structure.

The most important results related to the performance of *Che*Shift, and 4 other servers, are discussed here. Additional material related to the dependence of the accuracy and speed of DFT calculations of the ^{13}C$^{\alpha}$ chemical shifts of proteins on the size of the basis set used; analysis of x-ray and x-ray-NMR pairs of structures; and approximations used to interpolate computed ^{13}C$^{\alpha}$ chemical shift values are provided in *Sections I*, *II*, and *III*, respectively of the *SI Appendix*.

Results and Discussion

In the absence of a "gold standard" against which to compare the predictions obtained from any servers, we adopted the ^{13}C$^{\alpha}$ chemical-shift values computed at the DFT level of theory by using a large basis set as an 'internal standard reference' (see *Materials and Methods*).

Determination of the Sensitivity of All Servers for Members of a Set of 36 X-Ray-Derived Protein Structures. Results of the analysis of the ^{13}C$^{\alpha}$ chemical-shift predictions for each of 36 x-ray-derived protein models, based on the correlation coefficient *R* (24), obtained by using SHIFTS, SHIFTX, PROSHIFT, SPARTA, and *Che*Shift are shown in Table S3 (*Section II*, *SI Appendix*). The differences in the *R* ranges are small among all of the servers, around 0.04 depending on the protein, albeit these small differences could be important for an accurate prediction. Besides, these results indicate that, for all of the proteins, the *R* value obtained from any server is greater than the one obtained from *Che*Shift. This raises the following question: do these servers provide a more sensitive validation method than *Che*Shift? To answer this question, 2 of the 36 validation results are analyzed here in detail.

Protein 1RGE (Ribonuclease Sa). The structure of this protein was solved (25) at 1.15 Å resolution with an R-factor of 10.9%. The corresponding crystal structure contains 2 chemically identical but crystallographically independent molecules in the asymmetric unit, named here as A and B (25). The main-chain torsional angles (ϕ and ψ) of the independent molecules are very similar (25) with the C$^{\alpha}$ rmsd between them of 0.4 Å. On the other hand, the all-heavy-atom rmsd is 1.1 Å due to differences in side chains, especially those on the protein surface, occupying different rotameric states.

Comparison of the predicted ^{13}C$^{\alpha}$ chemical shifts, computed as the conformational-average (*ca*) (9) between the 2 chains, with the observed ^{13}C$^{\alpha}$ chemical shifts yields *R* values of 0.95, 0.98, 0.99, 0.97, and 0.97 for *Che*Shift, SHIFTX, SPARTA, SHIFTS, and PROSHIFT, respectively.

At first glance, all servers appear to be more accurate than *Che*Shift. However, it is necessary to determine whether all servers are sensitive enough to detect differences between the independent molecules A and B. To answer this question, we carried out an additional test that does not require a comparison with the observed ^{13}C$^{\alpha}$ chemical shifts. Thus, we computed the correlation coefficient *R* between the ^{13}C$^{\alpha}$ chemical-shift predictions obtained for molecules A and B, respectively, by using each of the 5 servers. The results of this test give the following *R* values: 0.96, 1.00, 1.00, 0.98, and 1.00 for *Che*Shift, SHIFTX, SPARTA, SHIFTS, and PROSHIFT, respectively (see Table S3, *SI Appendix*). Except for *Che*Shift (0.96*3*) and SHIFTS (0.98*1*), none of the servers is able to discriminate, beyond doubt, between molecules A and B. From a statistical point of view, the *R* values obtained from SHIFTX (0.99*7*), SPARTA (0.99*7*), and PROSHIFT (0.99*6*) servers indicate that molecules A and B are practically indistinguishable protein models with which to compute the ^{13}C$^{\alpha}$ chemical shifts. In other words, these 3 servers cannot detect the conformational difference between molecules A and B.

This test enables us to conclude that a lower *R* value between predicted and observed ^{13}C$^{\alpha}$ chemical shifts does not necessarily mean poorer accuracy; on the contrary, it could mean higher sensitivity to detect subtle structural differences.

If this were a valid conclusion, a similar analysis carried out with a larger basis set, namely using the results from the more-accurate "internal standard reference," should lead to a lower correlation, *R*, between ^{13}C$^{\alpha}$ chemical shifts predicted for molecules A and B. Indeed, this is the case. The *R* value (0.93) computed with the larger basis set is significantly lower than the *R* value obtained with *Che*Shift (0.96) or any other server, namely, 1.00, 1.00, 0.98, and 1.00 for SHIFTX, SPARTA, SHIFTS, and PROSHIFT, respectively.

The previous analysis demonstrates that *Che*Shift is a more sensitive server to detect subtle structural differences than any other servers, although this analysis does not reveal the origin of such sensitivity. To detect this origin, we first carried out a graphic analysis of the correlation between corresponding torsional angles in molecules A and B, and, second, from these graphs we determined the distribution of differences between predicted ^{13}C$^{\alpha}$ chemical shifts for each of these 2 molecules by using the *Che*Shift server. Fig. S2 (*Section II*, *SI Appendix*) shows the well-known (25) strong correlation between the corresponding backbone torsional angles derived from molecules A and B of protein 1RGE. Consistently, with the low value of the C$^{\alpha}$ rmsd (0.4 Å) between these 2 molecules, the correlation coefficient, *R*, computed between backbone torsional angles of the molecules A and B of 1RGE, is greater than 0.99.

On the other hand, Fig. 1 shows the correlation between corresponding side-chain torsional angles χ_1 for molecule A and B. The *R* value (0.92, obtained after removing the, approximately, identical $\chi 1 = \approx 180^{o}$ and $\sim -180^{o}$, see Fig. 1) is lower than the one obtained for the backbone torsional angles (>0.99), indicating the significantly higher side-chain all-heavy-atom rmsd between molecules A and B of 1RGE (1.1 Å). To determine whether the observed differences in the side-chain $\chi 1$ torsional angles, shown in Fig. 1, are the origin of the *R* values yielded by *Che*Shift for molecules A and B, we highlighted those residues showing differences between predicted ^{13}C$^{\alpha}$ chemical shifts for molecules A and B of 1RGE, greater than 2.0 ppm. Among all 8 residues, 5 (highlighted as black-filled stars in Fig. 1) show a significant departure from the linear regression. These 5 residues possess significantly different side-chain $\chi 1$ torsional angles in molecules A and B and, hence, significantly different ^{13}C$^{\alpha}$ chemical shift predictions. Two out of these 5 highlighted residues in Fig. 1, namely Asp-25 and Arg-40, were reported (25) to have higher temperature factors or partial disorder of the side-chains. Another 2 (Ser-48 and Thr-76) of these 5 residues

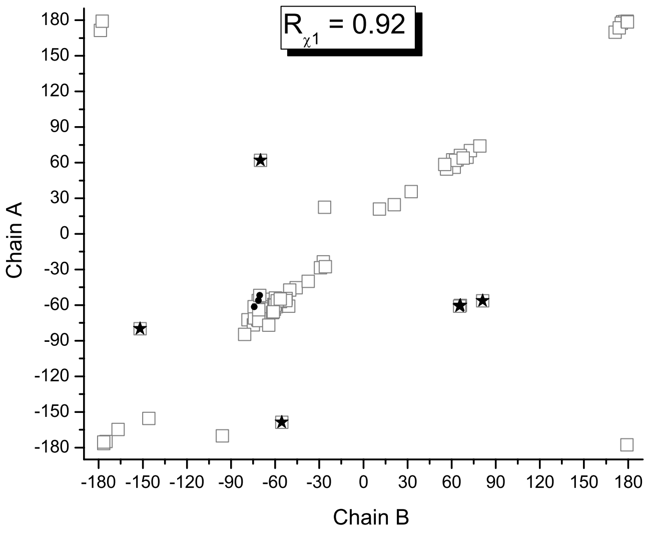


Fig. 1. Plot of the $\chi1$ torsional angles in degrees (as open-squares) from chain A versus chain B of the x-ray-determined structure of PDB ID 1RGE. We highlighted those residues showing differences greater than 2.0 ppm between predicted $^{13}C^{\alpha}$ chemical shifts from molecule A and B of 1RGE by *Che*Shift with filled stars and filled circles. For details about the distribution of the black-filled symbols, see *Protein 1RGE (Ribonuclease Sa).*

are in different crystal environment, i.e., these residues of molecule A are part of a well-ordered hydrogen-bond network, while they are oriented toward the solvent in molecule B.[25] In particular, the influence of 2 different torsional angles $\chi1$, for a fixed $\chi2$, on the predicted *Che*Shift value of any Ser residue is illustrated in Fig. 2*A*. Finally, the remaining residue, Arg-63, of these 5 is located in the loop region (Gly-61-Thr-64) whose conformation is different between the 2 molecules (25).

There are 3 other residues in Fig. 1, namely Gln-32, Ile-71 and Gln-77, close to the regression line and highlighted as filled-black circles. Their $\chi1$ torsional angles are very similar, although all 3 residues show significant differences in the side-chain $\chi2$ torsional angles, namely ($\approx g+$, $\approx g-$) for Gln-32 and Ile-71, and ($\approx g-$, $\approx t$) for Gln-77. To illustrate the influence of 2 different torsional angles $\chi2$ for a fixed $\chi1$ on the predicted *Che*Shift value, we selected Gln (in Fig. 2*B*).

Finally, it is important to note that, for a given molecule, A or B of 1RGE, some residues are reported to show 2 discrete side-chain conformations (25). Differences up to 4.0 ppm between *Che*Shift-predicted chemical shifts are obtained by using these alternative side-chain conformations, as for Thr5A, Val6A, and Ser42A from molecule A or Ser3B and Thr5B from molecule B. For some of these residues, the alternative side-chain conformational dilemma can be resolved easily by inspection of the occupancy, e.g., Thr5A shows 1 of the 2 conformations with much higher occupancy (approximately 80%) (25). However, in other residues, such as Val6A and Ser42A of molecule A or Ser3B and Thr5B of molecule B, the alternative conformations show very similar occupancies (approximately 50%) (25) although significantly different chemical shifts; if the occupancy does not offer conclusive evidence, and if it is necessary to select one conformation, then the *Che*Shift predictions could be a useful criterion with which to decide which 1 of the 2 conformations should be selected.

The results derived from the analysis of 2 chains of protein 1RGE enable us: (*i*) to illustrate that a higher correlation coefficient, *R*, obtained for the $^{13}C^{\alpha}$ chemical shift prediction between molecules A and B could mean less sensitivity to detect subtle structural differences, rather than more accurate predictions; and (*ii*) to determine the origin of the difference between $^{13}C^{\alpha}$ chemical shift predictions for the 2 molecules; i.e., although the main contribution determining the predicted $^{13}C^{\alpha}$ chemical shifts comes from backbone torsional angles, a proper consideration of the side-chain torsional angles ($\chi1$ and $\chi2$) is very important for an accurate $^{13}C^{\alpha}$ chemical shift validation (see Fig. 2 *A*–*B*). The latter conclusion is in agreement with evidence (11, 26–28) indicating the role of side-chain conformations in the computation of accurate $^{13}C^{\alpha}$ chemical-shift values.

Protein Interleukin 1*β* (Human). The computed $^{13}C^{\alpha}$ chemical shifts for 2 different x-ray structures of this protein solved at 2.0 Å resolution and refined to a crystallographic R-factor of 19.0% (4I1B) (29) and 17.2% (2I1B) (30), are compared with the observed $^{13}C^{\alpha}$ chemical shifts in solution [Biological Magnetic Resonance data Bank (BMRB) accession no. 1061(31)]. The all-heavy-atom rmsd between these 2 x-ray structures is 1.1 Å with a difference,

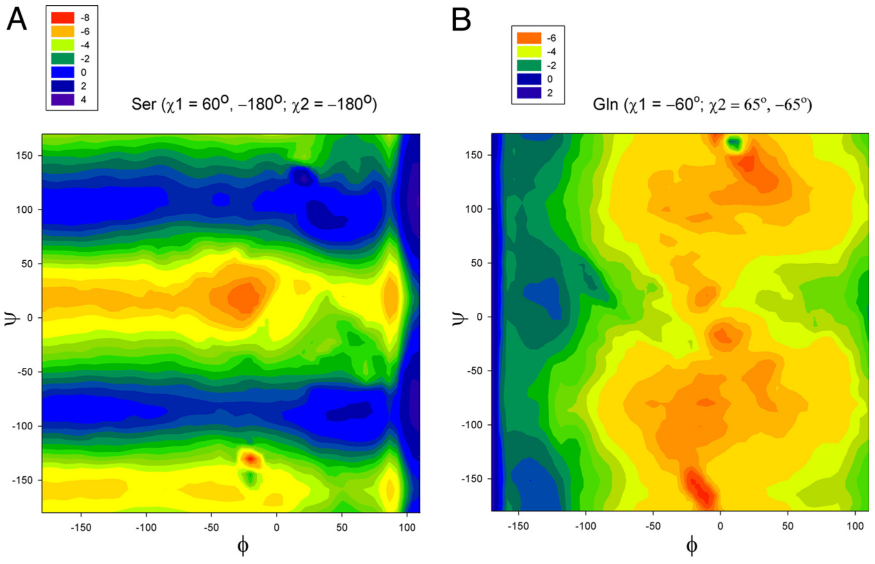


Fig. 2. Map of the differences in the computed $^{13}C^{\alpha}$ chemical shifts (in ppm according to the color scale) between an arbitrarily selected pair of side-chain torsional angles. The color indicates the difference in $^{13}C^{\alpha}$ chemical shifts (in ppm) for any pair of ϕ and ψ. (*A*) for Ser with $\chi1$ = 160° and −180°, and a fixed $\chi2$ = −180°; and panel (*B*) for Gln with $\chi2$ = 65° and −65°, and a fixed $\chi1$ = −60°.

mainly, in the loop regions (29), e.g., the all-heavy-atom rmsd between loop residues His-30-Val-41, Val-47-Asp-54, Val-85-Glu-96, and Gly-136-Asp-145 are 1.5 Å, 1.6 Å, 1.3 Å, and 0.9 Å, respectively. The results for *R* obtained with *Che*Shift, SHIFTX, and SPARTA point to 2I1B, rather than 4I1B, as a better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution (see *Section II*, Table S3, *SI Appendix*). However, only *Che*Shift indicated that 2I1B ($R = 0.91$) is a significantly better model than 4I1B ($R = 0.87$) to reproduce the observed $^{13}C^{\alpha}$ chemical shifts. In fact, the *Che*Shift *R* value indicated that approximately 83% of the observed $^{13}C^{\alpha}$ chemical shifts in solution are reproduced by the 2I1B protein model, compared to only approximately 76% of protein 4I1B. A similar analysis, carried out with SHIFTX and SPARTA, indicates that approximately 92% and approximately 96% of the observed $^{13}C^{\alpha}$ chemical shifts in solution are reproduced by the 2I1B protein model and, a slightly smaller, approximately 90% and approximately 94% by 4I1B, respectively. The SHIFTS server points to protein 4I1B (approximately 90%), rather than 2I1B (approximately 88%), as a better representation of the observed $^{13}C^{\alpha}$ chemical shifts. On the other hand, according to PROSHIFT predictions, both proteins are equivalent models with which to reproduce (approximately 92% of) the observed $^{13}C^{\alpha}$ chemical shifts (despite the differences between these 2 structures, mainly, in the loop regions). Clearly, for this protein too, *Che*Shift provides a more sensitive discrimination between different models.

As an additional test, the $^{13}C^{\alpha}$ chemical shifts computed by using the internal standard reference indicated that protein 2I1B is, in fact, a significantly better representation of the observed chemical shifts in solution ($R = 0.87$), than protein 4I1B ($R = 0.81$), in agreement with the *Che*Shift predictions.

Regarding the disagreement between *Che*Shift and SHIFTS, the latter server contains, besides a database of DFT-computed shifts for residues populating helical and sheet conformations, a "coil" database (with coil designating residues belonging to neither helical nor β-sheet region) computed as an average of helix and sheet data. Conceivably, this could be the reason for this disagreement, since the 4I1B and 2I1B proteins differ, mainly, in the loop (i.e., coil) regions. In other words, although quantum-mechanical calculations are a common feature of both *Che*Shift and SHIFTS, these calculations were limited to only some regions of the Ramachandran map for SHIFTS but not for the *Che*Shift server.

Are the Servers Sensitive Enough to Determine Differences Between X-Ray and NMR Models? Test on 15 Pairs of X-Ray and NMR-Derived Sets of Protein Conformations. The results obtained from the validation analysis involving 15 pairs of x-ray and sets of NMR-derived protein models are shown in Table S4 (*Section II*, *SI Appendix*). For the NMR-derived conformations, the *R* values were computed between the observed $^{13}C^{\alpha}$ chemical shifts and the predicted conformational-averaged ones (9), i.e., among all structures of the NMR-derived ensemble (see *Computation of the Conformationally-Averaged rmsd* in *Section I* of *SI Appendix*). As already noted for an x-ray set of structures, all *R* values computed by *Che*Shift are systematically lower than those of the other servers. However, for several pairs of x-ray and NMR-derived conformations most of the servers, but not *Che*Shift, do not show differences between x-ray and NMR-derived structures, or the differences are very small, in terms of the correlation coefficient, *R*. To understand whether these results reflect real similarity between the x-ray and NMR models or arise from the low sensitivity of the servers, 2 cases have been selected for further analysis, namely protein PDB ID 3LZT (32) solved by x-ray diffraction at 0.92 Å resolution, and 50 conformations of PDB ID 1E8L (33) (solved by NMR spectroscopy), and protein PDB ID 1UBQ (34) (x-ray derived structure at 1.8 Å resolution) and 128 conformations of PDB ID 1XQQ (35) (NMR-derived ensemble).

A Comparative Validation Analysis of Proteins 1E8L and 3LZT. The NMR solution structure of hen Lysozyme (PDB ID 1E8L) (33), determined with the aid of residual dipolar coupling data, shows "…conformational disorder within the NMR ensemble in some regions of the structure, most notably in the long loop and involving residues in the turns between helices A and B and between the first two strands in the β-sheet (33)." Model 1 of the 50 models of the NMR-derived ensemble (PDB ID 1E8L) (33) has an all-heavy-atom rmsd of 2.20 Å from the corresponding x-ray structure (PDB ID 3LZT), solved at 0.92 Å resolution (32), indicating major conformational differences between these 2 structures. The results shown in Table S4 (*Section II*, *SI Appendix*) indicate that only *Che*Shift, SHIFTX, and SPARTA point to the x-ray structure, but not the NMR-derived ensemble, as a better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution. However, only *Che*Shift shows a significant difference between these 2 protein models, namely $R = 0.89$ and $R = 0.94$ for proteins 1E8L and 3LZT, respectively (i.e., in agreement with the existence of significant differences between these 2 models in some regions of the structure, such as the long loop, and also involving residues in the turns between helices), while the differences obtained using SHIFTX or SPARTA are minimal, i.e., $R = 0.95$ and 0.96, and R = 0.96 and 0.97 for proteins 1E8L and 3LZT, respectively. On the other hand, SHIFTS does not discriminate between these 2 proteins ($R = 0.95$ for both 1E8L and 3LZT, respectively).

If the 8 cysteines are excluded (to make a fair comparison with the results obtained from the SHIFTS server which does not consider cysteines), the following results are obtained: (*1*) for *Che*Shift, the difference is slightly smaller than the one obtained with the cysteines included, i.e., $R = 0.91$ and 0.95 for proteins 1E8L and3LZT, respectively, although the x-ray structure remains as a much better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution; (*2*) SHIFTX does not discriminate between these protein models ($R = 0.97$); and (*3*) SPARTA provides improved agreement for both proteins, keeping the difference to a minimum, i.e., $R = 0.97$ and 0.98 for proteins 1E8L and 3LZT, respectively.

Overall, the *Che*Shift and SPARTA servers point to the same conclusion, with or without cysteines, but only *Che*Shift shows higher discriminative power, as was obtained in the analyses carried out for the 2 proteins whose structures were determined by x-ray diffraction. In other words, *Che*Shift indicates that the x-ray-derived structure, 3LZT, is a *significantly* better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the NMR-derived ensemble of 1E8L.

A Comparative Validation Analysis of Proteins 1UBQ and 1XQQ. The structure of ubiquitin solved by x-ray diffraction at 1.8 Å resolution, PDB ID 1UBQ (34), and 128 conformations obtained using NMR-derived information, PDB ID 1XQQ (35), were compared according to their ability to reproduce the observed $^{13}C^{\alpha}$ chemical shifts in solution. Among all servers (see Table S4, *Section II*, *SI Appendix*), *Che*Shift and SHIFTS predictions indicate that the NMR-derived ensemble (1XQQ) is a better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the x-ray structure (1UBQ), but only *Che*Shift shows a significant difference between them, i.e., $R = 0.95$ (NMR) and 0.91 (x-ray). Even more important, the results obtained with *Che*Shift are consistent with previous calculations (36) carried out by using the internal standard reference, indicating that the 1XQQ ensemble is a significantly better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the 1UBQ single protein model.

On the other hand, SHIFTX does not discriminate between these x-ray and NMR models, but SPARTA shows slightly better agreement between observed and predicted $^{13}C^{\alpha}$ chemical shifts for the x-ray-derived structure (1UBQ), rather than the NMR-derived ensemble (1XQQ). This result should not be surprising since the x-ray structure of ubiquitin (1UBQ) is included in the SPARTA database.

Are the Servers Sensitive Enough to Discriminate Decoys from Native Conformations? To answer this question, sets of decoys for proteins 1AIL (20), 1RNB (21), and 1UBI (22) for which the $^{13}C^{\alpha}$ chemical shifts of the proteins are available, were taken from the ROSETTA@HOME decoys set (19), and are considered here. All decoys analyzed here are close to the x-ray determined conformation, i.e., within an arbitrary rmsd cutoff of 3 Å. The x-ray-determined conformations from which the decoys were generated are termed "native" conformations here, although an x-ray structure may, or may not, be the best model with which to represent the observed $^{13}C^{\alpha}$ chemical shifts in solution (9).

If an attempt to discriminate decoys from the native structure, based only on $^{13}C^{\alpha}$ chemical shift information, would include conformations with an rmsd beyond the 3 Å cutoff value, then addition of other selection criteria, such as NOE-derived distance constraints, would be necessary because the $^{13}C^{\alpha}$ chemical shift is only a local property of the residue, i.e., a given $^{13}C^{\alpha}$ chemical shift can correspond to more than one set of backbone and side-chain torsional angles.

The ability to discriminate decoys from the native structures, using *Che*Shift, SHIFTS, SHIFTX and SPARTA for the proteins PDB ID 1AIL (20) and 1RNB (21) is illustrated in Figs. S3*A*-*D* and S4*A*-*D*, respectively (*SI Appendix*). The results show that only *Che*Shift and SPARTA are able to discriminate the decoys from the native conformation for both proteins. On the other hand, SHIFTS fails for both proteins while SHIFTX was able to discriminate the native structure of only 1AIL. Analysis of the 1UBI decoys is discussed in detail in the next subsection.

Further analysis of the results shown in Figs. S3 and S4 (*SI Appendix*) indicates that none of the servers are able to discriminate among all of the decoys, i.e., as to which one is closest to the "native" structure, indicating that, for this purpose, another scoring function is necessary.

Is the X-Ray "Native" Conformation the Best Model with Which to Represent the Observed $^{13}C^{\alpha}$ Chemical Shifts in Solution? Test on Decoys Derived from Ubiquitin (1UBI). The ability of different servers to discriminate decoys of ubiquitin from the native structure [1UBI, solved at 1.8 Å resolution (22)] is illustrated in Fig. S5*A*–*D* (*SI Appendix*). Only SHIFTS and SPARTA discriminate all of the protein decoys from the native conformation, PDB ID 1UBI. In other words, *Che*Shift and SHIFTX do not recognize the native conformation (i.e., the x-ray model) as the best representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution. This failure poses the question whether the x-ray native conformation is indeed the best structure with which to represent the observed $^{13}C^{\alpha}$ chemical shifts in solution. To answer this important question, we computed the agreement between the observed and predicted $^{13}C^{\alpha}$ chemical shifts for 10 NMR-derived, high-resolution structures of ubiquitin, namely protein PDB ID 1D3Z (37); see results in Fig. S5 (*SI Appendix*). The prediction of *Che*Shift indicates that any model from the 1D3Z ensemble is a better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the native (PDB ID 1UBI) or any protein decoy. This is not a surprising result since previous calculations carried out with the internal standard reference indicated that the 1D3Z conformations are a better representation of the observed $^{13}C^{\alpha}$ chemical shifts than a single x-ray structure (9). A similar conclusion is obtained from the analysis with the SHIFTS and SHIFTX servers, but not with SPARTA which favors the 1UBI model over any of the 1D3Z conformations. The latter is not an unexpected result because the SPARTA database contains an x-ray model of ubiquitin (1UBQ).

Conclusions

We have shown that the quantum-mechanical basis of the *Che*Shift server enables us to predict the $^{13}C^{\alpha}$ chemical shifts with reasonable accuracy in seconds and, hence, provides a standard with which to evaluate the quality of any reported protein structure solved by either x-ray crystallography or NMR-spectroscopy, provided that the experimentally observed $^{13}C^{\alpha}$ chemical shifts are available. These conclusions are supported here by an extensive analysis of a large number of x-ray-determined structures, pairs of x-ray and NMR-determined conformations, and the power to discriminated protein decoys from "native" conformations. Moreover, a detailed comparison with the results obtained for these sets of conformations using other available servers illustrates one of the main advantages of *Che*Shift predictions: these predictions are significantly more sensitive than those of any of the tested servers to conformational differences between protein models. This was verified, in most cases, by comparing *Che*Shift predictions with those obtained using the internal standard reference.

Even though the *Che*Shift server has somewhat lower sensitivity to detect subtle conformational differences between protein models than the highly-accurate internal standard reference predictions, it is a thousand times faster and, hence, it overcomes the main limitation of this purely physics-based, $^{13}C^{\alpha}$-based method (9). Even more important, the *Che*Shift-server predictions can now be adopted as a validation routine by spectroscopists and crystallographers. In fact, members of the scientific community are invited to use it by uploading their protein models on a new web server.

Materials and Methods

Experimental Set of Structures. All of the information concerning the x-ray and NMR-derived set of conformations used, as well as the BMRB accession number from which the observed $^{13}C^{\alpha}$ chemical shifts were obtained, are listed on Tables S3 and S4 (*Section II*, *SI Appendix*). It is worth noting that no $^{13}C^{\alpha}$ chemical shift reference correction (15) was applied to the experimentally observed values.

Reproducing Observed $^{13}C^{\alpha}$ Chemical Shifts of Proteins: Dependence on the Basis Set Size. Five basis sets using the *locally-dense* basis-set approximation (38) (see Table S1, *SI Appendix*), viz., 6–31G/3–21G, 6–31G(d)/3–21G, 6–311G(d,p)/3–21G, 6–311+G(d,p)/3–21G, and 6–311+G(2d,p)/3–21G, and the uniform 3–21G/3–21G basis set were initially applied to 10 NMR-derived conformations of the 76-residue α/β protein ubiquitin [PDB ID 1D3Z (37)]. The results of this analysis for 3 proteins (see Table S2, *SI Appendix*) indicate that, first, the $^{13}C^{\alpha}$ chemical shifts in proteins, computed at the DFT level of theory with the large [6–311+G(2d,p)/3–21G] basis set, can be reproduced accurately (within an average error of approximately 0.4 ppm) and approximately 9 times faster by using the small (6–31G/3–21G) basis set with an effective TMS value of 195.4 ppm and extrapolating it with: $^{13}C^{\alpha} = -1.597 + 1.040 \times {}^{13}C^{\alpha}_{\mu}$, where $^{13}C^{\alpha}_{\mu}$ represents the $^{13}C^{\alpha}$ chemical shifts computed for a given residue μ with the small basis set, and, second, the results provide evidence that the conclusions reached apply to proteins of any size or class. Moreover, in *Section I* of the *SI Appendix*, an analysis of the magnitude of the errors is provided.

Internal Standard Reference. As an internal standard reference, the values computed at the DFT level of theory by using a large basis set [6–311+G(2d,p)/3–21G], as a ''basis set limit result,'' were adopted. This arbitrary reference was chosen because this physics-based method is extremely sensitive to small conformational changes (10) and, hence, it represent a very accurate (9, 36) method with which to computed the $^{13}C^{\alpha}$ chemical shifts for a given protein structure model.

Building the CheShift Database. For the generation of the 696,916 conformations, as function of the ϕ, ψ, $\chi1$, and $\chi2$ torsional angles, for all 20 naturally occurring amino acids, the following sampling procedure was used: (*i*) the backbone torsional angles ϕ and ψ were sampled every 10°; (*ii*) all ω torsional angles were assumed to be 180°, except for Pro residues for which the *cis* conformation (0°) was also considered; (*iii*) all cysteines were considered nonbonded; (*iv*) all χ^1 side-chain torsional angles were sampled every 30°; (*v*) all χ^2 side-chain torsional angles were sampled according to the most 'frequently-seen' torsional values (39); and (*vi*) any conformation with a total ECEPP05 (40) internal energy >30 kcal/mol was rejected. This cutoff value in the total internal energy was chosen because it is large enough to cover a broad range of conformations populating the Ramachandran map, i.e., to account for the existence of conformations generated with several different force-fields used to determine x-ray and NMR-derived structures. For each of these 696,916 conformations, the $^{13}C^{\alpha}$ chemical shifts were computed using a small basis set and linearly extrapolated to the large basis set by using the above mentioned linear regression.

Approximations Used to Interpolate Computed $^{13}C^{\alpha}$ Chemical Shift Values. Since the database is a $^{13}C^{\alpha}$ chemical shift coarse-grained representation of the continuum variable space in ϕ, ψ, $\chi1$, and $\chi2$ torsional angles, an accurate interpolation method must be used to compute $^{13}C^{\alpha}$ chemical shift for any arbitrary combination of these 4 torsional angles. Among all possible options that do not require adjustable parameters, we tested 2: a Gaussian (41) and a linear interpolation, respectively (see *Section III*, *SI Appendix*). To decide whether the Gaussian or the linear interpolation provides a more accurate representation of the $^{13}C^{\alpha}$ chemical shift hypersurface, the following test was carried out. An arbitrary selected fraction of the accessible torsional angle space ($\phi = [-150°, -160°]$; $\psi = [160°, 170°]$; and $\chi1 = [-180°, 180°]$) for the tripeptide Ac-GXG-NMe, with X = Ser, was sampled by using a fine grid, namely 2° for ϕ and ψ backbone torsional angles, and 5° for the $\chi1$ side-chain torsional angle. For this fine grid, the $^{13}C^{\alpha}$ chemical shift map was computed using a small basis set and linearly extrapolated to a large basis set. The ability of the Gaussian and linear interpolations to reproduce these fine-grid values by using the corresponding coarse grain data, namely at 10° steps for ϕ and ψ backbone torsional angles, and 30° for $\chi1$, was analyzed graphically (see Fig. S6, *SI Appendix*) and by the frequency of the error distribution (see Fig. S7, *SI Appendix*). Both, the graphic analysis and the standard deviation of the frequency of the error distribution indicate that the linear interpolation is a significantly better approximation ($\sigma = 0.09$ ppm for the fine grid mesh results) than the Gaussian interpolation ($\sigma = 0.27$ ppm) and, hence, the linear interpolation was adopted. The accuracy of the linear interpolation to reproduce the values obtained for the fine-grid mesh is due to the small-torsional-angle variations chosen to build the coarse-grained *Che*Shift database which captures the most important dependence of the $^{13}C^{\alpha}$ chemical shift on the backbone and side-chain torsional angles.

ACKNOWLEDGMENTS. We thank Dr. H.A. Baldoni of the Universidad Nacional de San Luis, Instituto de Matemática Aplicada San Luis, Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina for suggestions and discussion regarding the material presented in *Section I* of the *SI Appendix*. We also thank all spectroscopists, crystallographers, and those in charge of both the PDB and the BRMB databases because, without their effort, this work would not be possible. 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.

1. Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins* 66:778–795.
2. Billeter M, Wagner G, Wüthrich K (2008) Solution NMR structure determination of proteins revisited. *J Biomol NMR* 42:155–158.
3. Williamson MP, Craven CJ (2009) Automated protein structure calculation from NMR data. *J Biomol NMR* 43:131–143.
4. Vriend G (1990) WHAT IF: A molecular modeling and drug design program. *J Mol Graphics* 8:52–56.
5. Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356:83–85.
6. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26:283–291.
7. Huang YJ, Powers R, Montelione GT (2005) Protein NMR Recall, Precision, and F-measure scores (RPF scores): Structure quality assessment measures based on information retrieval statistics. *J Am Chem Soc*, 127:1665–1674.
8. Davis IW, et al. (2007) MolProbity: All atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35:W375–W383.
9. Vila JA, Scheraga HA (2009) Assessing the accuracy of protein structures by quantum mechanical computations of $^{13}C^{\alpha}$ chemical shifts. *Acc Chem Res*, in press.
10. Arnautova YA, Vila JA, Martin OA, Scheraga HA (2009) What can we learn by computing $^{13}C^{\alpha}$ chemical shifts for X-ray protein models? *Acta Crystallogr D* 65:697–703.
11. Xu X-P, Case DA (2001) Automated prediction of ^{15}N, $^{13}C^{a}$, $^{13}C^{b}$ and $^{13}C'$ chemical shifts in proteins using a density functional database. *J Biomol NMR* 21:321–333.
12. Xu X-P, Case DA (2002) Probing multiple effects on ^{15}N, $^{13}C^{a}$, $^{13}C^{b}$ and $^{13}C'$ chemical shifts in peptides using density functional theory. *Biopolymers* 65:408–423.
13. Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1H, 13C, and 15N chemical shifts. *J Biomol NMR* 26:215–240.
14. Meiler J (2003) PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J Biomol NMR* 26:25–37.
15. Shen Y, Bax Ad (2007) Protein backbone chemical shifts predicted from searching a database for torsional angle and sequence homology. *J Biomol NMR* 38:289–302.
16. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690.
17. Vila JA, et al. (2008) Quantum chemical $^{13}C^{\alpha}$ chemical shift calculations for protein NMR structure determination, refinement, and validation. *Proc Natl Acad Sci USA* 105:14389–14394.
18. Ramelot TA, et al. (2008) Improving NMR protein structure quality by Rosetta refinement: A molecular replacement study, *Proteins* 75:147–167.
19. All Atom Decoy Sets from Rosetta@home. Available at http://depts.washington.edu/bakerpg/, 2007. Accessed on May 2009.
20. Liu J, et al. (1997) Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. *Nat Struct Biol* 4:896–899.
21. Baudet S, Janin J (1991) Crystal structure of a barnase-d(GpC) complex at 1.9 Å resolution. *J Mol Biol* 219:123–132.
22. Ramage R, et al. (1994) Synthetic, structural and biological studies of the ubiquitin system: The total chemical synthesis of ubiquitin. *Biochem J* 299:151–158.
23. Berman HM, et al. (2000) The protein data bank. *Nucleic Acids Res* 28:235–242.
24. Press HW, Teukolsky SA, Vetterling WT, Flannery BP (1992) in *Numerical Recipes in FORTRAN 77. The Art of Scientific Computing*, 2nd Ed, (Cambridge Univ Press, Cambridge, UK), pp 630–633.
25. Sevcik J, Dauter Z, Lamzin VS, Wilson KS (1996) Ribonuclease from streptomyces aureofaciens at atomic resolution. *Acta Crystallogr D* 52:327–344.
26. Havlin RH, Le H, Laws DD, deDios 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.
27. Iwadate M, Asakura T, Williamson MP (1999) C^{α} and C^{β} carbon-13 chemical shifts in protein from an empirical database. *J Biomol NMR* 13:199–211.
28. 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.
29. Veerapandian B, et al. (1992) Functional implications of interleukin-1β based on three-dimensional structure. *Proteins* 12:10–23.
30. Priestle JP, Chär H-P, Grütter MG (1989) Crystallographic refinement of interleukin 1β at 2.0 Å resolution. *Proc Natl Acad Sci USA* 86:9667–9671.
31. Ulrich EL, et al. (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408.
32. Walsh MA, et al. (1998) Refinement of triclinic hen egg-white Lysozyme at atomic resolution. *Acta Crystallogr D* 54:522–546.
33. Schwalbe H, et al. (2001) NMR solution structure of hen Lysozyme. *Protein Sci* 10:677–688.
34. Vijay-Kumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. *J Mol Biol* 194:531–544.
35. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433:128–132.
36. Vila JA, Villegas ME, Baldoni HA, Scheraga HA (2007) Predicting $^{13}C^{\alpha}$ chemical shifts for validation of protein structures. *J Biomol NMR* 38:221–235.
37. Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc* 120:6836–6837.
38. Chesnut DB, Moore KD (1989) Locally dense basis-sets for chemical-shift calculations. *J Comp Chem* 10:648–659.
39. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. *Proteins* 40:389–408.
40. Arnautova YA, Jagielska A, Scheraga HA (2006) A new force field (ECEPP05) for peptides proteins and organic molecules. *J Phys Chem B* 110:5025–5044.
41. Kuszewski J, Qin J, Gronenborn AM, Clore MG (1995) The impact of direct refinement against $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts on protein structure determination by NMR. *J Magn Reson B* 106:92–96.

Supporting Information Appendix

Section I

Dependence of the accuracy and speed of DFT calculations of the $^{13}C^{\alpha}$ chemical shifts of proteins on the size of the basis set used.

The purpose of this section is to study the dependence of the accuracy and speed of DFT calculations of the $^{13}C^{\alpha}$ chemical shifts on the size of the basis set used. Six basis sets (see Table S1), viz., five *locally-dense* basis-set approximation 6-31G/3-21G, 6-31G(d)/3-21G, 6-311G(d,p)/3-21G, 6-311+G(d,p)/3-21G, and 6-311+G(2d,p)/3-21G, and the uniform 3-21G/3-21G basis set were initially applied to 10 NMR-derived conformations of the 76-residue α/β protein ubiquitin (Protein Data Bank id 1D3Z[1]). For each of these six basis sets, combined with the OB98 functional,[13] the $^{13}C^{\alpha}$ shielding was computed for 760 amino acid residues by treating each amino acid **X** in the sequence as a terminally-blocked tripeptide with the sequence Ac-G**X**G-NMe in the conformation of the regularized experimental protein structure. Analysis of the results (see Table S1), in terms of the agreement between the computed and observed $^{13}C^{\alpha}$ chemical shifts, shows that the accuracy, with which the observed $^{13}C^{\alpha}$ chemical shifts are reproduced by using either the small basis set (6-31G/3-21G) or the larger basis set [6-311+G(2d,p)/3-21G], is very similar, although use of the small basis set leads to a significant decrease in computational time. An additional analysis was carried out here for: (*a*) two other proteins with different numbers of residues and topology solved by NMR spectroscopy and X-ray diffraction (PDB id 2JVD[2]; 1NS1[3] and 3HBP[4]; 1AIL[5]), and (*b*) Val and Arg hypersurfaces constructed by calculating a grid of 6,864 and 6,794 points, respectively, corresponding to different combinations of the ϕ, ψ, $\chi1$ (and $\chi2$ only for Arg) torsional angles. The results of this analysis, reported in Table S2, and Figures S1 provide evidence that the conclusions derived here apply to proteins of any size or class (see section *Transferability of the results*, below). The results also indicate that the $^{13}C^{\alpha}$ chemical shifts computed with the small basis set (6-31G/3-21G), and extrapolated by an empirically-determined linear regression formula to reproduce the values obtained with a larger basis set [6-311+G(2d,p)/3-21G], constitute an adequate compromise between accuracy (within the average error of ~0.4 ppm) and computational cost (~9 times faster) with which to reproduce the observed $^{13}C^{\alpha}$ chemical shifts of proteins in solution.

Method used to compute the $^{13}C^{\alpha}$ chemical shifts. All the experimentally determined conformations were *regularized*,[6] i.e., all residues were replaced by the standard ECEPP/3[7] residues in which bond lengths and bond angles are fixed (rigid-body geometry approximation) at the standard values,[7] and hydrogen atoms are added, if necessary.

The computations of the $^{13}C^{\alpha}$ chemical shifts involve a series of approximations. For each amino acid residue **X** in the protein sequence: (*a*) it is assumed that the observed $^{13}C^{\alpha}$ chemical shift is a conformational-averaged one (see *Computation of the conformationally-averaged rmsd* section); (*b*) computation of the $^{13}C^{\alpha}$ shielding was carried out on a terminally-blocked tripeptide with the sequence Ac-G**X**G-NMe in the conformation of the regularized experimental protein structure; (*c*) computation of the $^{13}C^{\alpha}$ shielding for each residue **X** was carried out with a Γ *locally-dense* basis set approach,[8] with Γ = 6-31G, 6-31G(d), 6-311G(d,p), 6-311+G(d,p) or 6-311+G(2d,p), while the remaining residues in the tripeptide were *always* treated with the 3-21G basis set. From here on, this combination of uniform and locally-dense basis sets is referred to as: Set_1, Set_2, Set_3, Set_4, Set_5 and Set_6, respectively; (*d*) all ionizable residues were considered neutral during the gas-phase quantum chemical calculations;[9] (*e*) no geometry optimization is necessary since such optimization by *ab-initio* (HF) or DFT methods has only a small effect on the computed chemical shifts;[10] (*f*) the computed $^{13}C^{\alpha}$ shieldings ($\sigma_{\mathrm{subst},th}$) were converted to $^{13}C^{\alpha}$ chemical shifts (δ) by employing the equation $\delta_{th} = \sigma_{\mathrm{ref}} - \sigma_{\mathrm{subst},th}$ where the indices denote a theoretical (*th*) computation, the reference substance (*ref*), and the substance of interest (*subst*), i.e., the $^{13}C^{\alpha}$ shielding of a given amino acid residue **X**. The observed shielding value of tetramethylsilane (TMS) in the gas phase[11], namely 188.1 ppm, was adopted as a reference value.

All the computed $^{13}C^{\alpha}$ shielding ($\sigma_{\mathrm{subst},th}$) values were calculated using the gauge-invariant atomic orbital (GIAO) method at the DFT level of theory as implemented in the GAUSSIAN 03 suite of programs.[12] We have used only one exchange-correlation functional in this section, namely OB98, because it was shown that this functional is, among others, one of the most accurate *and* faster one with which to reproduce the observed $^{13}C^{\alpha}$ chemical shift of proteins in solution.[13]

Determination of an effective TMS shielding value. The determination of an *effective* TMS shielding value follows the procedure introduced recently[13] and, hence, only a brief description will be offered here. By adopting the observed TMS value of 188.1 ppm as a reference, it is possible to find the characteristic mean (x_{o}) and standard deviation (σ) of the Normal (or Gaussian) fit of the frequency of the error distribution for each of the six basis sets. In other words, for each basis set it is feasible to find an 'effective' TMS shielding value for which the Normal (or Gaussian) fit shows a zero displacement, i.e., an *effective* TMS value that gives $x_o = 0.0$. The following *effective* TMS values were obtained by applying this procedure and are used throughout this work: 195.4 ppm and 184.5 ppm,[13] for the small (Set_2) and larger (Set_6) basis sets, respectively.

Computation of the conformationally-averaged rmsd. Although the methodology has been published in several previous papers[4,6,9,13] we reproduce it here for the reader's convenience. A protein in solution exists as an ensemble of conformations. As a consequence, 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 from which the conformations were derived.[6] Thus, the following quantity can be computed: $^{13}C^{\alpha}{}_{\text{computed},\mu} = \sum_{i=1}^{\Omega} \lambda_i \ {}^{13}C^{\alpha}{}_{\mu,i}$, where $^{13}C^{\alpha}{}_{\mu,i}$ is the computed chemical shift for amino acid μ in conformation i out of Ω protein conformations, 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 assumptions, the computation of the *ca*-rmsd for a protein containing N amino acids residues, is straightforward:[26] $ca\text{-rmsd}^{\alpha} = (1/N) \sum_{\mu=1}^{N} ({}^{13}C^{\alpha}{}_{\text{observed},\mu} - \langle {}^{13}C^{\alpha}{}_{\text{computed},\mu} \rangle)^2]^{½}$ with $\langle {}^{13}C^{\alpha}{}_{\text{computed}} \rangle_{\mu} = (1/\,\Omega) \sum_{i=1}^{\Omega} {}^{13}C^{\alpha}{}_{\mu,i}$. Naturally, if $\Omega = 1$, *ca*-rmsd ≡ rmsd, as for any single structure.

In addition, for each amino acid μ, we define an error function $\Delta^{\alpha}{}_{\mu} \cong ({}^{13}C^{\alpha}{}_{\text{observed},\mu} - <{}^{13}C^{\alpha}{}_{\text{computed}}>_{\mu})$.

Computation of the average CPU time. The average computational time (reported in Table S1) was computed as an average over all 76 residues of conformation 1 out of 10 of 1D3Z:

$$\textit{Averaged-CPU time} = (1/N) \sum_{\mu=1}^{N} \mathrm{T}_{\mu} \text{ , with } N = 76$$

where T_{μ} represents the total cpu time (in seconds) for residue μ, as reported by the output file of the GAUSSIAN 03 suite of programs.[12]

Transferability of the results. The current methodology[4,6,9,13] relies on a crucial observation that once the residue conformations are established by their interactions with the rest of the protein, the ${}^{13}C^{\alpha}$ shielding of each residue depends, mainly, on its backbone and its side-chain conformation, with no significant influence of either the amino acid sequence or the position of the given residue in the sequence. This observation allows us to parallelize the ${}^{13}C^{\alpha}$ shielding calculations in proteins and, hence, to make them feasible. In addition, this means that a given set of accurately-determined amino acid residue conformations, representing the accessible conformational space for all the 20 naturally occurring amino acids and showing a good distribution of side-chain conformations, will constitute a reasonable ensemble with which to carry out tests of the current methodology. In other words, the results of such tests will not depend on whether such ensembles of residue conformations belong to a single or many proteins and, therefore, the results should be transferable to proteins of any class or size. For this purpose, three proteins solved by NMR and X-ray were chosen (see Table S2). The analysis of the three proteins includes information for *all* the 20 naturally occurring amino acid residues with their backbone torsional angles populating the α-helical, β-sheet, turn and extended regions of the Ramachandran map.

Statistical analysis for the six basis sets. For each basis set, we compute the correlation coefficient,[14] *R* (or *Pearson* coefficient), between the average, $<^{13}C^{\alpha}_{\mathrm{computed}}>_{\mu}$, chemical shift calculated for the 10 conformations of 1D3Z (as described in *Computation of the conformationally-averaged rmsd* section) and the observed $^{13}C^{\alpha}$ chemical shifts, the standard deviation of the correlation and the average CPU-time (see Table S1). Adopting the *R* value obtained with the largest basis set, i.e., Set_6, as a 'basis set limit result' enables us to conclude that the results obtained with a small basis set, Set_2, appear as the best tradeoff between accuracy and speed of the calculations. In other words, use of the small basis set leads to comparable correlation, in terms of *R*, to that obtained with the larger basis set but at a significantly lower computational cost (3,268/363 ~9 times; see Table S2). Hence, from here on, our work will be focused on a comparison of the results obtained with a small basis set (Set_2), rather than a large (Set_6) basis set.

The previous analysis enabled us to select the smaller basis set that provides similar accuracy as a 'basis set limit', to reproduce the computed shielding rather than chemical shifts, but at a significantly lower CPU time. However, such analysis says nothing about the accuracy of the $^{13}C^{\alpha}$ chemical shifts, not the shielding, obtained with the Set_2. The answer to this important question is provided in the next section.

Comparison of the results obtained with basis Set_2 and Set_6. An analysis of the correlation of the results obtained from Set_2 and Set_6 was carried out for 3 proteins, namely, 1D3Z (10 conformations), 2JVD (20 conformations), 1NS1 (32 conformations) and for 6,864 and 6,794 points corresponding to different combinations of the ϕ, ψ, $\chi 1$, and $\chi 2$ torsional angles for Arg and Val, respectively. The results of the analysis are shown in Figure S1a-e. For all the molecules shown in Figure S1, the slopes of the linear regression are very similar and, even more importantly, very close to the ideal value of 1.0. The largest difference appears on the *y* intercept ($-2.23 < y < -0.620$) of the linear regression. Nevertheless, the $^{13}C^{\alpha}$ chemical shifts computed with Set_6 can be obtained from Set_2 by using the following linear regression: $^{13}C^{\alpha} = -1.597 + 1.040 \times \mathbf{^{13}}C^{\alpha}_{\mu}$, where $^{13}C^{\alpha}_{\mu}$ represents the $^{13}C^{\alpha}$ chemical shifts computed for a given residue μ with the Set_2 and, −1.597 and 1.040 representing the averaged values over the five linear regressions, namely from each panel of Figure S1, for both the *y* intercept and the slope of the regression, respectively.

The main goal of this section I is to determine the basis set size that would enable us to compute the $^{13}C^{\alpha}$ chemical shifts in proteins accurately and fast. Once the small basis set has been chosen, namely Set_2, the next step is to determine how accurately the extrapolated values of the $^{13}C^{\alpha}$ chemical shifts computed with such basis set reproduce the 'basis set limit' results, i.e. the results obtained with the larger basis set (Set_6). The results of such an analysis are shown in Table S2. The accuracy of the results is evaluated here in terms of the *ca*-rmsd. As can be seen from Table S2, the quality of the protein structures in terms of these scoring parameters, computed by using either Set_2 or Set_6, is comparable. In other words, extrapolating the $^{13}C^{\alpha}$ chemical shifts computed with Set_2 enables us to reproduce the results obtained with the more computationally expensive basis set (Set_6) with high accuracy.

In Table S2, we also list the average error (Δ) obtained between the $^{13}C^{\alpha}$ chemical shifts computed with the small basis set (after extrapolation by using the linear relationship: $^{13}C^{\alpha} = -1.597 + 1.040 \times \mathbf{^{13}}C^{\alpha}_{\mu}$) and the values obtained with the large basis set. Notably, the average error (Δ) among all the proteins listed in Table S2 is quite low, namely ~0.4 ppm.

The results of this Section I indicate that the ^{13}C$^{\alpha}$ chemical shifts in proteins, computed at the DFT level of theory with the large (Set_6) basis set, can be reproduced accurately (within an average error of ~0.4 ppm; see Table S2) and ~9 times faster by using the small (Set_2) basis set with an *effective* TMS value of 195.4 ppm and extrapolating it with: $^{13}C^{\alpha} = -1.597 + 1.040 \times \mathbf{^{13}}C^{\alpha}_{\mu}$.

Table S1

Test of Six Basis Sets[a]

Basis sets [b]	Correlation Coefficient R[c]	Average CPU time[d] (sec)
Set_1: 3-21G/3-21G	0.892 (2.13)	201
Set_2: 6-31G/3-21G	**0.903 (2.04)**	**230** **(363; 118)**
Set_3: 6-31G(d)/3-21G	0.894 (2.12)	338
Set_4: 6-311G(d,p)/3-21G	0.900 (2.06)	568
Set_5: 6-311+G(d,p)/3-21G	0.903 (2.03)	1,092
Set_6: 6-311+G(2d,p)/3-21G	**0.908 (1.97)**	**1,535** **(3,268; 372)**

a) The entire test was carried out with 10 conformations of the protein Ubiquitin (PDB id 1D3Z).[1] Two basis sets, selected from the results of column 2 and 3 for further test (see results in Table S2), are represented in boldface.

b) As described in the *Method used to compute the $^{13}C^{\alpha}$ chemical shifts* section.

c) The correlation coefficient,[14] R (or *Pearson* coefficient), between the average chemical shifts, $<^{13}C^{\alpha}_{computed}>_{\mu}$ computed from the 10 conformations of 1D3Z, and the observed $^{13}C^{\alpha}$ chemical shifts. The standard deviation of the correlation is in parenthesis.

d) As an average over 76 residues computed from model 1 out of 10 models of 1D3Z,[1] as explained in the *Computation of the average CPU time* section. The maximum and minimum CPU times of the two selected basis set are shown in parentheses in boldface.

Table S2

Test on Proteins Structures with Two Selected Basis Sets[a]

Protein[b]		Basis sets		
		Set_6	Set_2	
		ca-rmsd[c] (ppm)	*ca*-rmsd[d] (ppm)	Δ[e] (ppm)
Ubiquitin (6457)	1UBQ (X-ray) [76] {1;*i*}	2.60	2.57	0.63±0.40
YnzC protein (15476)	2JVD (NMR) [46] {20; *ii*}	1.64	1.59	0.36±0.22
	3HBP (X-ray) [52] {1;*iii*}	2.51	2.55	0.34±0.22
		1.86	1.85	0.37±0.20
		1.88	2.00	0.40±0.21
Non-structural Protein 1 (15117)	1NS1 (NMR) [73] {32; *iv*}	2.48	2.47	0.26±0.17
	1AIL (X-ray) [70] {1; *v*}	2.07	2.09	0.34±0.29

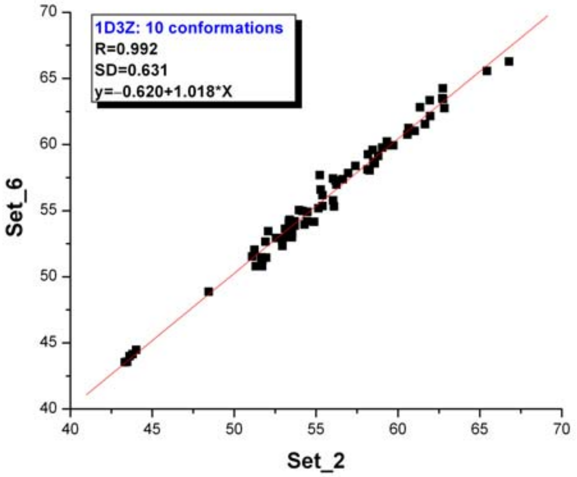
a) Carried out for the two highlighted basis sets in Table S1, namely Set_6 and Set_2.

b) First column list the set of proteins used and, in parenthesis, the BMRB[24] accession number under which the observed $^{13}C^{\alpha}$ Chemical shifts can be found. Second column, in parentheses, the experimental method used; in brackets, the numbers of residues for each protein; and, in braces, the number of conformations and the reference for the protein, namely {*i*} Vijay-Kumar *et al.*[17]; {*ii*} Aramini *et al.*[2]; {*iii*} Vila *et al.*[4]; {*iv*} Chien *et al.*[3]; {*v*} Liu *et al.*[5]

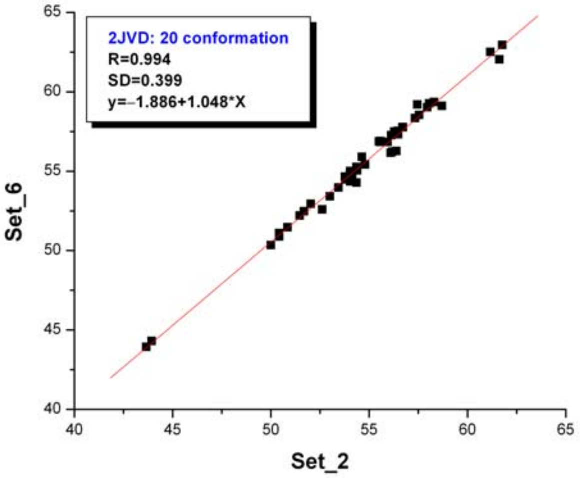
c) The *ca*-rmsd, computed as described in *Computation of the conformationally-averaged rmsd* section, with Set_6 using an *effective* TMS value of 184.5 ppm. Note that the *ca*-rmsd ≡ rmsd for the single X-ray conformations.

d) Same as item (c) but with *all* the $^{13}C^{\alpha}$ chemical shifts given by: $^{13}C^{\alpha} = -1.597 + 1.040 \times \mathbf{^{13}}C^{\alpha}_{\mu}$ where $^{13}C^{\alpha}_{\mu}$ represent the $^{13}C^{\alpha}$ chemical shifts computed for the residue μ with a small (Set_2) basis set, and using an *effective* TMS value of 195.4 ppm.

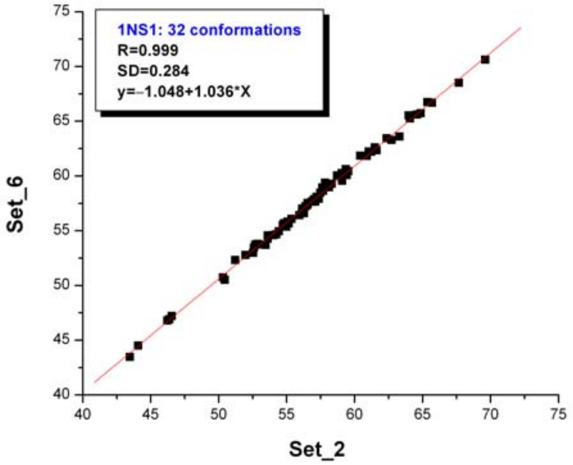
e) The value of the absolute averaged error per-residue, Δ, between the $^{13}C^{\alpha}$ chemical shifts computed with the small basis set (after extrapolation by using the linear relationship) and the values obtained with the large basis set.


1D3Z: 10 conformations
R=0.992
SD=0.631
y=-0.620+1.018*X
Set_6
Set_2
40
45
50
55
60
65
70

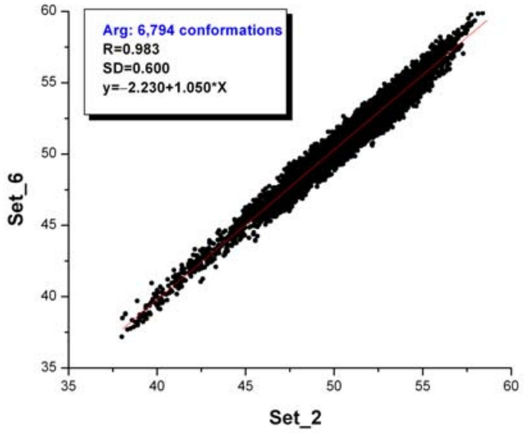
(a)


2JVD: 20 conformation
R=0.994
SD=0.399
y=-1.886+1.048*X
Set_6
Set_2
40
45
50
55
60
65

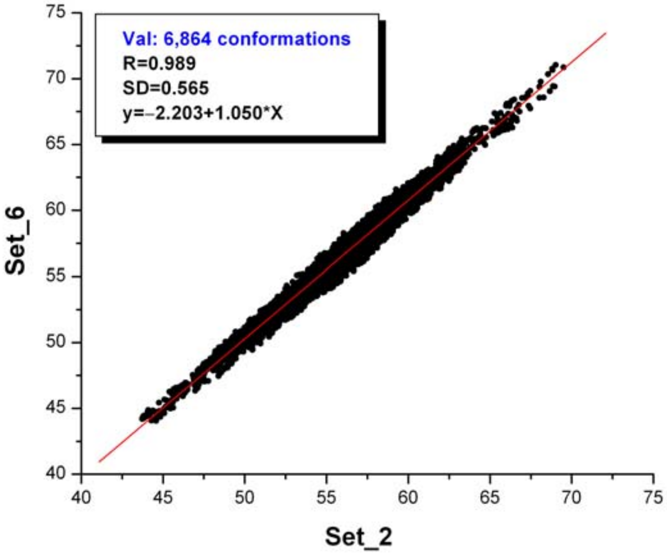
(b)


1NS1: 32 conformations
R=0.999
SD=0.284
y=-1.048+1.036*X
Set_6
Set_2
40
45
50
55
60
65
70
75

(c)


Arg: 6,794 conformations
R=0.983
SD=0.600
y=-2.230+1.050*X
Set_6
Set_2
35
40
45
50
55
60

(d)



(e)

Figure S1. (a) Correlation between average chemical shifts, $<^{13}C^{\alpha}_{computed}>_{\mu}$, computed from the 10 conformations of 1D3Z[1] with Set_6 versus Set_2. The red line represents the linear regression. Values for the correlation coefficient[14] (*R*), the standard deviation from the linear regression (SD) and the slope and the *y*-intercepts of the linear regression are inserted in the panel; **(b)** same as (**a**) for 20 conformations of 2JVD[2]; **(c)** same as (**a**) for 32 conformations of 1NS1[3]; **(d)** same as (**a**) for 6,864 points of Arg obtained by sampling the ϕ, ψ, χ1, and χ2 Ramachandran space; and **(e)** same as (**d**) for 6,794 points for Val computed obtained by sampling the ϕ, ψ, χ1 Ramachandran space.

Section II

Analysis of X-ray and X-ray-NMR pairs of structures

In this section, we provide information about the quality of the prediction, in terms of the correlation coefficient[14] (*R*), between observed and predicted $^{13}C^{\alpha}$ chemical shifts from different databases, for a set of X-ray derived structures (listed in Table S3), and for a set of pairs of X-ray and NMR-determined structures (listed in Table S4). The analysis of the set listed in Table S3 was carried out with five databases, namely SHIFTS,[18,19] SHIFTX,[20] PROSHIFT,[21] SPARTA[22] and *Che*Shift. PROSHIFT predictions were not carried out for NMR-derived ensembles, because this database web server provides predictions for *only* one structure at a time, making the analysis of a large number of structures very tedious, e.g., as for PDB id 1XQQ[23] containing 128 conformers.

Table S3

Set of X-ray Structures and Corresponding $^{13}C^{\alpha}$ Chemical Shifts

PDB id[a] [Resolution (Å); BMRB accession]		SERVERS[b]				
		*Che*Shift	SHIFTX	SPARTA	SHIFTS	PROSHIFT
1A6K [1.10; 4061]		0.94	0.97	0.97	0.96	0.97
1BKF [1.60; 4077]		0.93	0.98	0.99	0.96	0.96
1CEX [1.00; 4101]		0.95	0.98	0.99	0.98	0.96
1CLL [1.70; 547]		0.91	0.97	0.99	0.97	0.98
1DMB [1.80; 4354]		0.93	0.98	1.00	0.96	0.97
1RGE[c] **[1.15; 4259]**		0.95	0.98	0.99	0.97	0.97
		0.96	**1.00**	**1.00**	**0.98**	**1.00**
1HFC [1.56; 4064]		0.94	0.97	0.99	0.95	0.97
1HKA [1.50; 4299]		0.94	0.97	0.99	0.95	0.97
1ONC [1.7 ;4371]		0.90	0.96	0.99	0.95	0.95
1HCB [1.60; 4022]		0.92	0.97	0.98	0.96	0.96
1RUV [1.30; 4031]		0.91	0.96	0.99	0.95	0.95
1TOP [1.78; 4401]		0.95	0.97	(0.97	0.95	0.96
3LZT [0.92; 4562]		0.94	0.96	0.97	0.95	0.96
4FGF[d] [1.60; 4091]		0.93 / 0.95	0.98	0.99	0.98	0.97
Interleukin 1β[e] **(human)**	4I1B [2.00;1061]	**0.87**	0.95	0.97	0.95	0.96
	2I1B [2.00;1061]	0.91	0.96	0.98	0.94	0.96
5PTI [1.00; 46]		0.95	0.97	1.00	0.98	0.97

PDB id[a] [Resolution (Å); BMRB accession]	SERVERS[b]				
	*Che*Shift	SHIFTX	SPARTA	SHIFTS	PROSHIFT
1AIL [1.90; 4317]	0.96	0.99	1.00	0.98	0.98
1BSY [2.24; 10010]	0.90	0.95	0.97	0.93	0.95
1BV1 [2.00; 4417]	0.94	0.97	0.98	0.97	0.97
1CHN [1.76; 4083]	0.95	0.96	0.97	0.95	0.96
1EKG [1.80; 4342]	0.94	0.97	1.00	0.96	0.96
1FIL [2.00; 4082]	0.92	0.97	0.99	0.97	0.97
1GSV [1.75; 6321]	0.94	0.98	0.98	0.96	0.97
1HB6 [2.00; 5351]	0.92	0.95	0.95	0.94	0.96
1HOE [2.00; 1642]	0.94	0.97	0.97	0.95	0.97
1I27 [1.02; 5685]	0.92	0.98	0.98	0.96	0.97
1JF4 [1.40;4083]	0.95	0.97	0.97	0.97	0.98
1RBV [1.80; 4012]	0.94	0.98	0.98	0.97	0.98
1RNB [1.90; 7126]	0.93	0.96	0.98	0.96	0.95
1UBI [1.80; 5387]	0.91	0.97	0.99	0.97	0.97
1ZON [2.00; 4553]	0.92	0.96	0.98	0.95	0.96
2CI2 [2.00; 4974]	0.92	0.98	0.99	0.96	0.97
2CPL [1.63; 2208]	0.97	0.98	1.00	0.97	0.97
2OVO [1.50; 5473]	0.93	0.97	0.97	0.97	0.97
3ICB [2.30; 6699]	0.93	0.97	0.98	0.96	0.97

a) PDB ID code identifying the X-ray structure with which the correlations, *R*, between observed and predicted $^{13}C^{\alpha}$ chemical shifts (otherwise noted) were used for each of the servers. In brackets the resolution at which the protein was solved, and the BMRB[24] accession number under which the observed $^{13}C^{\alpha}$ Chemical shifts can be found. Details for each of the mentioned structures can be found in the cited manuscript in the PDB web site.

b) Server name used for automatic prediction of the $^{13}C^{\alpha}$ chemical shifts. The agreement is analyzed here in terms of the correlation coefficient *R*. The *R* values are reported with only two digits to facilitate a fast comparison of the results among servers. All the *R* values from the servers *Che*Shift, SHIFTX, SPARTA and SHIFTS include all residues, except the first and last one for *Che*Shift, SPARTA and SHIFTS. All correlations computed with SHIFTS do not include predictions for the first and last one as well as for cysteines. If the inclusion/exclusion of cysteines leads to a significant difference in the *R* value computed with the *Che*Shift server, as with protein 4FGF, two correlation coefficients are reported (see footnote *d*, below). All *R* values lower than an arbitrary selected cut off value of 0.90 are highlighted in red because they indicate that more than 20% of the observed $^{13}C^{\alpha}$ chemical shifts cannot be explained by a given protein model and, hence, further analysis may be required.

c) See discussion about the results for this protein in section: *Protein 1RGE (Ribonuclease Sa)*, of the main text. First row shows the correlation coefficient, *R*, between observed and predicted $^{13}C^{\alpha}$ chemical shifts. Second row shows, in bold face and green color, the correlation coefficient, *R*, between $^{13}C^{\alpha}$ chemical shift predictions for molecule A and B, respectively, of 1RGE, for each of the servers.

d) There are two values of *R* listed for prediction by the *Che*Shift server. In the first one, all residues were included and, in the second one, *all* cysteines were omitted from the calculations of *R* (as with the SHIFTS calculations).

e) The first and second rows of this entry list the *R* values, obtained from each server, for protein 4I1B and 2I1B, respectively. For a discussion of the results obtained for this protein, see section: *Protein interleukin 1β (human)*, of the main text.

Table S4

Pairs of X-ray Structures and sets of NMR Structures

PDB id[a]	BMRB[24] accession	SERVERS[b]			
		*Che*Shift	SHIFTX	SPARTA	SHIFTS
1A6K 1MYF(12)	4061	0.94 0.95	0.97 0.96	0.97 0.96	0.96 0.96
1BKF 1FKR(20)	4077	0.93 0.95	0.98 0.97	0.99 0.97	0.96 0.96
1CLL[c] 2BBN(21)	1634	0.92 0.94	0.97 0.98	0.98 0.98	0.96 0.97
1DMB 1EZP(10)	4354	0.93 0.91	0.98 0.96	1.00 0.97	0.96 0.95
1RGE 1C54(20)	4259	0.95 **0.89**	0.98 0.95	0.99 0.95	0.97 0.94
1HFC 4AYK(30)	4064	0.94 0.91	0.97 0.96	0.99 0.98	0.95 0.95
1HKA 1EQ0(20)	4299	0.94 0.91	0.97 0.96	0.99 0.97	0.95 0.95
1ONC 1PU3(20)	4371	0.90 **0.88**	0.96 0.95	0.99 0.97	0.95 0.95
1RUV 2AAS(32)[d]	4031	0.91 / 0.92 **0.89** / 0.91	0.96 / 0.96 0.94 / 0.96	0.99 / 0.99 0.96 / 0.97	0.95 0.94
1TOP 1SKT(40)	4401	0.95 0.92	0.97 0.97	0.97 0.97	0.96 0.96
3LZT **1E8L(50)**[e]	4562	0.94 / 0.95 **0.89** / 0.91	0.96 / 0.97 0.95 / 0.97	0.97 / 0.98 0.96 / 0.97	0.95 0.95
4I1B / 2I1B 7I1B (32)[f]	1061	**0.87** / 0.91 0.90	0.95 / 0.96 0.96	0.97 / 0.98 0.97	0.95 / 0.94 0.95
5PTI 1UUA(20)	46	0.95 0.93	0.97 0.96	1.00 0.97	0.98 0.96
2CI2 3CI2 (20)	4974	0.92 0.93	0.98 0.97	0.99 0.98	0.96 0.96
1UBQ[g] **1XQQ (128)**	6457	0.91 0.95	0.98 0.98	0.99 0.98	0.97 0.98
2B95(20)[h] 1TGQ (40)	6210	0.93 **0.87**	0.95 0.91	0.97 0.95	0.97 0.95

a) PDB ID code identifying the X-ray- and NMR-derived conformations, in the first and second lines, respectively, for each protein. The total number of structures in the NMR ensemble for which the conformational-average was computed for each database is given in parentheses.

b) Server name used for automatic prediction of the $^{13}C^{\alpha}$ chemical shifts. The agreement is analyzed here in terms of the correlation coefficient, *R*,[14] between observed and predicted $^{13}C^{\alpha}$ chemical shifts. As with Table S3, only two significant digits are reported for each R value to facilitate the comparison among the servers. All *R* values lower than an arbitrary selected cut off, namely 0.90, are highlighted in red because they indicate that more than 20% of the observed $^{13}C^{\alpha}$ chemical shifts cannot be explained by a given protein model and, hence, further analysis may be required.

c) The observed $^{13}C^{\alpha}$ chemical shifts for the protein Calmodulin (BMRB[24] accession 1634) were obtained by NMR for Calmodulin complexed with a 26-residue synthetic peptide, while the X-ray structure was obtained for a peptide-free Calmodulin at 1.7 Å resolution. There are conformational differences between the X-ray (1CLL) and NMR-derived structures (2BBN) along the whole sequence but the larger differences occur for the long central helix (residues 65-93, in the X-ray structure) which is disrupted into two helices connected by a long flexible loop in the NMR-determined conformations (2BBN).

d) The solution NMR-derived conformations for RNase A (2AAS) and the observed $^{13}C^{\alpha}$ chemical shifts (BRMB accession 4031) are for the wild-type protein. However, the X-ray structure (1RUV) was solved at 1.30 Å resolution for the ribonuclease A-Uridine Vandate (UV) complex. Analysis of the correlation between the observed $^{13}C^{\alpha}$ chemical shifts (from the wild-type protein) and the phosphate-free ribonuclease A (7RNS) solved at 1.26 Å, shows, not surprisingly, a slightly better correlation coefficient than 1RUV, namely R = 0.92. The close agreement, in terms of R, between 1RUV (0.91) and 7RNS (0.92) indicates that both structures are very similar. In fact, the overall rmsd between the C^{α} positions of these two structures is only 0.2 Å, indicating no major conformational change upon UV binding.[25] In other words, the X-ray structure for Ribonuclease A solved with (1RUV) or without (7RNS) ligand (0.91 and 0.92, respectively) are better representations of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the NMR-derived ensemble (2AAS, *R* = 0.89). For some residues of the ensemble of conformations of 2AAS the predictions with *Che*Shift fail because those residues are in high energy regions of the Ramachandran map, for which the DFT method fails to converge, e.g., Asn34 of conformation N° 12 shows a backbone $\phi = -10.9^{o}$ and $\psi = 20.6^{o}$. In such cases, the whole conformation was removed from the analysis using *Che*Shift. Despite this, all of the other servers, namely SHIFTS, SHIFTX or SPARTA, do predict $^{13}C^{\alpha}$ chemical shifts for residues populating high-energy regions of the Ramachandran map. The second value in each row, for the X-ray and NMR-derived conformations, denotes the *R* value without 8 cysteines in the sequence (for a straightforward comparison with the SHIFTS predictions which do not include values for cysteines).

e) For a discussion of the results, see section: *A comparative validation analysis of proteins 1E8L and 3LZT*, of the main text. For each server, except SHIFTS, there are two *R* values for both the X-ray and the NMR-derived models. The first one was computed by using all residues in the sequence, and the second one without the cysteines (for a straightforward comparison with the SHIFTS predictions which do not include values for cysteines).

f) The two values in the first row are the *R* values obtained for the X-ray structures of 4I1B and 2I1B, respectively.

g) For a discussion of the results see section: *A comparative validation analysis of proteins 1UBQ and 1XQQ*, of the main text.

h) There is no X-ray-derived protein structure here, i.e., both ensembles of conformations were obtained by NMR-spectroscopy, and the main differences between these two sets of conformations and the relevance of their analysis was recently discussed.[26] The reported *R* values belong to a segment of 27 residues, from Asp 45 to Asp 71 of protein 1TGQ (now obsolete) and the corresponding segment of protein 2B95. In very good agreement with previous calculations, using the 'internal standard reference',[26] only the results of *Che*Shift, among all the servers, indicates that careful attention should be paid to the fold of this segment in the protein 1TGQ. In other words, this segment of protein 2B95 (*R* = 0.93) is a significantly better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution than 1TGQ (*R* = 0.87).

Section III

Approximations used to interpolate computed $^{13}C^{\alpha}$ chemical shift values

A Gaussian[27] and a linear interpolation function, described in detail for a one-dimensional case by Eq. (1) and (2) below, were used to reproduce the DFT results obtained on a fine grid (see section *Approximations used to interpolate computed $^{13}C^{\alpha}$ chemical shift values*, in the main text). The fill-red circles in each panel of Figure S5 correspond to the results computed using the Gaussian [panels (**a**)-(**b**)] and linear [panels (**c**)-(**d**)] interpolations, respectively. Generalization of these equations to higher dimension of the torsional angle space (i.e. for ϕ, ψ, χ1, and χ2), as used here, is straightforward. The frequencies of the error distribution obtained using three-dimensional interpolations are shown in Figure S6.

I) Gaussian[27] interpolation

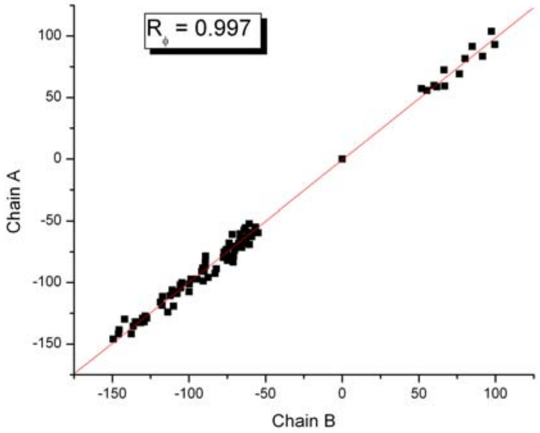
$$ {}^{13}C^{\alpha}_{computed}(\chi_1^n)\,|_{\phi,\psi,\chi^2} = \frac{\sum_{k=1}^{2} {}^{13}C^{\alpha}_{k} \exp[-(\chi_1^n - \chi_{1,k}^o)^2 / S]}{\sum_{k=1}^{2} \exp[-(\chi_1^n - \chi_{1,k}^o)^2 / S]} \qquad (1) $$

with $n = 1, \ldots, 16$; $\chi_1^1 = \chi_{1,1}^0$ and $\chi_1^{16} = \chi_{1,2}^0$; and S a Gaussian scale factor.

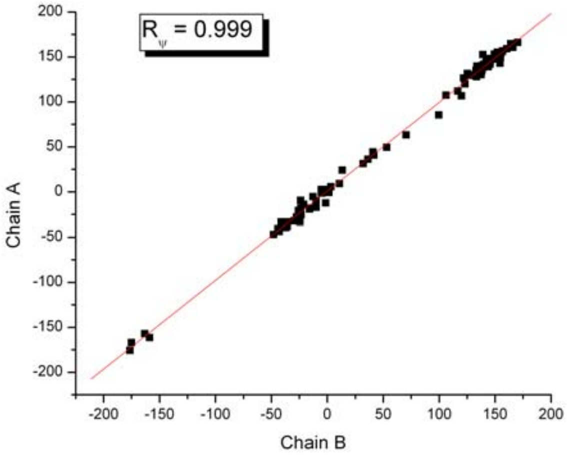
II) Linear interpolation

$$ {}^{13}C^{\alpha}_{computed}(\chi_1^n)\,|_{\phi,\psi,\chi^2} = \sum_{k=1}^{2} {}^{13}C^{\alpha}_{k} \left[\frac{(\chi_{1,k+1}^o - \chi_1^n)}{\Delta\chi_1}\right]^{2\text{-}k} \left[\frac{(\chi_1^n - \chi_{1,k-1}^o)}{\Delta\chi_1}\right]^{k-1} \textbf{with } n = 1,\ldots.,16; \qquad (2) $$

$(\chi_{1,2}^o - \chi_1^n)$ and $(\chi_1^n - \chi_{1,1}^o)$ equal to $\Delta\chi_1$ $(= 30^o)$ for n = 1 and 16, respectively.

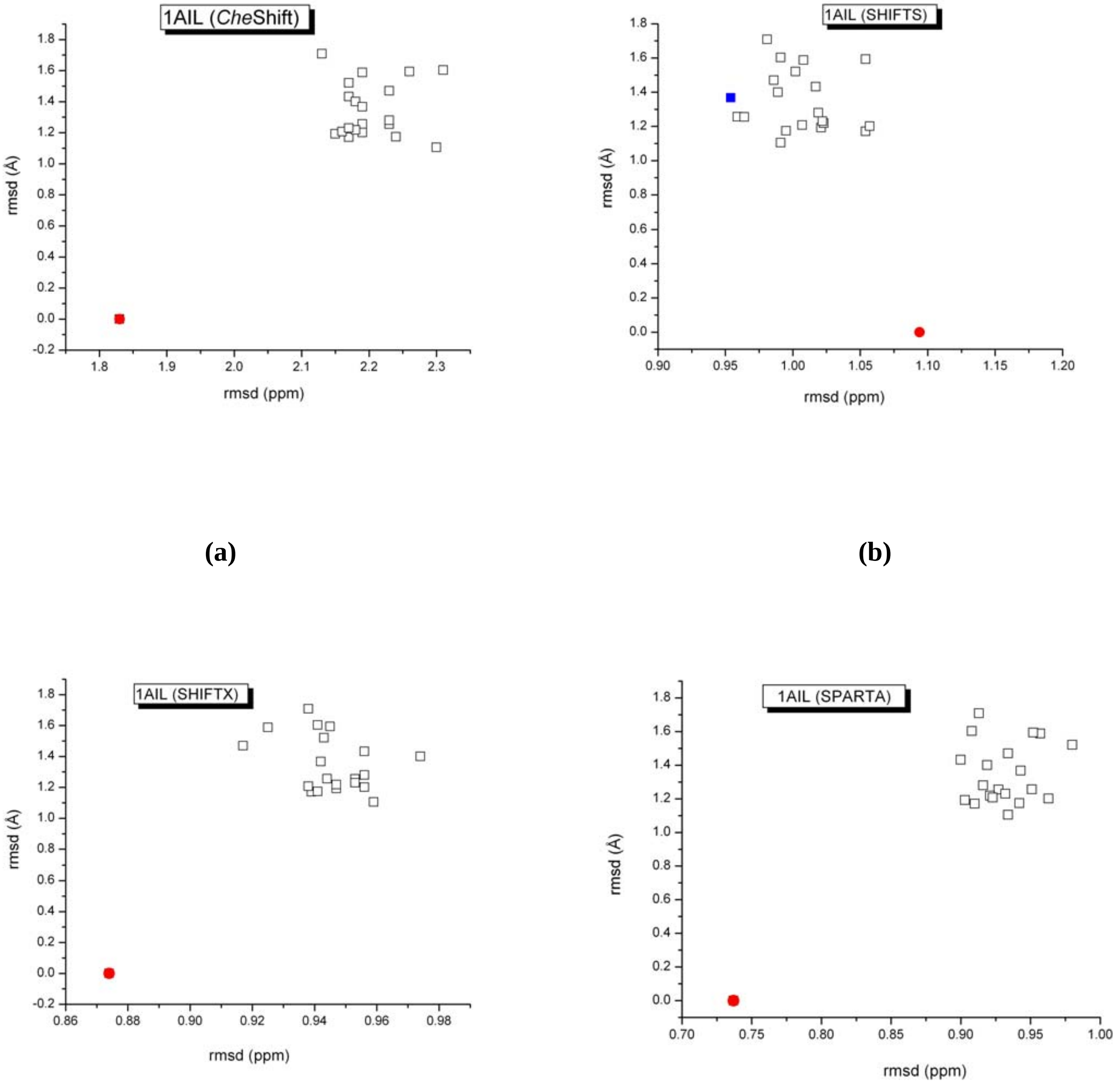


(**a**)



(**b**)

Figure S2. Correlation between the identical backbone torsional angles derived from the molecules A and B of protein PDB id 1RGE, from Ribonuclease Sa. Panel (**a**) for the ϕ torsional angle and (**b**) for the ψ torsional angle.



(a)

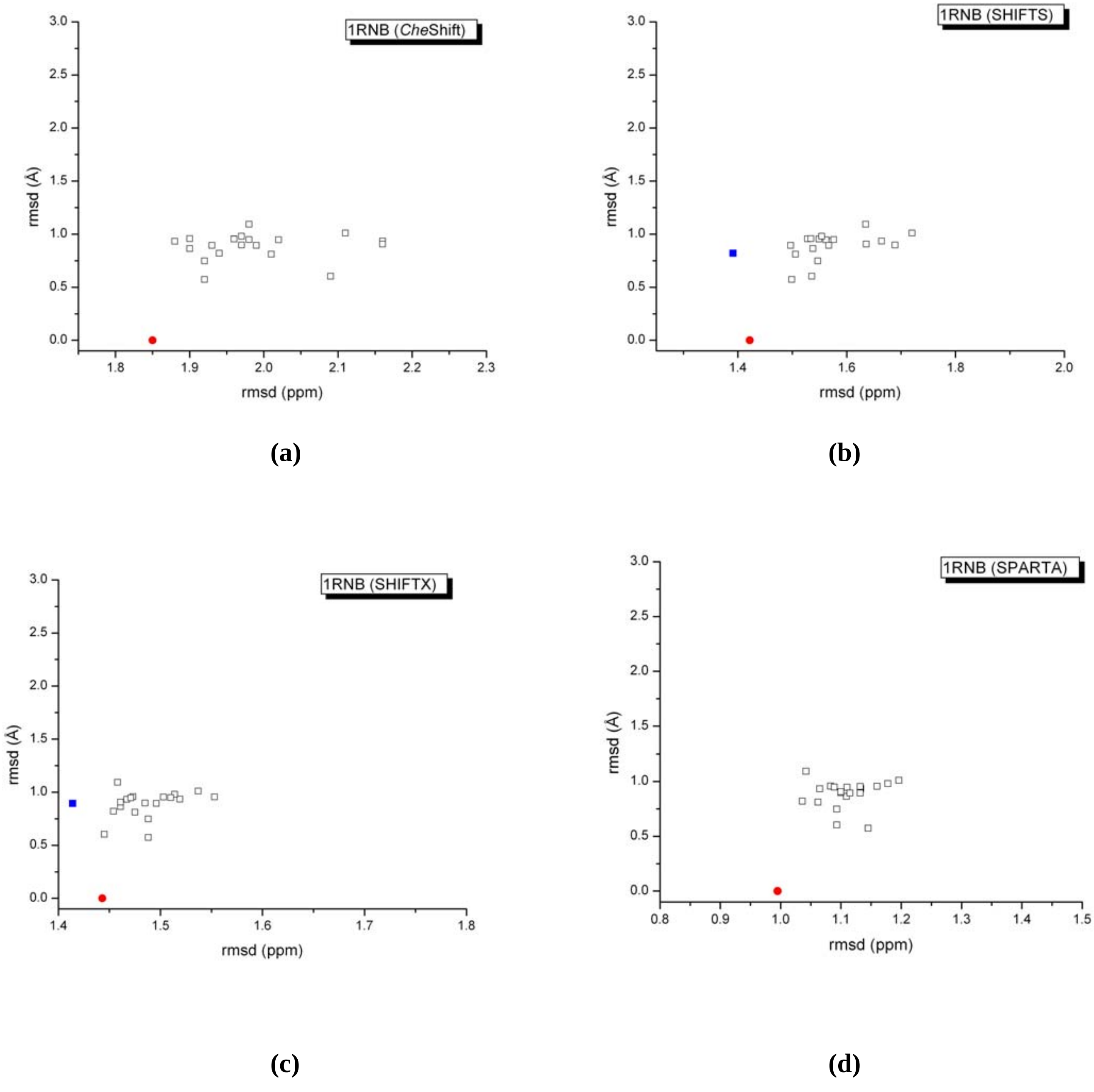
(b)

(c)

(d)

Figure S3. Plot showing rmsd of the protein decoys (open squares), lower than 3 Å from the 'native' structure (PDB id 1AIL), versus the rmsd (ppm) between observed and predicted $^{13}C^{\alpha}$ chemical shifts, computed with four different servers, namely (**a**) *Che*Shift; (**b**) SHIFTS;

(**c**) SHIFTX; and (**d**) SPARTA. In each panel, the red-filled circle denotes the ‘native’ structure (1AIL) from which the decoys were obtained; if one decoy shows a lower rmsd (ppm) than that of the ‘native’ structure, it is denoted by a blue-filled square.



(a)

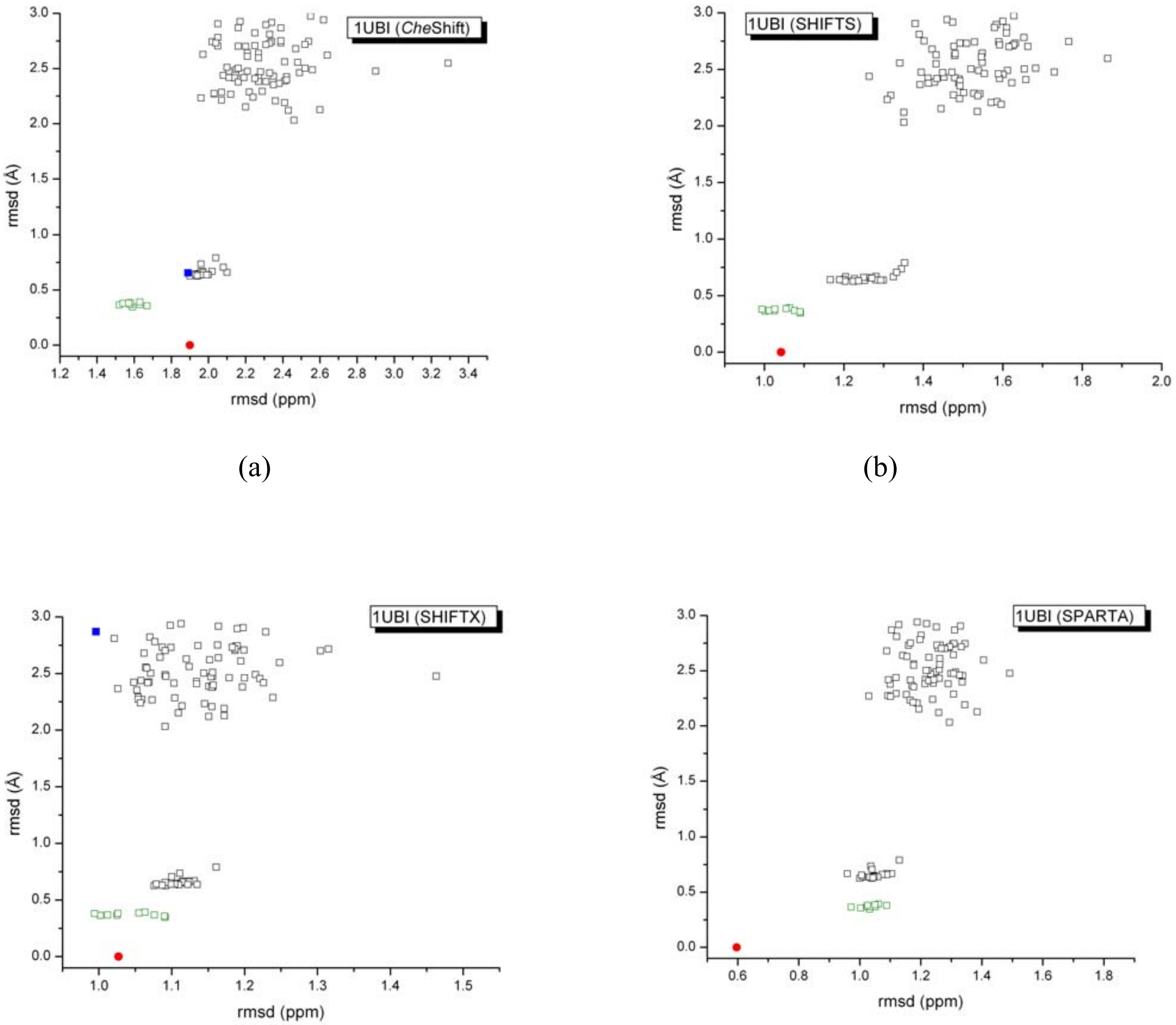
(b)

(c)

(d)

Figure S4. Plot showing rmsd of the protein decoys (open squares), lower than 3 Å from the ‘native’ structure (PDB id 1RNB), versus the rmsd (ppm) between observed and predicted

$^{13}C^{\alpha}$ chemical shifts, computed with four different servers, namely (**a**) *Che*Shift; (**b**) SHIFTS; (**c**) SHIFTX; and (**d**) SPARTA. In each panel, the red-filled circle denotes the 'native' structure (1RNB) from which the decoys were obtained; if one decoy shows a lower rmsd (ppm) than that of the 'native' structure, it is denoted by a blue-filled square.



(a)

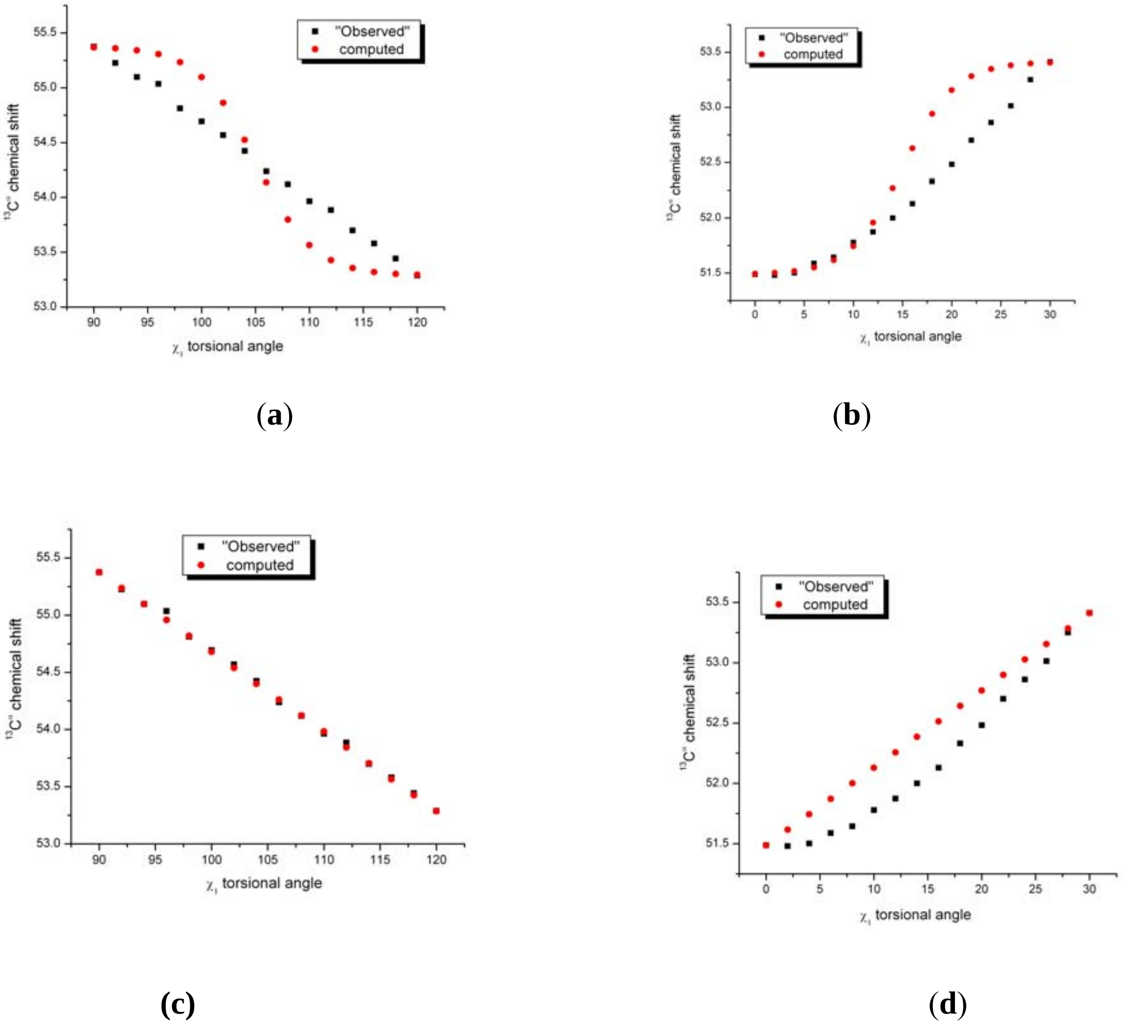
(b)

(c)

(d)

Figure S5.- Plot showing rmsd of the protein decoys (open squares), lower than 3 Å from the 'native' structure (PDB id 1UBI), versus the rmsd (ppm) between observed and predicted $^{13}C^{\alpha}$ chemical shifts, computed with four different servers, namely (**a**) *Che*Shift; (**b**) SHIFTS; (**c**) SHIFTX; and (**d**) SPARTA. In each panel, the red-filled circle denotes the 'native'

structure (1UBI) from which the decoys were obtained; if one decoy shows a lower rmsd (ppm) than that of the 'native' structure, it is denoted by a blue-filled square. Each of the open green squares in each panel, were computed from 10 conformations of protein PDB id 1D3Z solved by NMR spectroscopy at very-high resolution.[1]

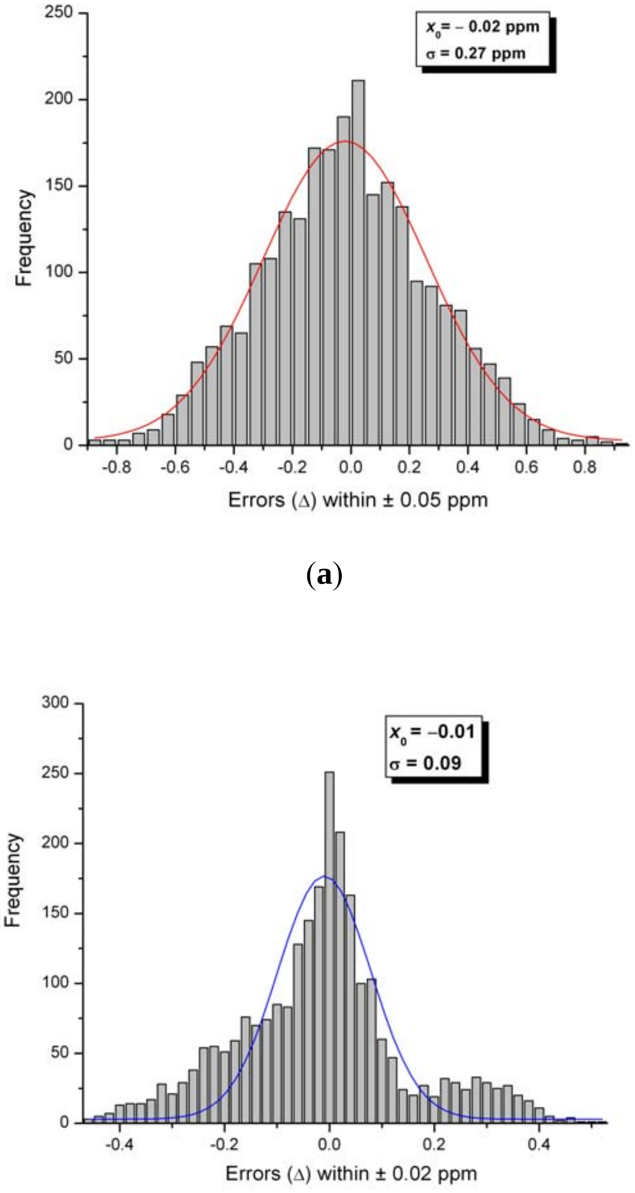


(**a**) (**b**)

(c) (**d**)

Figure S6. In each panel, the black-filled squares indicate the "observed" values, i.e., the $^{13}C^{\alpha}$ chemical shifts computed on a fine grid, namely with 5 degree steps for $\chi 1$, using a small basis set and linearly extrapolated to a large basis set (see Section I, Supporting Information). The red-filled circles indicated the interpolated $^{13}C^{\alpha}$ chemical shifts, between the two known values, namely the first and last one of the 30° interval, by using a Gaussian [panels (**a**) and (**b**)] or linear [panels (**c)** and (**d**)] interpolation, respectively. For a given set of (ϕ, ψ, $\chi 2$) torsional angles, there is a total of 9 panels, using an interval of 30° as in the *Che*Shift database, to fully describe the $\chi 1$ torsional angle variations (−180°, 180°). Among all these 9 panels we selected the best and worst results for each of the two interpolation

methods, i.e., shown in FigureS5 (**a**) and (**b**) for the Gaussian and (**c**) and (**d**) the linear interpolation, respectively. Five out of 9 panels for the linear interpolations are ‘similar’ to the results shown in Figure S5 (**c**); the good performance of the linear interpolation is reflected in the standard deviation, σ, of the frequency of the error distribution [see Figure S7 (**b**)].



(**a**)

(**b**)

Figure S7. Each panel denotes the frequency of the error distribution, between the "observed" $^{13}C^{\alpha}$ chemical shifts and those computed by using (**a**) Gaussian or (**b**) linear interpolation. The frequency of the error distribution can be fitted by a Gaussian curve with the mean value, x_o, and standard deviation, σ shown in the inserted panels of each Figure. A total of 2,500 conformations of Ser were evaluated, by sampling the φ and ψ every 2^o and χ1

every 5^{o}, in the following torsional angle ranges (−150^{o}; −160^{o}), (160^{o}, 170^{o}) and (−180^{o}, 180^{o}), respectively. The computed interpolation values were obtained by using a three-dimensional generalization of Equations (1) and (2) of Section III.

References

1. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase, 1998**,** *J. Am. Chem. Soc.***120**, 6836-6837.
2. Aramini JM, Sharma S, Huang YJ, Swapna GVT, Ho CK, Shetty K, Cunningham K, Ma L-C, Zhao L, Owens LA, Jiang M, Xiao R, Liu J, Baran MC, Acton TB, Rost B and Montelione GT. Solution NMR structure of the SOS response protein YnzC from *Bacillus subtilis* (2008) *Proteins* 72:526-530.
3. Chien C.-y., Tejero R., Huang Y., Zimmerman D.E., Rios C.B., Krug R.M., Montelione G.T. A novel RNA-binding motif in influenza A virus non-structural protein 1, 1997, Nature Structural Biology, **4**, 891-895.
4. Vila, J. A.; Aramini J. M.; Rossi P.; Kuzin A.; Su M.; Seetharaman J.; Xiao R.; Tong L.; Montelione G. T.; H. A. Scheraga. Quantum Chemical $^{13}C^{\alpha}$ Chemical Shift Calculations for Protein NMR Structure Determination, Refinement, and Validation. *Proc. Natl. Acad. Sci. USA*. 2008, **105**, 14389-14394.
5. Liu J., Lynch P.A., Chien C.-y., Montelione G.T., Krug R.M., Berman H.M. (1997) Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. *Nature Structural Biology* **4**, 896-899.
6. Vila JA, Villegas ME, Baldoni HA and Scheraga HA. Predicting $^{13}C^{\alpha}$ chemical shifts for validation of protein structures (2007) *J Biomol NMR* 38:221-235.
7. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, and Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical

parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to praline-containing peptides. (1992) *J Phys Chem* 96:6472-6484.

8. Chesnut DB and Moore KD. Locally dense basis-sets for chemical-shift calculations (1989) *J Comp Chem* 10:648-659.

9. Vila JA and Scheraga HA. Factors affecting the use of $^{13}C^{\alpha}$ chemical shifts to determine, refine, and validate protein structures (2008) *Proteins*, 71:641-654.

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

11. Jameson A. K.; Jameson C.J. J Chem Phys Lett 1997, 134, 461

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

C. Gonzalez, and J. A. Pople, Gaussian 03, Revision E.01, Gaussian, Inc., Wallingford CT, 2004.

13. Vila J.A., Baldoni H.A. and Scheraga H.A. Performance of Density Functional Models to Reproduce Observed $^{13}C^{\alpha}$ Chemical Shifts of Proteins in Solution. *J. Comp. Chem.*, **30**, 884-892 (2009).
14. Press H.W.; Teukolsky S.A.; Vetterling W.T.; Flannery BP, in Numerical Recipes in FORTRAN 77. The Art of Scientific Computing, Second Edition, Cambridge University Press **1992**, Chapter 14, page 630-633.
15. Wang Y.J., Jardetzky O. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci.* 2002, **11**, 852-861.
16. Lovell S.C., Word J.M, Richardson J.S. and Richardson D.C. The penultimate rotamer library. Proteins, 2000, **40**, 389-408.
17. Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 Å resolution (1987) *J Mol Biol* 194:531-544.
18. Xu X.-P. and Case D.A. Automated prediction of ^{15}N, $^{13}C^{a}$, $^{13}C^{b}$ and $^{13}C'$ chemical shifts in proteins using a density functional database. *J. Biomol. NMR*, 2001, **21**, 321-333.
19. Xu X.-P. and Case D.A. Probing multiple effects on ^{15}N, $^{13}C^{a}$, $^{13}C^{b}$ and $^{13}C'$ chemical shifts in peptides using density functional theory. *Biopolymers*, 2002, **65**, 408-423.
20. Neal S., Nip A.M., Zhang H. and Wishart D.S. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, 2003, **26**, 215-240.
21. Meiler J. PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, 2003, **26**, 25-37.

22. Shen Y. and Bax Ad. Protein backbone chemical shifts predicted from searching a database for torsional angle and sequence homology. J. Biomol. NMR, 2007, **38**, 289-302.

23. Lindorff-Larsen K., Best R.B., Depristo M.A., Dobson C.M. and Vendruscolo M. Simultaneous determination of protein structure and dynamics, 2005, Nature, **433**, 128-132.

24. Ulrich E.L., Akutsu H., Doreleijers F.J., Harano Y., Loannidis E.Y., Lin J., Livny M., Mading S., Maziuk D., Miller Z., Nakatani E., Schulte C.F., Tolmie D.E., Wenger R.K., Yao H. and Markley J.L. BioMagResBank, Nucleic Acids Res., 2007, **36**, D402-D408.

25. Ladner J.E., Wladkowski B.D., Svensson L.A., Sjölin L. and Gilliland G. X-ray structure of Ribonuclease A-uridine vanadate complex at 1.3 Å resolution. Acta Cryst., 1997, D**53**, 290-301.

26. Vila J.A. and Scheraga H.A. Assessing the accuracy of protein structures by quantum mechanical computations of $^{13}C^{\alpha}$ chemical shifts. *Accounts of Chemical Research*, 2009, in press.

27. Kuszewski J., Qin J., Gronenborn A.M., Clore M.G. The impact of direct refinement against $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts on protein structure determination by NMR. J. Mag. Res., 1995, B**106**, 92-96.