^{13}C Chemical Shifts in Proteins: A Rich Source of Encoded Structural Information

Jorge A. Vila* and Yelena A. Arnautova

Abstract. Despite the formidable progress in Nuclear Magnetic Resonance (NMR) spectroscopy, quality assessment of NMR-derived structures remains as an important problem. Thus, validation of protein structures is essential for the spectroscopists, since it could enable them to detect structural flaws and potentially guide their efforts in further refinement. Moreover, availability of accurate and efficient validation tools would help molecular biologists and computational chemists to evaluate quality of available experimental structures and to select a protein model which is the most suitable for a given scientific problem. The ^{13}C$^{\alpha}$ nuclei are ubiquitous in proteins, moreover, their shieldings are easily obtainable from NMR experiments and represent a rich source of encoded structural information that makes ^{13}C$^{\alpha}$ chemical shifts an attractive candidate for use in computational methods aimed at determination and validation of protein structures. In this chapter, the basis of a novel methodology of computing, at the quantum chemical level of theory, the ^{13}C$^{\alpha}$ shielding for the amino acid residues in proteins is described. We also identify and examine the main factors affecting the ^{13}C$^{\alpha}$-shielding computation. Finally, we illustrate how the information encoded in the ^{13}C chemical shifts can be used for a number of applications, viz., from protein structure prediction of both α-helical and β-sheet conformations, to determination of the fraction of the tautomeric forms of the imidazole ring of histidine in proteins as a function of pH or to accurate detection of structural flaws, at a residue-level, in NMR-determined protein models.

Jorge A. Vila
IMASL-CONICET, Universidad Nacional de San Luis,
Ejército de Los Andes 950-5700 San Luis, Argentina
Baker Laboratory of Chemistry and Chemical Biology,
Cornell University, Ithaca; NY 14853-1301, USA
email: jv84@cornell.edu

Yelena A. Arnautova
Molsoft L.L.C., 11199 Sorrento Valley Road,
S209, San Diego; CA, 92121

* Corresponding author.

A. Liwo (Ed.): *Comput. Methods to Study the Struct. & Dyn. of Biomolecules*, SSBN 1, pp. 639–683.
DOI: 10.1007/978-3-642-28554-7_19

Abbreviations

AMBER = Assisted Model Building with Energy Refinement
ca-rmsd = conformational average-root mean square deviation
*Che*Shift = Chemical Shifts
CPU = Central Processing Unit
DFT = Density Functional Theory
ECEPP = Empirical Conformational Energy Program for Peptides
MBE = Multiple Boundary Element
MCM = Monte Carlo with Minimization
MD = Molecular Dynamics
NMR = Nuclear Magnetic Resonance
NOE = Nuclear Overhauser Effect
PARSE = Parameters for Solvation Energy
PDB = Protein Data Bank
QM = Quantum mechanics
TMS = Tetramethylsilane
VTF = Variable Target Function

1 Introduction

Before a protein structure can be analyzed in light of its biological function it is necessary to validate it, i.e., to have a clear understanding of its reliability in terms of both the overall structure and of its details at per-residue level. However, an accurate and fast validation of protein structures constitutes a long-standing problem in Nuclear Magnetic Resonance (NMR) spectroscopy (Williamson et al., 1995; Bhattacharya et al., 2007; Billeter et al., 2008; Williamson and Craven, 2009). For this reason, investigators have proposed a plethora of methods to determine the accuracy and reliability of protein structures in recent years (Vriend, 1990; Lüthy et al., 1992; Laskowski et al., 1993; Lovell et al., 2003; Huang et al., 2005, 2006; Nabuurs et al., 2006; Davis et al., 2007). Despite this progress, there is a growing need for more sophisticated, physics-based, and fast structure-validation methods (Huang et al., 2005, 2006; Nabuurs et al., 2006; Bhattacharya et al., 2007; Billeter et al., 2008).

The $^{13}C^{\alpha}$ chemical shifts provide important information about conformations of peptides and proteins in solution (Spera and Bax, 1991; de Dios et al., 1993a, 1993b; Lee and Oldfield, 1994; Kuszewski et al. 1995; Luginbühl et al., 1995; Wishart et al., 1995a, 1995b; Havlin et al., 1997; Iwadate et al. 1999; Cornilescu et al., 1999; Xu and Case, 2001, 2002; Meiler 2003; Neal et al., 2003; Vila et al., 2003, 2004; Berjanskii et al., 2005, 2007; Shen and Bax, 2007; Villegas et al., 2007; Vila et al., 2007a,b, 2008a,b; Vila and Scheraga, 2009; Vila et al., 2009;

Shen et al., 2008; Han et al., 2011; Frank et al., 2012) and, therefore, can be used as an exquisitely sensitive probe with which to assess the *quality* of protein models. We developed recently a new, physics-based methodology (Vila and Scheraga, 2009), that makes use of observed and computed {at the Density-functional theory (DFT) level of theory [Parr and Yang, 1989]} $^{13}C^{\alpha}$ chemical shifts for an accurate validation of protein structures in solution and in crystal (Arnautova et al., 2009). The first step in the development of this new methodology involved determining the factors that affect $^{13}C^{\alpha}$ shielding calculations, such as the protonation/deprotonation state of distant ionizable groups, sequential nearest-neighbor, or covalent geometry effects (i.e., due to variations in the bond lengths and bond angles of residues) and the sensitivity of the shielding/deshielding of $^{13}C^{\alpha}$ nuclei to changes in side-chain conformation. Once all these factors affecting $^{13}C^{\alpha}$-shielding have been properly identified and considered, a very important test is to determine the accuracy and speed of the computation of the $^{13}C^{\alpha}$-shielding as a function of the size of the basis set chosen and the Density Functional Theory (DFT) model adopted. These are important tests because DFT-based quantum mechanical (QM) calculations are very CPU demanding, despite the ever-increasing computational power available.

The new DFT-based method has been applied to study a number of problems, such as unblocked statistical-coil tetrapeptides in aqueous solution (Vila et al., 2003), polyproline II helix conformation in a proline-rich environment (Vila et al., 2004), the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of cysteines in disulfide-bonded cysteine (Martin et al., 2010) or determination of the fraction of the tautomeric forms of histidine in proteins as a function of pH (Vila et al., 2011). This new strategy also provides a unified, self-consistent method to determine high-quality protein structures, without relying on knowledge-based information (Vila et al. 2007b). Thus, a β-sheet or an all α-helical protein structure can be accurately determined by simply identifying a set of conformations which simultaneously satisfy a number of constraints, namely $^{13}C^{\alpha}$-dynamically derived torsional angle constraints and Nuclear Overhauser Effect (NOE) derived distance constraints (Vila et al., 2007b,2008a).

The currently used $^{13}C^{\alpha}$ chemical shift-based validation and determination protocol (Vila and Scheraga 2008; 2009; Vila et al. 2007a, b; 2008a) exploits the following features: (*a*) the assignment of chemical shifts is a fundamental step in a protein structure determination by NMR spectroscopy (Wüthrich, 1986), and no extra experimental work is needed; (*b*) in addition to the impact of the covalent structure, $^{13}C^{\alpha}$ chemical shifts are modulated mainly by the intraresidue backbone and side-chain dihedral angles (Spera and Bax 1991; de Dios et al. 1993a, 1993b; Kuszewski et al. 1995; Luginbühl et al. 1995; Havlin et al. 1997; Pearson et al. 1997; Iwadate et al. 1999; Xu and Case 2001; Sun et al. 2002; Villegas et al. 2007), with no significant influence of the amino acid sequence (Vila et al., 2010); (*c*) $^{13}C^{\alpha}$ is ubiquitous in proteins; and, (*d*) $^{13}C^{\alpha}$ chemical shifts can be computed with high accuracy at the QM level of theory.

This chapter is intended to be an overview of the author's contribution to the field of protein structure determination and validation using, *mainly*, information decoded from the $^{13}C^{\alpha}$ chemical shifts. Consequently, the chapter is organized as follows: first, the method used to compute the $^{13}C^{\alpha}$ chemical shifts and to analyze the results are briefly described; second, the main factors affecting the $^{13}C^{\alpha}$ chemical shifts computation are enumerated and discussed; third, the capabilities of the computed $^{13}C^{\alpha}$ chemical shifts, as a rich source of encoded structural information, are illustrated by a series of applications that involves, but is not limited to, the determination of protein structures; and finally a new protein-structure validation server, *Che*Shift-2 (Martin et al., 2012), with which NMR spectroscopists can assess the quality of their protein models, before they are deposited in the Protein Data Bank (PDB) [Berman et al., 2000], is presented. It is worth noting that the theory, and details, behind alternative protein structure determination and validation methods are not discussed here and, hence, the reader is referred instead to an extensive collection of such methods (Vriend, 1990; Günter et al., 1991; Lüthy et al., 1992; Laskowski et al., 1993; Günter 1998; Cornilescu et al., 1998; Brünger et al., 1998; Lovell et al., 2003; Huang et al., 2005, 2006; Nabuurs et al., 2006; Brünger, 2007; Davis et al., 2007; Bhattacharya et al, 2007; Cavalli, et al., 2007; Shen et al., 2008; Günter 2009; Rosato et al., 2009; Frank et al., 2011; Guerry and Herrmann, 2011; Rosatto et al., 2012).

2 Methods

2.1 Calculation of $^{13}C^{\alpha}$ Chemical Shifts

All the experimentally determined conformations, unless noted otherwise, were *regularized*, i.e., all residues were replaced by the standard Empirical Conformational Energy Program for Peptides (ECEPP) [Némethy et al., 1992] residues in which bond lengths and bond angles are fixed (rigid-body geometry approximation) at the standard values, (Némethy et al., 1992) and hydrogen atoms were added, if necessary.

Computations of the $^{13}C^{\alpha}$ chemical shifts involve a series of approximations. For each amino acid residue **X** in the protein sequence: (*a*) the $^{13}C^{\alpha}$ shielding depends, mainly, on its own backbone conformations (Spera and Bax, 1991; deDios et al., 1993; Kuszewski et al., 1995) and side-chain (Iwadate et al., 1999; Havlin et al., 1997; Pearson et al., 1997; Villegas et al., 2007), with no significant influence of either the amino acid sequence or the position of the given residue in the sequence, except for residues preceding proline (Vila et al., 2010); (*b*) each amino acid residue **X** in the protein sequence can be treated as a terminally blocked tripeptide with the sequence Ac-G**X**G-NMe, with **X** in the conformation of the protein structure; (*c*) the $^{13}C^{\alpha}$ isotropic shielding values (σ) for each amino acid residue **X** can be computed at the OB98/6-311+G(2d,p) level of theory (Vila *et al.*, 2009) with the Gaussian 03 package (Frisch *et al.*, 2004). The remaining residues in each tripeptide are treated at the OB98/3-21G level of theory, i.e., by using the *locally dense basis set* approach (Chesnut and Moore, 1989); (*d*) all ionizable residues can be considered neutral during the QM calculations (Vila and Scheraga, 2008), unless noted otherwise; (*e*) no geometry optimization is necessary because such optimization by *ab-initio* (HF) or DFT methods has only a small effect on the computed chemical shifts (Pearson et al., 1997).

The computed $^{13}C^{\alpha}$ shieldings ($\sigma_{\mathrm{subst},th}$) are 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 (Jameson and Jameson, 1987), namely 188.1 ppm, was adopted as an initial (see below) reference value. All the computed $^{13}C^{\alpha}$ shielding ($\sigma_{\mathrm{subst},th}$) values are calculated using the Gauge-Invariant Atomic Orbital method at the DFT level of theory as implemented in the GAUSSIAN 03/09 suite of programs (Frisch et al., 2003). For all purposes, in this chapter, we have used only one exchange-correlation functional, OB98, because it was shown (Vila et al., 2008b) to be one of the most accurate *and* fast functionals with which to reproduce the observed $^{13}C^{\alpha}$ chemical shifts of proteins in solution (see section, 3.2, below).

2.2 Determination of an Effective TMS Shielding Value

Determination of a proper TMS shielding value for each functional is crucial for an accurate computation of the $^{13}C^{\alpha}$ chemical shifts because it will enable us to minimize the presence of systematic errors which might *bias* the chemical shifts-based analysis. From this point of view the *effective* TMS value will provide the most accurate approach to solve the problem because it will not require further adjustments. Consequently computation of an effective TMS values is central to our calculations.

By adopting the observed TMS value of 188.1 ppm (Jameson and Jameson, 1987) as a reference it is possible to find for any functional, the characteristic mean (x_o) and standard deviation (σ) of the Normal (or Gaussian) fit of the frequency of the errors distribution. For all functionals tested in our work the characteristic mean value (x_o) appears displaced from its ideal value of 0.0 by a positive, or negative, amount, e.g., for OB98 a $x_o = +3.6$ ppm was found. Further analysis (Vila et al., 2008b) indicates that for *any* of the 10 functionals tested a straightforward use of the observed TMS shielding value (188.1 ppm) is not appropriate, if no further corrections are introduced. Hence, for each functional and basis set chosen 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 a $x_o = 0.0$. For example, use of OB98 with a large [6-311+G(2d,p)/3-21G] basis set leads to an *effective* TMS of 184.5 ppm, i.e., by subtracting 3.6 ppm from 188.1 ppm (Vila et al., 2008b), that gives a $x_o = 0.0$ ppm. Likewise, use of a small (6-31G/3-21G) basis set leads to an *effective* TMS of 195.4 ppm.

2.3 Computation of the CA-RMSD Model

The observed chemical shift for each residue *i*, ${}^{13}\mathrm{C}^{\alpha}{}_{\mathrm{observed},i}$, represents contributions from an ensemble of rapidly interconverting conformers that coexist in solution. Then, an accurate comparison between the observed and computed ${}^{13}\mathrm{C}^{\alpha}$ chemical shifts requires consideration of an ensemble of NMR-derived conformers, rather than of a single conformation (Vila et al., 2007a; Arnautova et al., 2009). Consequently, for each amino acid residue in the sequence, *i*, the average of the chemical shifts calculated for the individual residues in the ensemble of Ω conformers representing the NMR structure, $< {}^{13}\mathrm{C}^{\alpha} >_i$, is computed as:

$$< {}^{13}\mathrm{C}^{\alpha} >_i = (1/\Omega) \sum_{k=1}^{\Omega} {}^{13}\mathrm{C}^{\alpha}{}_{i\,k}\,, \tag{1}$$

where ${}^{13}\mathrm{C}^{\alpha}{}_{i,k}$ is the computed chemical shift for residue *i* in conformer *k*, with $1 \leq i \leq N$, where *N* is the number of residues in the sequence. Derivation of Equation (1) was obtained through the following approximation: for each residue *i* the quantity to be computed must, in principle, be $< {}^{13}\mathrm{C}^{\alpha} >_i = \sum_{k=1}^{\Omega} \lambda_k \, {}^{13}\mathrm{C}^{\alpha}{}_{i,k}$, where λ_k is the Boltzmann factor for conformer *k*, with $\sum_{k=1}^{\Omega} \lambda_k \equiv 1$. But, computation of the Boltzmann factors at QM level of theory is not possible, with the existing computational facilities, because it would require computation of the total energy at the QM level of theory for each of the conformers in the ensemble used to represent the NMR structure. Therefore, the following approximation was used: $\lambda_k = 1/\Omega$ (Vila et al., 2010); in other words, in this approximation each conformer contributes equally to the average chemical shift obtained by fast conformational averaging. Whether a computation of a Boltzmann average, rather than the arithmetic average, would lead to a more accurate representation of the ${}^{13}\mathrm{C}^{\alpha}$ chemical shifts needs further investigation.

The $< {}^{13}\mathrm{C}^{\alpha} >_i$ value obtained from Eq. (1) is used to compute the conformational-average difference Δ_i between the observed and computed ${}^{13}\mathrm{C}^{\alpha}$ chemical shifts for each amino acid residue *i*,

$$\Delta_i = ({}^{13}C^{\alpha}_{observed,i} - < {}^{13}C^{\alpha} >_i) \tag{2}$$

Hereafter, the conformational-average root-mean-square-deviation (rmsd) parameter, *ca*-rmsd (Vila et al. 2010), is obtained as:

$$ca\text{-rmsd} = [(1/N)\sum_{i=1}^{N}\Delta_i^2]^{1/2}, \tag{3}$$

which is a global property of the protein NMR structure given as the weighted average of the differences between the experimental $^{13}C^{\alpha}$ chemical shifts and the $< {}^{13}C^{\alpha} >_i$ - values for all the residues in the protein.

2.4 $^{13}C^{\alpha}$-Based Protein Structure Determination Method

The $^{13}C^{\alpha}$-based procedure used for determination of protein structures consists of three steps. The flow chart of this protocol (Vila et al., 2007b) is shown in Figure 1 and a brief description of each step follows.

Step 1: The Variable-Target-Function (VTF) approach with a simplified soft-sphere potential function (Vásquez and Scheraga, 1988) is used to generate an ensemble of conformations at random that simultaneously satisfy a set of long-range distance constraints derived from the experimental NOEs and (φ,ψ) torsional constraints, derived from the observed $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ conformational shifts (Spera and Bax, 1991). The derived torsional constraints are *only* for those amino acids residues in the sequence that pertain to a regular structure, i.e., to a α-helix or β-sheet. Consequently, these $(\varphi,\psi)_{\alpha,\beta}$ torsional constraints (shown in Figure 1) are limited to, on average, ~50% of the amino acids residues in proteins because the remaining ones populate non-regular structures.

Then, a clustering procedure, e.g., the Minimal Spanning Tree method (Kruskal, 1956), is used to select a small sub-set of the total number of the VTF-derived conformations, namely those possessing a maximum NOE-derived distance violation lower than some arbitrary fixed value. For each of these conformations the $^{13}C^{\alpha}$ chemical shifts are computed as described in section 2.1. Examination of the chemical shifts of all the amino acids in the ensemble of conformations enables us to identify the amino acid at each position in the sequence whose computed chemical shifts most closely match the observed ones, among all these conformations. This identified set of individual amino acid conformations corresponds to only one conformation of the whole chain: the '*theoretical minimal-rmsd model*' (Vila et al., 2007a). In this model, the $^{13}C^{\alpha}$ chemical shift of each residue individually best matched the experimental one, thereby providing a *new* set of ϕ, ψ, and χ torsional angle constraints for all amino acid residues in the sequence, i.e., not just for the amino acid residues in regular

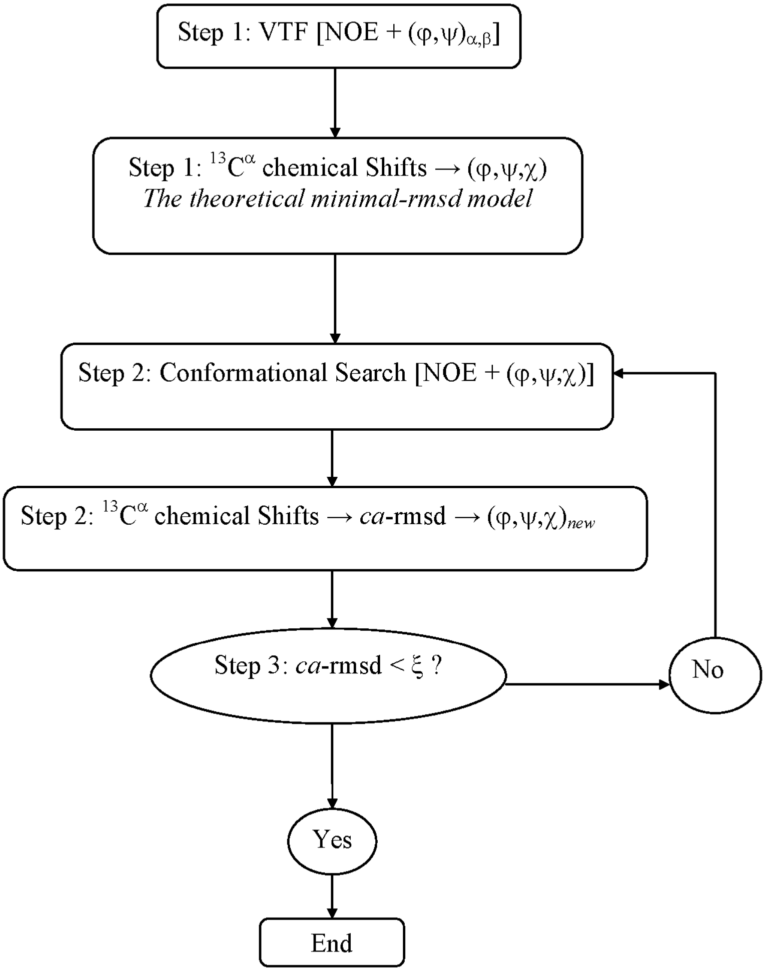
Step 1: VTF [NOE + $(\varphi,\psi)_{\alpha,\beta}$]
Step 1: $^{13}C^{\alpha}$ chemical Shifts → (φ,ψ,χ)
The theoretical minimal-rmsd model
Step 2: Conformational Search [NOE + (φ,ψ,χ)]
Step 2: $^{13}C^{\alpha}$ chemical Shifts → *ca*-rmsd → $(\varphi,\psi,\chi)_{new}$
Step 3: *ca*-rmsd < ξ ?
No
Yes
End

Fig. 1 Flow-chart of the $^{13}C^{\alpha}$-based protein structure determination protocol described in the Methods section. Figure adapted from Vila et al. (2007b). Copyright 2007 American Chemical Society.

structures. Because the chemical shifts are a multivalued function of the ϕ, ψ, and χ torsional angles, the set of torsional angles derived from the '*theoretical minimal-rmsd model*' does not, necessarily, represent a unique solution to a given set of observed $^{13}C^{\alpha}$ chemical shifts values.

Step 2: Only one conformation among all the conformations produced in **Step 1** is selected, for example, the conformation possessing the lowest rmsd between the computed and observed $^{13}C^{\alpha}$ chemical shifts. The selected conformation is used as a starting one in a new conformational search with the Monte Carlo with Minimization (MCM) method (Li and Scheraga, 1987; Li and Scheraga, 1988). The MCM search is carried out with two types of constraints: the original set of NOE-derived distance constraints and the *new* set of ϕ, ψ, χ torsional angles derived in **Step 1**. This time the conformational search is carried out using a complete force-field including the internal potential energy described by ECEPP/05 (Arnautova et al., 2006), the solvent free energy calculated by using a solvent-accessible surface area model (Vila et al., 1991), and an additional energy terms aimed at penalizing violations of the distance and torsional angle constraints (Ripoll and Ni, 1988). Convergence of the determination protocol is monitored using the *ca*-rmsd between the computed and observed $^{13}C^{\alpha}$ chemical shifts.

Step 3: If the computed *ca*-rmsd is lower than certain, arbitrary chosen, cutoff value (ξ), then the procedure is ended. Otherwise, the **Step 2** is repeated using a *new* set of (ϕ,ψ,χ) derived from the *minimal-rmsd-model* of the previous step.

It is worth noting that after our *physics-based* protocol was published (Vila et al., 2007b) an alternative *knowledge-based* method that makes use of ^{1}H, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and ^{15}N chemical shifts as restraints, was successfully applied to structure determination of several proteins (Cavalli et al., 2007). A blind test of computational methods, included several that use also chemical shifts as restraints, aimed at fully automated determination of protein structures has been carried out recently (Rosato et al, 2012).

2.5 Computation of the $^{13}C^{\alpha}$ Chemical Shifts as Function of the pH

For a given residue *i*, of a protein in a conformation *k*, the average charge distribution, $< \rho_{i,k} >$, could be determined by solving the Poisson equation by considering the 2^{ξ} ionization states, with ξ being the number of ionizable groups in the molecule. Regarding this problem, it is worth noting that ξ could be a large number because ~30% of *all* residues in a protein sequence are, on average, ionizable and, hence, an accurate solution would require a fast algorithm. Consequently, in all the applications mentioned in this chapter, we used the Multiple Boundary Element (MBE) method (Vorobjev *et al.*, 1997; 2008), in which the free energy associated with the state of ionization of the ionizable groups at a fixed pH value, namely 6.5, is calculated with the general multi-site titration formalism (Ripoll et al., 1996; Vila et al., 2005). The charges and atomic radii from the PARSE (Parameters for Solvation Energy) algorithm (Sitkoff et al., 1994) were used for the solvation free energy calculations using the MBE method, and the internal (ε_{int}) and solvent ($\varepsilon_{\text{solv}}$) dielectric constants of 2 and 80, respectively (Vila et al., 2005) were adopted for the calculations of $< \rho_{i,k} >$.

The value of $\varepsilon_{\text{int}} = 2$ is consistent with the use of PARSE charges (Barth et al., 2007) and is also commonly assumed as an adequate representation of the protein interior. Following these approximations, for a given conformation *k,* the average degree of ionization of the *i*th ionizable group of this conformation is computed as:

$$< \rho_{i,k} >= Z^{-1} \sum_{n=1}^{2^{\xi}} \rho_{i,k}^{n} [-\Delta G(P_k, x_k^n) / k_B T] \tag{4}$$

where Z is the partition function, k_B is the Boltzmann constant, T is the absolute temperature, x_k^n = $(\rho_{1,k}^n, ..., \rho_{i,k}^n, ..., \rho_{N,k}^n)$ with $\rho_{i,k}^n$ = (1 or 0) is the *n*th protonation microstate of conformation *k* for protein P$_k$. $\Delta G(P_k, x_n^k)$ is the free energy of ionization of the *n*th microstate of protein P$_k$ in conformation *k* (Ripoll et al, 1996).

It should be noted that for any ionizable residue *i* of a single conformation *k*, equation (4) can lead to a non-integer average degree of charge, although we know that such non-integer charges do not make physical sense. Due to the Boltzmann nature of the averaged value computed by equation (4), a fractional charge should physically be interpreted as follows: for a given conformation *k,* there are many identical replicas of such a conformation in solution and, hence, a fractional charge computed by equation (4), e.g., 0.75, means that 75% of these replicas possess the ionizable group *i* protonated/deprotonated with an integral charge while the remaining 25% of the replicas possess the same ionizable group as deprotonated/protonated, depending on whether the ionizable group is basic or acidic.

Assuming that the protonation/deprotonation reactions are instantaneous on the NMR time scale, i.e., microsecond to millisecond (Hass et al., 2008), the theoretical ^{13}C$^{\alpha}$ chemical shifts, $\delta_i^{computed}(pH)$, for a given residue *i* in the sequence (except for histidine that possess 2 tautomers) are computed as a function of the pH using the following equation:

$$\delta_i^{computed}(pH) = (1/\Omega) \sum_{k=1}^{\Omega} \{< \rho_{i,k} > \delta^{+,i,k} + (1- < \rho_{i,k} >) \delta^{0,i,k}\} \tag{5}$$

where $\delta^{+,i,k}$ and $\delta^{0,i,k}$ are the computed ^{13}C$^{\alpha}$ chemical shifts, for the amino acid *i* in conformation *k*, with fully charged and neutral side chains, respectively, Ω is the number of conformers in the protein ensemble, and $< \rho_{i,k} >$ the averaged degree of charge, as given by equation (4).

3 Factors Affecting the Calculation of $^{13}C^{\alpha}$ Chemical Shifts

3.1 Transferability of the Results

The current methodology (Vila et al., 2007a; Vila and Scheraga 2009) relies on a crucial observation: once 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 side-chain conformations, with no significant influence by the nature of the nearest-neighbor amino acids, except for residues immediately preceding proline (Vila et al., 2010).

The above observation allows us to parallelize the $^{13}C^{\alpha}$ shielding calculations in proteins and, hence, to make them computationally feasible. Moreover, 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. The results of these tests should be transferable to proteins of any class or size. Consequently, we used structures of three proteins solved by NMR and X-ray, namely PDB id 1D3Z, 2JVD, and 1NS1 to evaluate the performance of different DFT functionals *and* basis sets, as explained below.

3.2 Performance of Different DFT Functionals to Reproduce Observed $^{13}C^{\alpha}$ Chemical Shifts

DFT has become a method of choice for QM calculations of the electronic structure and properties of many molecular and solid systems. Because the exact exchange-correlation functional is unknown, a large number of approximations has been proposed in the literature making it essential to pursue more accurate and reliable approximate functional, a process which, on the other hand, depends on the applications. Selection of the most appropriate density functional model for a particular application becomes one of the main problems of the DFT method. For this reason we decided (Vila et al., 2009) to test several density functional models (namely B3LYP, OLYP, PBE1PBE, OPBE, O3LYP, OPW91, OB98, BPW91, BPBE, and B971). The benchmarking was intended to find not only the most accurate functional with which to reproduce the observed $^{13}C^{\alpha}$ chemical shifts in solutions but also the fastest one, in terms of CPU time, because speed of DFT calculations could severely limit their applicability to proteins. The test was applied to 10 NMR-derived conformations of the 76-residue α/β protein ubiquitin (PDB id 1D3Z).

Comparison of the observed and computed $^{13}C^{\alpha}$ chemical shifts shows that there are five functionals, namely OPW91, OB98, OPBE, OLYP, and O3LYP, which are among the faster ones and, even more importantly, behave *very similarly* in their ability to reproduce accurately the observed $^{13}C^{\alpha}$ chemical shifts. In particular, we observe that OB98 appears to be slightly better than any other of the five functionals in terms of both the correlation coefficient, *R*, (or *Pearson* coefficient) between the observed and the conformational-averaged $^{13}C^{\alpha}$ chemical shifts and the standard deviation of the computed conformational-averaged $^{13}C^{\alpha}$ chemical shifts from a linear regression. Consequently, we chose the OB98 for all the applications (Vila et al., 2008b).

We also compared the results obtained using OB98 with those obtained with B3LYP, a very popular functional that has been used extensively in our group, and elsewhere. The correlation existing between averaged $^{13}C^{\alpha}$ chemical shift values obtained for the 10 conformations of 1D3Z with OB98 and B3LYP functional, is excellent (Vila et al., 2008b), i.e., showing a correlation coefficient $R = 0.998$ and standard deviation of 0.300 ppm. This test provides solid evidence that the results and conclusions obtained using B3LYP do not need to be revised if the OB98 functional is adopted (Vila et al., 2008b).

3.3 Performance of Different Basis Sets to Reproduce Observed $^{13}C^{\alpha}$ Chemical Shifts

To study 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, six basis sets, 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 *locally-dense* basis-set approximations, and uniform 3-21G/3-21G set were initially applied (Vila et al., 2009) to 10 NMR-derived conformations ubiquitin (Cornilescu et al., 1998). For each of these six basis sets, combined with the OB98 functional, 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 (Vila et al., 2009), 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.

The results also indicates that the $^{13}C^{\alpha}$ chemical shifts computed with the large [6-311+G(2d,p)/3-21G] basis set, can be reproduced accurately (within an average error of ~0.4 ppm) and faster (by ~9 times) by using the small (6-31G/3-21G) basis set after extrapolating it with: $^{13}C^{\alpha} = -1.597 + 1.040 \times \mathbf{^{13}}C^{\alpha}_{\mu}$. In effect, the correlation existing between averaged $^{13}C^{\alpha}$ chemical shift values computed for the 32 conformations of 1NS1 with these two basis sets, is excellent (Vila et al., 2009), i.e., showing a correlation coefficient $R = 0.999$ and standard deviation of 0.284 ppm. Even more important, an analysis of the magnitude of the errors and their distribution carried out for 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, indicates that ~70% of them are within ~0.6 ppm and that the most populated regions of the Ramachandran map are not affected by errors higher than ~1.0 ppm (Vila et al., 2009).

In conclusion, the described analysis enabled us to select the smaller basis set (6-31G/3-21G) that provides accuracy similar to that of a 'basis set limit' [6-311+G(2d,p)/3-21G] to reproduce the computed chemical shifts, but at a significantly lower computational cost (Vila et al., 2009).

3.4 Effect of Sequential Nearest-Neighbors on the $^{13}C^{\alpha}$ Chemical Shifts Calculations

The ^{13}C$^{\alpha}$ chemical shifts for a residue **X** in the model peptide Ac-G-**X**-G-NMe has always been computed (Vila et al., 2007b; Vila and Scheraga, 2009) considering that all the torsional angles of the residue **X** are *exactly* those of the residue in the protein conformation and that the surrounding Gly residues and the end-blocking groups are free to rotate. It is implicit in this approach that the ^{13}C$^{\alpha}$ chemical shifts of residue **X** do not depend on the identity of the nearest-neighbor residues. This assumption needs to be proved.

The structure of the Nucleic Acid Binding (NAB) protein of the SARS coronavirus (Serrano et al. 2009), a 116-residue α/β protein containing 9 Prolines (Pro) and with 50% of its residues in loops and turns, was chosen to further evaluate the origin of differences between computed and observed ^{13}C$^{\alpha}$ chemical shifts, as well as to study the influence of the nearest-neighbor residues on the computed ^{3}C$^{\alpha}$ chemical shifts.

The results (Vila et al., 2010) indicate that computation of the ^{13}C$^{\alpha}$ chemical shifts of a given residue in the sequence of the NAB protein is not influenced significantly, i.e., within ~0.5 ppm, by the nature of the nearest-neighbor amino acids, except for residues immediately preceding proline (see Figure 2a). For such residues, Pro must be considered during the computation of the ^{13}C$^{\alpha}$ chemical shifts; otherwise, an overestimation of the computed ^{13}C$^{\alpha}$ chemical shifts by about +1.7 ppm occurs. This finding is in good agreement with both the experimental evidence (Wishart et al. 1995a; Schwarzinger et al. 2001; Wang and Jardetzky 2002a) and the empirical observations (Wishart et al. 1995b; Schwarzinger et al. 2001). It is equally important to emphasize the physical nature of this effect: "*…an imide bond formed by an Xxx–Pro pairing is generally thought to be much less electron-withdrawing than an amide bond…*" (Wishart et al., 1995b).

Overall, except for the Pro effects, use of the Ac-G-**X**-G-NMe model peptide for the computation of the ^{13}C$^{\alpha}$ chemical shifts of residue **X** is a good approximation because the computed values are accurate within ±0.5 ppm for all residue-types, if neither the subsequent nor precedent residue-type effects are taken into account (see Figure 2).

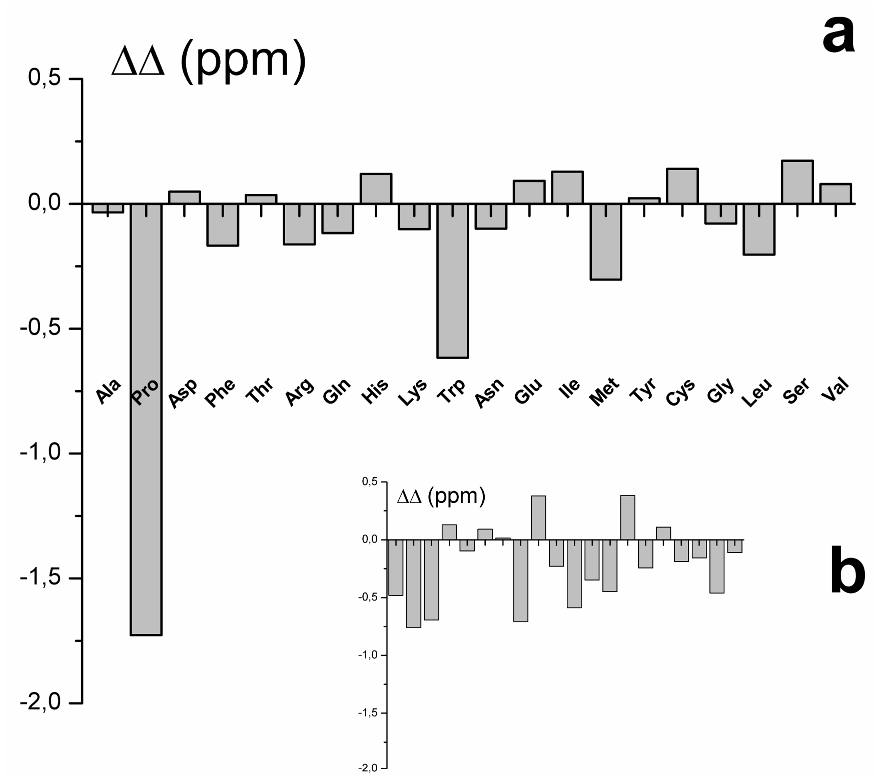

Fig. 2 Histogram of the average, over all 20 conformers of the protein PDB id 2K87, second-order differences ΔΔ: (**a**) with $\Delta\Delta = <(\Delta_X - \Delta_{YX})>$ arising from the nature of the sequentially preceding residue-type (**Yyy**). Δ_X and Δ_{YX} are the differences between the observed chemical shifts and those computed using the Ac-Gly-**Xxx**-Gly-NMe and Ac-Gly-Yyy-**Xxx**-Gly-NMe model peptides, respectively; (**b**) with $\Delta\Delta = <(\Delta_X - \Delta_{XY})>$ for the differences arising from the nature of the subsequent residue-type, i.e., with Δ_{XY} computed with Ac-Gly-**Xxx**-Yyy-Gly-NMe. Figure adapted from Vila et al., 2010 (with permission of Springer).

3.5 Rigid-Geometry Approximation and Accuracy of the Calculations of $^{13}C^{\alpha}$ Chemical Shifts

Experimental protein structures are often solved using force fields which allow variation of bond lengths and bond angles. However, it is known that QM calculations are very sensitive to bond lengths and bond angles (de Dios *et al.*, 1993a). Therefore, we have explored the dependence of the computed $^{13}C^{\alpha}$-chemical shifts on the bond lengths and bond angles to establish whether a rigid- rather than non-rigid geometry approximation is a more accurate representation with which to compute the chemical shifts.

For this test, the structure of ubiquitin deposited in the PDB (PDB id 1UBQ) was chosen because it possesses non-regularized geometry and has been solved by X-ray diffraction at 1.8 Å resolution (Vijay-Kumar et al., 1987). We have also examined the corresponding structure with regularized geometry, i.e., the one with all the residues replaced by the standard ECEPP residue geometry (Némethy et al., 1992), named here as 1UBQ^{regular}. Analysis of the differences between the computed and observed ^{13}C$^{\alpha}$ chemical shifts for the 1UBQ and 1UBQ^{regular} structures, leads to rmsd of 3.28 ppm and 2.38 ppm, respectively. The better agreement obtained with 1UBQ^{regular}, rather than 1UBQ, is consistent with the long-time recognition that the bond lengths and bond angles of both X-ray and NMR-derived structures are not as highly accurately defined as in studies of small molecules (de Dios et al., 1993a), with which the ECEPP geometry (Némethy et al., 1992) has been parameterized. Further analysis of the agreement of the two ubiquitin structures with the deposited electron density data (Vijay-Kumar et al., 1987) of 1UBQ, in terms of the *R*-factor, leads to 19.2% and 23.1% for 1UBQ and 1UBQ^{regular}, respectively; while the all-heavy-atom rmsd between these two structures is 0.142 Å (Vila and Scheraga, 2009).

Overall, the use of *regularized geometry*, i.e., ECEPP geometry, is an accurate approximation with which to compute the ^{13}C$^{\alpha}$ chemical shifts in proteins and, hence, is used in most of the application discussed in this chapter.

3.6 $^{13}C^{\alpha}$ Chemical Shifts as a Function of the Charge Distribution

Among the factors that affect ^{13}C$^{\alpha}$-shielding, which are important for an accurate computation of chemical shifts, is the sensitivity of ^{13}C$^{\alpha}$ nuclei to the shielding/deshielding induced by changes in the protonation/deprotonation of distant ionizable groups (Quirt et al., 1974; Sayer et al., 1976; Rabenstein and Sayer, 1976; Surprenant et al., 1980). However, these factors have not been taken into account explicitly in current computations of ^{13}C$^{\alpha}$ chemical shifts in proteins at the QM level of theory because, usually, the calculations are carried out in the gas phase, and the ionizable residues are treated as neutral groups.

The question of whether the use of neutral, rather than charged, side chains is more accurate for computation of the ^{13}C$^{\alpha}$ chemical shifts of ubiquitin, at a given fix pH, was investigated as follows (Vila and Scheraga, 2008). For a given ionizable residue *i* in a conformation *k*, first, the average charge distribution, $< \rho_{i,k} >$, was computed by using Equation (4), i.e., by explicit consideration of the 2^{ξ} ionization states for every conformation (Ripoll et al., 1996), with ξ being the number of ionizable groups in the molecule, namely 22; and second, the ^{13}C$^{\alpha}$ chemical shifts as a function of the pH, $\delta_i\ (pH)$, were computed by using Equation (5). This analysis was applied to 139 conformations of ubiquitin: 138 (10 conformations from PDB id 1D3Z plus 128 conformations from PDB id 1XQQ) NMR-derived conformations (Cornilescu et al., 1998; Lindorff-Larsen et al., 2005), while the remaining one is an X-ray structure (PDB id 1UBQ) solved at 1.8 Å resolution (Vijay-Kumar et al., 1987).

Additionally, an extra set of 50 randomly generated conformations for each amino acid residue **X,** in the terminally blocked tripeptide with the sequence Ac-G**X**G-NMe, with **X** being Lysine (Lys), Ornithine (Orn), Diaminobutyric acid (Dab), Glutamic acid (Glu) or Aspartic (Asp) acid, were also obtained. This set of randomly generated conformations was used to determine: (*i*) the range of shielding/deshielding of the $^{13}C^{\alpha}$ nucleus of free acidic/basic amino acid residues in solution, in their fully charged and neutral forms, respectively; (*ii*) how these ranges of shielding/deshielding variations compare with those derived from 3,058 ionizable groups of the 139 conformations of the protein ubiquitin; and (*iii*) how the computed shielding/deshielding range of variations are influenced by the distance between the charged side-chain group and the $^{13}C^{\alpha}$ nucleus (for example, there are two chemical bonds in Asp, rather than three in Glu, separating the deprotonated carboxyl group from the $^{13}C^{\alpha}$ nucleus). To examine an analogous effect for a basic side-chain group, such as Lys, use was made of the non-natural amino acids Orn and Dab because, for these amino acids, the protonated amino group is separated from the $^{13}C^{\alpha}$ nucleus by four and three chemical bonds, rather than by five in Lys.

The results of this study (Vila and Scheraga, 2008), based on the analysis of 139 conformations of ubiquitin at pH 6.5, indicate that use of neutral, rather than charged, amino acids is a significantly better approximation of the observed $^{13}C^{\alpha}$ chemical shifts in solution for the acidic groups, and a slightly better representation, though significantly less expensive computationally, for the basic groups (see Figure 3).

Additionally, our analysis of Lys, Orn, and Dab revealed a significantly greater deshielding of the $^{13}C^{\alpha}$ nucleus (due to the deprotonation of the acidic groups) than the shielding due to the protonation of the basic groups. The origin of such a difference can be found in the distance between the ionizable groups and the $^{13}C^{\alpha}$ nucleus, which is shorter for the acidic than for the basic groups.

3.7 $^{13}C^{\alpha}$ Chemical Shifts as a Function of Side-Chain Flexibility

To what extent are the chemical shifts of the amino acid residues in a protein affected by the side-chain orientation? The basis for such a query arises from the fact that the three torsion angles ϕ, ψ, and $\chi1$ are not independent on each other over the whole range because they involve a common N–C$^{\alpha}$ bond [Dumbrack and Karplus 1993, 1994; Chakrabarti and Pal 1998]. To find an answer to this question, the dependence of the ^{13}C chemical shifts on side-chain orientation was investigated (Villegas et al., 2007), at DFT level of theory, for two-strand antiparallel β-sheet model peptide with the amino acid sequence Ac-A$_3$-**X**-A$_{12}$-NH2 where **X** represents any of the 17 naturally occurring amino acids considered here, i.e., not including alanine, glycine, and proline. Because the majority of

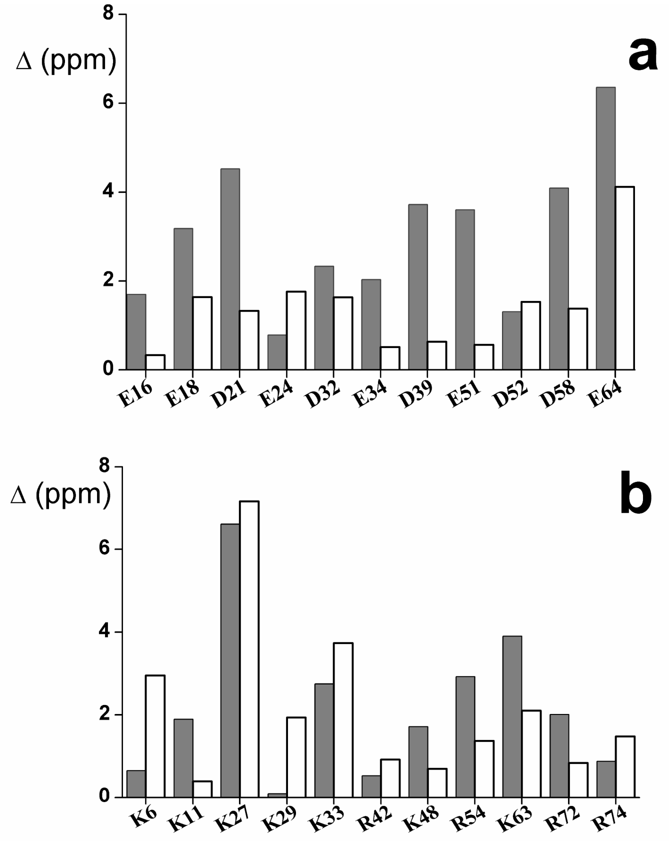

Fig. 3 Average difference, Δ, computed over a set of 9 conformations of protein ubiquitin using Equation (2) for: (**a**) acidic and (**b**) basic groups, respectively. Grey and white bars denote charged and neutral side-chain, respectively. Figure adapted from Vila and Scheraga, 2008 (with permission of John Wiley and Sons).

β-sheets are twisted, rather than planar, with a right-hand twist in the approximately ± 30° range for the backbone dihedral angles (Chothia *et al.*, 1977; Chou and Scheraga, 1982; Chou et al., 1982; Creighton, 1984) conformational parameters for β-sheets may deviate from those for planar pleated sheets and, hence, are difficult to model by using canonical values. The fact that β-sheets in proteins appear as parallel or antiparallel strands, or a combination of both, only exacerbates the modeling problem. For this reasons, the dihedral angles adopted for the backbone were taken, and kept fixed, from the experimental structure of an antiparallel β-sheet, specifically from the 16-residue segment (G41-G56) of the B3 binding domain of protein G (PDB id 1P7E).

For the 17 naturally occurring amino acids considered the analysis indicates that there is: (*a*) good agreement between computed and observed $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, i.e., with correlations coefficient, *R*, of 0.95 and 0.99, respectively; (*b*) significant variability of the computed $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts as function of χ^1 for all 17 residues, except for Ser; and (*c*) a smaller compared to χ^1, although significant, dependence of the computed $^{13}C^{\alpha}$ chemical shifts of χ^{ξ} (with $\xi \geq 2$) for 11 out of 17 residues.

The above results obtained by Villegas et al., (2007) for an antiparallel (16-residue segment) β-sheet were later validated on a 76 residues α/β protein, i.e., by exploring the effects of side-chain conformation on the computed $^{13}C^{\alpha}$ chemical shifts (Vila and Scheraga, 2008). This validation process involved an exhaustive conformational search, starting from an arbitrary selected conformation of the NMR-determined ubiquitin protein (PDB id 1D3Z), in which *only* the torsional angles of the side chains were allowed to vary, i.e., all backbone dihedral angles (ϕ,ψ,ω) were fixed at their corresponding observed values. Furthermore, the correlation coefficient, *R*, between computed, by using the Karplus equation (Karplus, 1959), and observed vicinal coupling constants $^{3}J_{N\text{-}C\gamma}$ and $^{3}J_{C'\text{-}C\gamma}$ of 17 valine, threonine, and Isoleucine residues, was used to check the accuracy of the side-chain conformational search.

The obtained results on an antiparallel β-sheet segment *and* the ubiquitin protein enabled us to determine the role and impact of a proper side-chain conformation for an accurate computation of the observed $^{13}C^{\alpha}$ chemical shifts in solution.

4 Use of the Structural Information Decoded from ^{13}C Chemical Shifts

We have chosen three examples to illustrate how the structural information decoded from the observed ^{13}C chemical shifts can be used in practice: (*1*) to determine the fraction of the tautomeric forms of the imidazole ring of histidine (His) in proteins as a function of pH, provided that the observed $^{13}C^{\gamma}$ and $^{13}C^{\delta 2}$ chemical shifts and the protein structure, or the fraction of H^+ form are known; (*2*) to determine *either* all α-helical or all β-sheet protein structures in solution; and (*3*) to assess the reliability of NMR-determined protein models before they are published or deposited in the PDB. Each of these applications is described in the following subsections.

4.1 The Importance of Being His

In 1965 Mandel, in a pioneering NMR experiment, detected the imidazole (C2) protons of histidine (His) residues in Ribonuclease A and in 1966, Bradbury and Scheraga, were able to distinguish between the histidine residues of Ribonuclease A, i.e., they resolved the NMR-peaks of three out of four histidines of this enzyme. Subsequently, use of NMR spectroscopy, X-ray crystallography and theoretical studies, based on QM calculations, have continuously evolved in their ability to determine properties of the histidine residues in solution and in the solid state (Meadow et al., 1968; Reynolds et al., 1973; Markley 1975; Wüthrich, 1976; Schuster and Roberts, 1979; Harbison et al., 1981; Munowitz et al., 1986; Bachovchin, 1986; Farr-Jones et al., 1993; Pelton et al., 1993; Steiner, 1996; Steiner and Koellner, 1997; Shimba et al., 1998, 2003; Sudmeier e*t al.*, 2003; Strohmeier et al., 2003; Cheng et al., 2005; Shimahara et al., 2007; Jensen et al., 2007; Hass et al., 2008; Hass et al., 2009; Hu et al., 2010; Vila et al., 2011). The reason for this persistent interest in His is due to the fact that this residue is unique among all 20 naturally occurring amino acids because ~50% of all enzymes use His in their active sites (Ulrich *et al.*, 2008). This is, mainly, because of the versatility of imidazole His ring, which includes two neutral, chemically-distinct forms, referred to as $N^{\delta1}$–H and $N^{\varepsilon2}$–H tautomers, and a protonated form, the charged H^+ form, with one form favored over the other two by the protein environment and pH. In addition, His with a pK^o of 6.6 (Demchuk and Wade, 1996) is the *only* ionizable residue that titrates around neutral pH, allowing the non-protonated nitrogen of its imidazole ring to serve as an effective ligand for metal binding (Hass et al., 2008), or to play a crucial role in the proton-transfer process (Hu et al., 2010).

Certainly, determination of the fraction of the tautomeric forms of the imidazole ring of His in proteins in solution is an important problem for a number of reasons. At a given fixed pH proteins in solution exist as an ensemble of conformations and, hence, the form of each His residue among different protein conformers may vary significantly because the tautomeric equilibrium is determined by the environment (Vila et al., 2011). Moreover, because the exchange between different protonation states is assumed to occur in the fast exchange regime (Hass et al., 2008), the NMR resonances of a given nucleus, which include rotation, protonation and tautomerization, merge into a single average signal. Decoding the information from these exchange processes offers possibility to determine the extent to which the His residues in proteins behave as free His, where the $N^{\varepsilon2}$-H tautomer is favored over the $N^{\delta1}$-H tautomer in a ratio of 4:1 (Reynolds, 1973).

To find a solution to this long-standing problem in the biophysical chemistry of proteins, first, each form of His was treated as a terminally-blocked model tripeptide with the sequence: Ac-G**H**$^{\xi}$G-NMe, with $\mathbf{H}^{\xi}$ in the $N^{\delta1}$-H, the $N^{\varepsilon2}$-H tautomeric form or the protonated form H^+, respectively. For each of the forms, a set of ~35,000 conformations, representing a uniform sampling of the whole Ramachandran map as function of ϕ, ψ, ω, $\chi1$ and $\chi2$ torsional angles, was generated. Afterward, the gas-phase, isotropic shielding value was computed using the method described in section 2.1. Finally, the distribution of the computed shielding of the imidazole ring of His was analyzed in terms of all ^{13}C nuclei, namely $^{13}C^{\gamma}$, $^{13}C^{\delta2}$, and $^{13}C^{\varepsilon1}$ (see Figure 4). Specifically, the histogram of the shielding distribution (among all ~35,000 conformations) was fit by a Gaussian

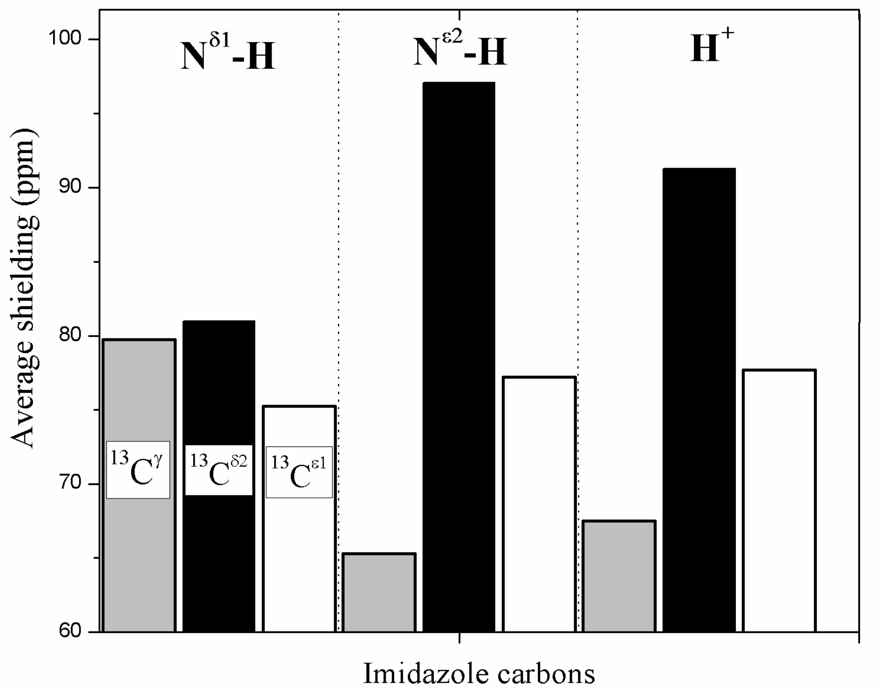

Fig. 4 Bar diagram of the average σ_o shielding values computed for each carbon of the imidazole ring of His for each of the two tautomers: $N^{\delta1}$-H, $N^{\varepsilon2}$-H, and for the H^+ form. The values were averaged over ~35,000 conformations of histidine in the model tripeptide Ac-GHG-NMe. Grey, black and white colors indicate the results obtained for the $^{13}C^{\gamma}$, $^{13}C^{\delta2}$ and $^{13}C^{\varepsilon1}$ nuclei, respectively. Figure adapted from Vila et al., 2011 (with permission of PNAS).

function with a mean value σ_o (shown as bars in Figure 4) and standard deviation *sd* (data not shown). A visual inspection of the histogram shown in Fig 4 revealed that the mean σ_o shielding values obtained for the $^{13}C^{\varepsilon1}$ nucleus is not sensitive to changes in the form of the imidazole ring and, therefore, we confine our interest to those nuclei that are sensitive to such changes, namely $^{13}C^{\delta2}$ and $^{13}C^{\gamma}$.

Use of first-order shielding differences for a pair of selected nuclei, $^{13}C^{\delta2}$ and $^{13}C^{\gamma}$, rather than chemical shifts, is a very convenient approach because the experimental referencing problem may be a source of errors (Cheng et al., 2005). Consequently, we define the first-order shielding difference, Δ^{ξ}, as $\Delta^{\xi} = |\sigma_o^{\delta2} - \sigma_o^{\gamma}|^{\xi}$, with ξ denoting the form of the imidazole ring, and $\sigma_o^{\delta2}$ and σ_o^{γ} are the computed mean values of the shielding distribution for the $^{13}C^{\delta2}$ and $^{13}C^{\gamma}$ nuclei, respectively. In other words, the following convention is adopted: $\xi = \delta, \varepsilon$, or +, to designate the $N^{\delta1}$-H, $N^{\varepsilon2}$-H, or the H^+ form, respectively.

Analysis of the first-order shielding differences indicates that the following inequality holds: $\Delta^{\varepsilon} > \Delta^{+} > \Delta^{\delta}$, and $\Delta^{\delta} \sim 0$. Therefore, once the fraction of protonated H$^+$ form, $f^{+} = \ <\rho>$, computed with Eq. (4), and $\Delta^{obs} = |^{13}C^{\delta 2} - {}^{13}C^{\gamma}|$, with $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ being the observed chemical shifts in solution, at a given pH, are known, the fraction of the N$^{\varepsilon 2}$-H tautomer (f^{ε}) can be obtained assuming: (*a*) that all forms are in fast exchange on the NMR chemical shift time-scale (Hass et al., 2008), i.e., as: $\Delta^{obs} = f^{\varepsilon}\Delta^{\varepsilon} + f^{+}\Delta^{+} + f^{\delta}\Delta^{\delta}$; and (*b*) that $\Delta^{\delta} \equiv 0$.

Using these assumptions, together with some physical constraints, enable us to find an analytical expression with which to compute f^{ε}, namely as: $f^{\varepsilon} = \dfrac{\Delta^{obs}(1-\langle\rho\rangle)}{\Delta^{\varepsilon}}$, with Δ^{ε} the single-valued first-order shielding difference computed for the N$^{\varepsilon 2}$-H tautomer ($\Delta^{\varepsilon} \sim 31$ ppm). The fraction of the f^{δ} tautomer is obtained straightforwardly as: $f^{\delta} = 1 - <\rho> - f^{\varepsilon}$.

The above formulation was used to determine the tautomeric forms of His for each of 8 selected proteins for which both the structure and the $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ chemical shifts of the imidazole ring of His, are available. In each of these applications the average degree of protonation $<\rho>$ for all ionizable residues was computed by using Eq. (4). The tautomeric forms of His are determined by using the expressions for f^{δ} and f^{ε} given above (Vila et al., 2011). Likewise, using the observed values, Δ^{obs}, obtained from solid-state NMR for unblocked dipeptides, with the sequence His-Leu, His-Met, Gly-His, Leu-His, His-Ala, His-Glu, Ala-His, and His-Asp (Cheng et al., 2005), we also determined the tautomeric fractions of the imidazole ring of His for each of these 8 compounds.

Results obtained from the 8 proteins indicate that the protonated form is the most populated one while the distribution of the tautomeric forms for the imidazole ring varies significantly among different histidine residues in the same protein (see Figure 5a). Thus, His226 and His250 show comparable degree of protonation, $<\rho>$, although the tautomeric distribution is very different (see Figure 5a), i.e., showing the importance of the environment of the histidines in determining the tautomeric forms. Let us explain the origin of this observation. On one hand, the N$^{\delta 1}$ nucleus of H250 is located only 2.9 Å from the carbonyl backbone oxygen of S248 (see Figure 5b), presumably forming a hydrogen-bond (green dots in Figure 5b), while the N$^{\varepsilon 2}$ nucleus is exposed to the solvent but the imidazole ring is surrounded by fully protonated R264 and R266 (data not shown) and, hence, lowering the probability that a proton binds to N$^{\varepsilon 2}$, in good agreement with the computed tautomeric distribution for H250 in Figure 5a. On the other hand, the N$^{\varepsilon 2}$ nucleus of the imidazole ring of H226 is at 3.3 Å from a backbone carbonyl oxygen of W246 (see Figure 5c), while the N$^{\delta 1}$ is at 3.1 Å from a backbone amino group of H226 (see Figure 5c). As a result, a preference of N$^{\varepsilon 2}$-H over the N$^{\delta 1}$-H tautomeric form for H226 is expected, in agreement with the computed tautomeric fractions for this residue in Figure 5a.

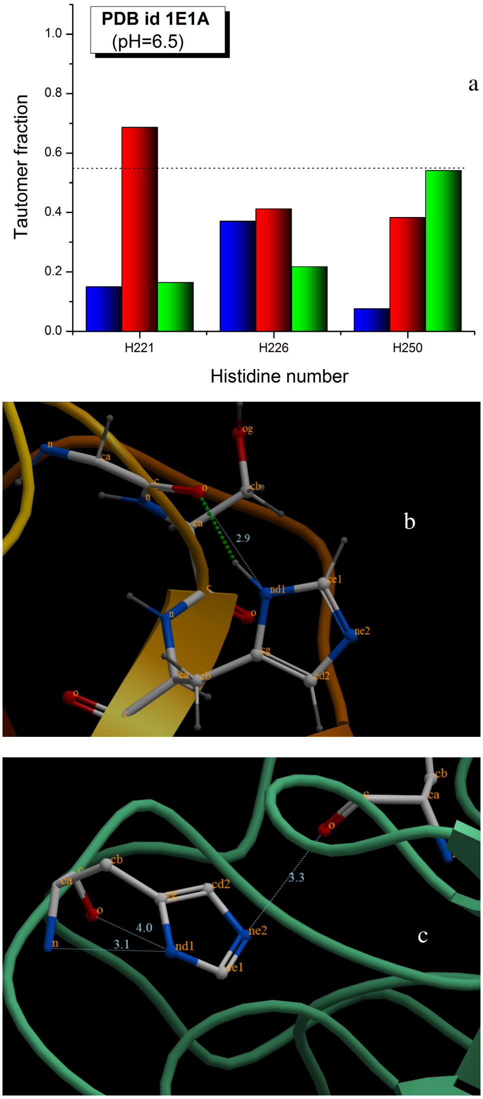

Fig. 5 (**a**) Fraction of His form distribution for 3 out of 6 His residues in protein PDB id 1E1A, for which the chemical shifts were determined in solution at pH 6.5. Blue and green bars represent the fraction of the $N^{\varepsilon2}$-H and $N^{\delta1}$-H tautomers, respectively, and the red bars represent the fraction of the protonated form, H^+. The dotted horizontal line indicates the fraction of the H^+ form that a free His residue would have in solution at pH 6.5; (**b**) Ball and stick representation of H250 in protein 1E1A. The grey, blue and red colors designate carbon, nitrogen and oxygen atoms, respectively. The background shows a ribbon diagram of part of protein 1E1A. The $N^{\delta1}$ nucleus of H250 is located at only 2.9 Å from the carbonyl backbone oxygen of S248, presumably forming a hydrogen-bond (indicated by green dotted line); (**c**) Same as (**b**) for H226. All displayed distances are in Angstroms. Figure (**a**) adapted from Vila et al., 2011 (with permission of PNAS).

In addition, our results show that for ~70% of the neutral histidine-containing dipeptides the method leads to fairly good agreement between the calculated and the experimental tautomeric form. Co-existence of different tautomeric forms in the same crystal structure may explain the disagreement obtained for the remaining 30% of dipeptides.

4.2 Protein Structure Determination

In this section we illustrate, with two examples, how the structural information encoded in the $^{13}C^{\alpha}$ chemical shifts can be used to determine an ensemble of conformations, provided that a set of NOE-derived distance constraints, is available. However, since the chemical shifts are sensitive to the dynamics of a protein on the microsecond time scale (Lindorff-Larsen et al, 2005) the question whether a single rather than an ensemble of conformations is a better representation of the NMR observables, such as the chemical shifts, must be investigated first.

4.2.1 The Crystallographer Dilemma: A Single Structure or an Ensemble of Conformations?

In protein crystallography it is conventional to represent the conformation of a protein by a single structure, although proteins are very flexible in solution, and, hence, the question whether a single structure, rather than an ensemble of conformations, is a more accurate representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution deserves to be investigated.

Proteins in solution are flexible molecules which exhibit anisotropic motion and exist as a dynamic ensemble of conformations. Although, protein flexibility in the crystalline state is reduced (compared to solution) as a result of crystal packing, some dynamics and heterogeneity still remain (Ringe and Petsko, 1986; DePristo *et al.*, 2004) because of the high solvent content in most protein crystals (Jensen, 1997). Despite this, protein structures solved by X-ray diffraction are traditionally represented by a single conformation. Crystallographic temperature (B) factors, which contain information about atomic displacements arising from the combined effects of dynamic, static, and lattice disorders within the crystal lattice, provide an important indication of protein motions in the crystalline state.

Consequently, consideration of an ensemble of protein conformations generated by using B-factor values as a guide *may* potentially improve the agreement between the NMR- and X-ray-derived protein models in terms of some NMR observables, such as $^{13}C^{\alpha}$ chemical shifts. To explore such possibility we selected ubiquitin, an α/β 76 residues protein. The structure of this protein was solved by X-ray [PDB id 1UBQ (Vijay-Kumar *et al.*, 1987)], and NMR [PDB id 1D3Z (Cornilescu et al., 1998)] methods, with the latter providing the available $^{13}C^{\alpha}$ chemical shifts.

Since the deposited PDB structures of 1UBQ were solved and refined by using software and force-field parameters different from those employed in our method, a new set of conformations was generated using MCM and rigid geometry starting from the corresponding regularized experimental X-ray structure (1UBQ^{regular}). During the MCM search, variations of the (ϕ, ψ, χ) torsional angles were allowed for all the residues in the sequence. The reported B-factors for 1UBQ were used to estimate the upper limit of the torsional angle variation adopted (±10^{o}). The generated set of conformations was subjected to several rounds of refinement using a standard procedure in X-ray crystallography, i.e., the Crystallography and NMR System (CNS) program (Brünger *et al*, 1998; Brünger, 2007). As a result 5 conformations were selected.

All the 5 generated models are quite different among themselves and from the corresponding starting structure, with an all-atom rmsd of 0.36-1.13 Å. Moreover, for all 5 models, no residues were in disallowed regions of the Ramachandran plot (Laskowski et al.,1993) and all unfavorable contacts occur between the atoms from the last five residues in the sequence, which were not visible in the electron-density map. In addition, the R and R_{free} factors of the 5 models are equivalent to or better than those of the one obtained for a Simulated Annealing Refined (SAR) structure of PDB 1UBQ. This refinement of the deposited 1UBQ structure i.e., named SAR structure, is a necessary step for a consistent comparison between the chemical shifts of the generated 5 models and the PDB structure, because C^{13} chemical shifts are very sensitive to small differences in bond lengths and bond angles (de Dios *et al.*, 1993a).

Figure 6 shows the rmsd values between the observed and computed ^{13}C$^{\alpha}$ chemical shifts obtained for each of the 5 new models (light-gray bars) and the SAR structure (black-filled bar). The *ca*-rmsd, computed from the ensemble of 5 new models, is shown as a horizontal solid line in Figure 6. The *ca*-rmsd (2.36 ppm) is lower than the value for the SAR structure (2.74 ppm) or for any of the new models. These results obtained for ubiquitin demonstrate that consideration of an ensemble of 5 conformations, derived from the regularized experimental X-ray (1UBQ^{regular}) structure, leads to better agreement with the observed ^{13}C$^{\alpha}$ chemical shifts than does a single conformation (the SAR structure).

The above conclusion is in line with the suggestion of crystallographers' that *"…a more suitable representation of a macromolecular crystal structure would be an ensemble of models..."* Furnham et al., 2006. Analysis of NMR-determined ensemble of conformations also lead to similar conclusion, i.e., use of the *ca*-rmsd value led to closer agreement with the observed ^{13}C$^{\alpha}$ chemical shifts in solution than when individual, or the mean, rmsd is used (Vila et al., 2007a). In other words, proteins in solution are conformationally labile, as indicated by both the *ca*-rmsd and the theoretical minimal-rmsd model analyses, and this must be taken into account to predict the ^{13}C$^{\alpha}$ chemical shifts most accurately.

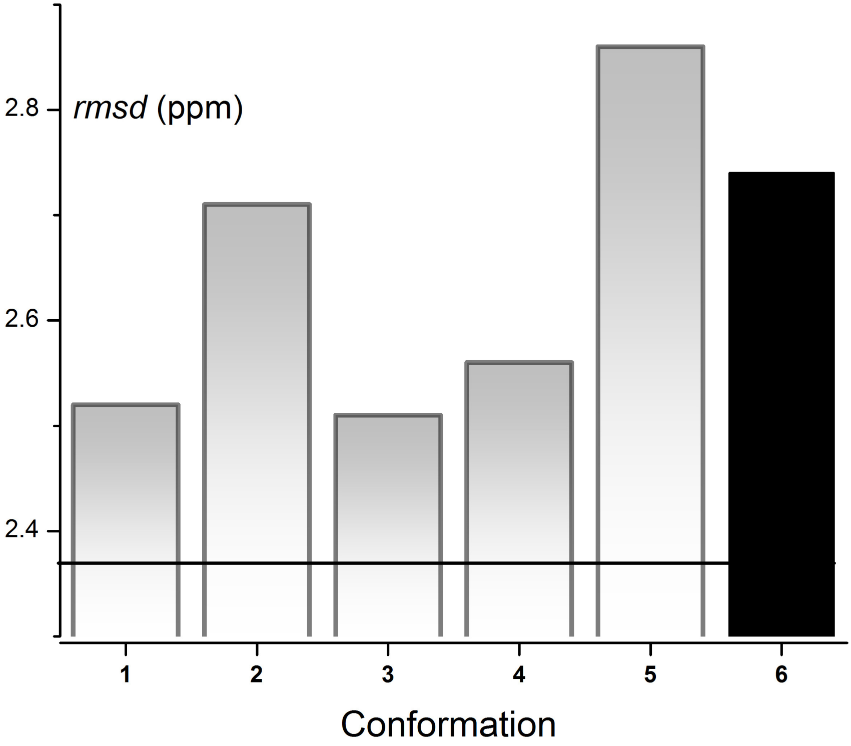

Fig. 6 Bar diagram of the rmsd (ppm) between the computed and observed $^{13}C^{\alpha}$ chemical shifts of ubiquitin. Black-filled bar (2.74 ppm) represents the results from the SAR structure. Grey-filled bars represent the rmsd for each of the generated 5 new models; the horizontal black line represents the *ca*-rmsd (2.36 ppm) computed from the ensemble of 5 new models. Figure adapted from Arnautova et al., 2009 (with permission of the International Union of Crystallography).

4.2.2 Determination of β-Sheet Structures

Evidence obtained from the probability-based secondary structure identification method of Wang and Jardetzky (2002b) suggests that the reliability to distinguish an α-helix from a statistical coil based on chemical shift information follows, for the heavy nuclei *only*, the ranking: $^{13}C^{\alpha} > {}^{13}C' > {}^{13}C^{\beta} > {}^{15}N$, whereas a different trend ($^{13}C^{\beta} > {}^{13}C^{\alpha} \sim {}^{13}C' \sim {}^{15}N$) was found for the corresponding reliability to distinguish a β-strand conformation from a statistical coil. This trend raises the question whether a mainly $^{13}C^{\alpha}$-driven methodology can be used to predict predominantly β-sheet structures and, if so, how well the corresponding $^{13}C^{\beta}$ chemical shift predictions would be.

To answer this question, our recently introduced physics-based protocol (see Figure 1) was applied to determine the structure a 20-residue peptide capable of forming a three-stranded antiparallel β-sheet in aqueous solution, i.e., the BS2 peptide with the sequence: TWIQN$_D$PGTKWYQN$_D$PGTKIYT, for which both a complete set of $^{13}C^{\alpha}$ chemical shifts and a reduced number of NOEs were reported. The experimental structure determination of small proteins and peptides, which are able to fold as monomers and do not contain disulfide bonds, is very valuable because such determinations can provide important information for force-field development and evaluation or improvement of search algorithms aimed at an efficient exploration of the conformational space (Jang et al, 2007; Zhou, 2003; Mohanty and Hansmann, 2006; Höfinger et al., 2007).

The results obtained indicate that an accurate *all* β-sheet structure can be determined by simply identifying a set of conformations which simultaneously satisfy a set of constraints including $^{13}C^{\alpha}$-dynamically derived torsional angle constraints for *all* amino acid residues in the sequence *and* a fixed set of NOE-derived distance constraints (Vila et al., 2008a). Among the thousands of conformations generated by the VTF approach, i.e., during the step 1 of the protein structure determination protocol shown in Figure 1, 25 of them (see Figure 7a) were selected by using a clustering procedure. This small set of conformation was used to determine the *theoretical minimal-rmsd model* that provides us with a set of φ, ψ, and χ torsional angle constraints for *all* the residues in the sequence not just for those in α-helix or β-sheet regions. Using this set of torsional angle constraints (φ, ψ, χ), combined with different number of NOE-derived constraints, 2 sets of conformations of the BS2 peptide were determined after the step 2 of the protocol. One set of 20 conformations (shown in Figure 7b) was obtained by using 118 NOE-derived distance constraints, while the other set of 10 conformations (shown in Figure 7c) was obtained by using 130 NOE-derived distance constraints. Regardless of the number of the NOE's-derived distance constraints used, addition of the $^{13}C^{\alpha}$-derived torsional constraints led to a noticeably lower *ca*-rmsd's (2.2 and 3.5 ppm, for the set of 20 and 10 conformations, respectively) compared to the 20 models obtained by Santiveri et al. (2004) who used a full set of 130 NOE's-derived distance constraints *but* no $^{13}C^{\alpha}$ chemical shift information (4.6 ppm). In line with this finding, graphical inspection of the results shown in Figure 7b-c also indicated that use of $^{13}C^{\alpha}$-derived torsional constraints led to sets of conformations with less side-chain torsional angle spreading, i.e., as can be seen from comparison of Figures 7b and 7c against 7d, with the latter obtained by Santiveri et al (2004). In addition, the correlation coefficient, *R*, between the observed and computed $^{13}C^{\beta}$ chemical shifts was somewhat better for the two sets obtained using the $^{13}C^{\alpha}$-based determination protocol (shown in Figure 1). Thus, *R* is 0.99 and 0.98 for the 20 and 10 conformation sets, respectively, while R is 0.97 for the set of conformation derived by Santiveri et al (2004).

Overall, analysis of the *ca*-rmsd, the NOE-derived distance violations, the $^{13}C^{\beta}$ chemical shifts, and some stereo chemical quality factors for these sets, as a measure of the closeness with which the calculations reproduce the structure in solution, indicates that our self-consistent physics-based method is able to produce a more accurate set of conformations (shown in Figure 7b and 7c) than that

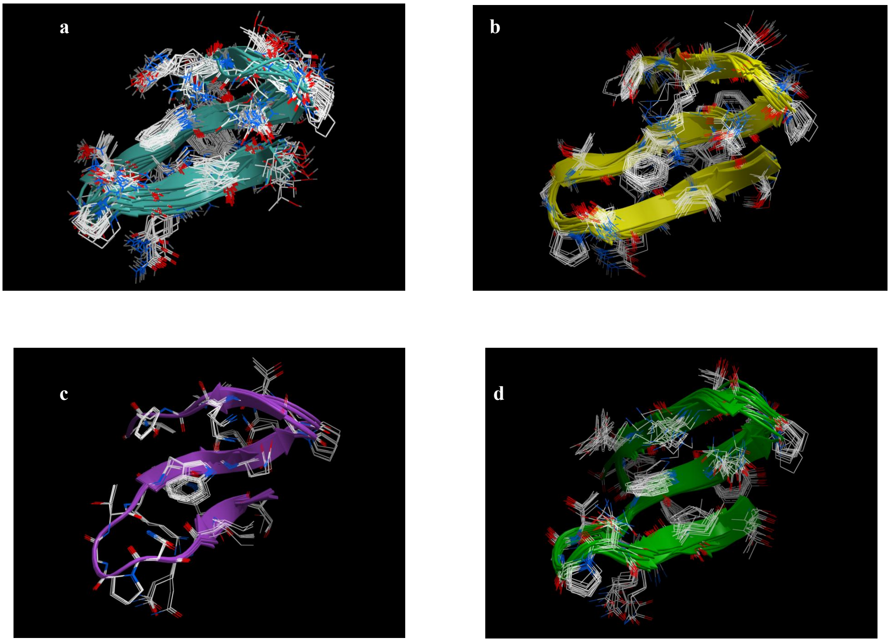

Fig. 7 (**a**) Superposition of 25 NMR-derived conformations of BS2 peptide (represented by ribbon diagrams) obtained in Step 1 after the VTF procedure (see Flow-chart in Figure 1); (**b**) Superposition of 20 NMR-derived conformations of BS2 obtained after the conformational search in Step 2 (see Flow-chart in Figure 1); 118 out of 130 NOE's distance constraints were used; (**c**) Same as (**b**) for 10 NMR-derived conformations; 130 NOE's distance constraints were used; (**d**) Superposition of 20 NMR-derived conformations obtained by Santiveri et al. (2004) using traditional methods

obtained with the traditional methods (Santiveri et al., 2004) [shown in Figure 7d]. Our results also suggest that for a flexible molecule in solution, like BS2, it may not be possible to determine a single structure that would satisfy *all* the constraints simultaneously. This is a consequence of the well-known fact (Constantine et al., 1995) that NMR parameters, such as the observed NOE-derived distances and the $^{13}C^{\alpha}$ chemical shifts, correspond to a dynamic ensemble of conformations and, therefore, may not be reproduced exactly by a limited set of static structures (Zhao and Jardetzky, 1994; Vila et al., 2007b).

Characterization of the structural flexibility of molecules in solution is of fundamental importance for the study of biological function, stability, and folding (Korzhnev et al., 1997; Palmer, 2004). Therefore, additional analysis of the per-residue average $^{13}C^{\alpha}$ *conformational shifts* was carried out and the results indicated that the third, C-terminal, strand in the β-sheet of the BS2 peptide is the most flexible strand, although less flexible than the turns. In addition, a 20 ns molecular dynamics simulations (MD) using the AMBER 8.0 package (Case et al.,

2004) were performed. The MD runs yielded a plausible atomic description of the motion of BS2 peptide in solution, as revealed by both the pattern of hydrogen bonds and the generalized Lindemann parameter (Zhou et al., 1999). The MD results were in line with the per-residue average $^{13}C^{\alpha}$ conformational shifts analysis, providing additional evidence of greater flexibility of the C-terminal strand.

The fact that the observed $^{13}C^{\alpha}$ chemical shifts, supplemented only by NOE-derived distance constraints, provide accurate information for validation and refinement of protein structures, as well as site-specific information about the flexibility of a molecule in solution, may be very useful for NMR spectroscopists and theoreticians interested in analysis of the stability and protein-folding mechanism.

4.2.3 A Blind Test to Determine an α–Helical Structure

The solution NMR structures of both full length (residues 1 to 77) and truncated (residues 1-46) forms of YnzC protein (PDB id 2JVD) from *Bacillus subtilis* (Kuzin et al., 2008), that is part of the small yneA SOS response operon that regulates cell division in this organism (Kawai, et al., 2003), have been determined recently (Aramini et al., 2008). The corresponding X-ray crystal structure (PDB ID, 3BHP) was solved by Kuzin et al. (2008) at 2.0 Å resolution. The unique two-helix monomeric structure of YnzC, with no disulfide bonds, makes it an attractive subject for testing our physics-based methodology for protein structure determination.

The goal of this application is two-fold. First, as a blind test, we attempted to determine whether it is possible to obtain an ensemble of conformations for which each individual conformer simultaneously satisfies the NOE-derived distance constraints and the $^{13}C^{\alpha}$-derived torsional constraints for the YnzC protein in solution (Vila et al., 2008c). Although the solution NMR structure (Aramini et al., 2008) of this protein had been solved at the time of this blind test, the *only* information provided was a full set of both the observed $^{13}C^{\alpha}$ chemical shifts and the NOE-derived distance constraints. In particular, no information about the coordinates of the solved structures of the YnzC protein (Aramini et al., 2008) or the heteronuclear ^{15}N-^{1}H NOE data was provided at the moment of the test.

Our second goal was to carry out a cross-validation test of high-quality sets of conformations obtained for the YnzC protein in solution by using alternative determination methods, namely, the solution NMR set of conformations (PDB id, 2JVD) obtained by using NOE-derived distance constraints, dihedral-angle constraints and hydrogen-bond constraints (Aramini et al., 2008), and the 2.0-Å X-ray crystal structure (PDB id, 3BHP) [Kuzin et al. 2008]. For this second goal, several validation scores were used (Vila et al., 2008c), including: (*i*) Recall, Precision, F-measure (RPF) analysis (Huang et al., 2005); (*ii*) several global quality score indicators provided by Verify3D (Lüthy et al., 1992), ProsaII (Sippl, 1993), Procheck (Laskowski et al., 1993), and MolProbity (Davis et al., 2007);

(*iii*) the *ca*-rmsd and rmsd between observed $^{13}C^{\alpha}$ chemical shifts and those computed at the DFT level, and (*iv*) the backbone rmsd between these refined structures and the mathematical average coordinates of the ensemble of NMR structures of YnzC(1-48) deposited in the PDB.

By carrying out a blind test we demonstrated (Vila et al., 2008c) that an accurate all α-helical set of protein structures can be determined by simply identifying conformations which simultaneously satisfy a set of constraints, including $^{13}C^{\alpha}$-dynamically derived torsional angle constraints for all amino acid residues in the sequence *and* a fixed set of 1,022 NOE-derived distance constraints. The protein structure determination was carried out as follows: after generation of thousands of conformations using the VTF procedure (step 1) 10 of them, shown in Figure 8b, were selected, i.e., those possessing a maximum NOE-derived distance violation lower than some fixed cutoff value; only one of the ten conformations produced in step 1 was selected. The selected conformation was used as a starting one in a conformational search carried out with two types of constraints: the original fixed limited NOE-derived distance constraints and the set of ϕ, ψ, χ torsional angles derived from step 1. The resulting new set of 10 conformations is shown in Figure 8c. Repetition of the step 2 with a tighter tolerance range, than in the previous iteration, for the torsional angle constraints enabled us to determine the final set of 10 conformations shown in Figure 8d, i.e., the so-called Set-NOE-CS.

A comparative analysis of the rmsd, between the computed and observed $^{13}C^{\alpha}$ chemical shifts values for the residues 1-46, for all three sets of conformations is shown in Figure 8a as a bar diagram, viz., the Set-NOE-CS (shown in Figure 8d), 2JVD (shown in Figure 8e), and the three chains of the X-ray crystallography structure 3HBP (shown in Figure 8f). The results shown in Figure 8a reveals that the two NMR-derived ensembles of structures (2JVD and Set-NOE-CS) are a better representation for the observed $^{13}C^{\alpha}$ chemical shifts in solution in terms of the *ca*-rmsd (solid horizontal black and red lines in Figure 8a), than any single conformer (red or yellow bars in Figure 8a), or any single chain of the X-ray structure (black, cyan and green bars in Figure 8a). This result is in line with previous calculations for 10 NMR-derived conformations (PDB id 1D3Z) and the X-ray structure (PDB id 1UBQ) of ubiquitin.

Since the *ca*-rmsd analysis might be biased by the fact that the 10 conformations of Set-NOE-CS were computed using a $^{13}C^{\alpha}$-based method while the others were not, a cross-validation quality test was also carried out. These structures consistently show good values for the RFP and DP-scores as well as for global structure quality indicators. This analysis reveals that *all* three sets of structures analyzed here display very good agreement with the experimental NOE data, as well as dihedral angle distributions and atomic clash scores typical of good quality protein structures. Taken together, these results indicate that the 20 conformations from the 2JVD set, the DFT-computed 10 conformations from Set-NOE-CS, and each of the three chains of the X-ray structure are highly accurate sets of conformations which represent the YnzC protein in solution.

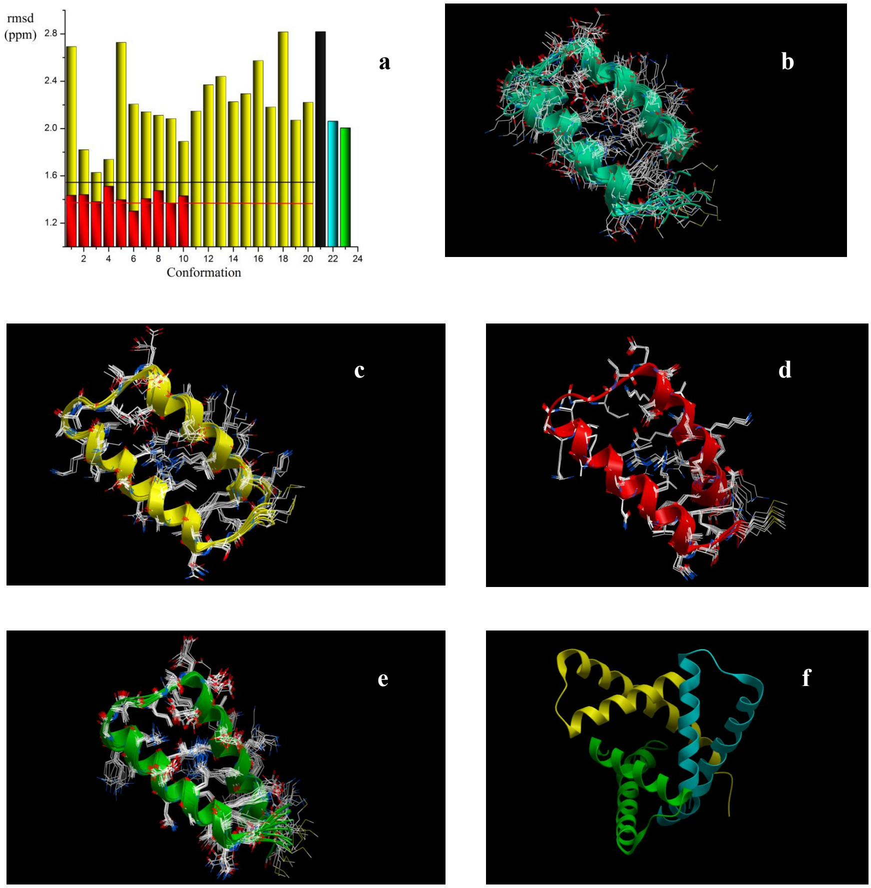

Fig. 8 Results for the 77-residue YnzC protein from *Bacillus subtilis.* (**a**) Bar diagram indicating the rmsd (ppm) between the computed and observed $^{13}C^{\alpha}$ chemical shifts for each of the 10 conformations from Set-NOE-CS (red bars), for the 20 conformations from 2JVD (yellow bars), and for each of the three chains in the 2.0 Å crystal structure of YnzC protein, PDB id 3BHP, namely chain *a*, *b* and *c* (black, cyan and green bars). Black (1.54 ppm) and red (1.38 ppm) horizontal lines show the *ca*-rmsd values computed for the residues 1-46 of 2JVD and Set-NOE-CS, respectively; (**b**) Superposition of 10 NMR-derived conformations of YnzC (represented by ribbon diagrams) obtained after the VTF procedure, in Step 1 (see Flow-chart in Figure 1); (**c**) Same as (**b**) after the conformational search in Step 2; (**d**) Same as (**c**) after repeating the conformational search in Step 2 (Set-NOE-CS), i.e., this time by using a *new* set of torsional angles (φ,ψ,χ) derived from the set of conformations shown in panel (**c**); (**e**) superposition of 20 NMR-derived conformations (PDB id 2JVD) of YnzC protein obtained by Aramini et al. (2008); and (**f**) Graphic representation of the X-ray determined structure of YnzC protein (PDB id 3HBP); the asymmetric unit contains 3 similar, but not identical, copies of the YnzC protein molecule, namely chain *a*, *b* and *c*. Figure (**a**) adapted from Vila et al., 2008c (with permission of PNAS).

4.3 Protein Structure Validation

The PDB is the most important archive of experimental protein structures solved by X-ray crystallography and NMR spectroscopy. The large number of structures deposited in PDB constitutes an extraordinary source of information that has been, and continuously is, used for a wide range of applications in structural drug design, molecular modeling, force-field parameterization, molecular biology applications, etc. Some deposited protein structures, showing few, or a large number, of flaws, are formally *withdrawn* from the data-base and, hence, considered as *obsolete*, even though their coordinates remain available in PDB. In most cases, a successor (or superseded) structure replaces the old obsolete one. The large number of obsolete structure indicates that development of accurate validation protocols remains an important task.

4.3.1 A Chemical-Shift-Based Server

An *ideal* validation method should meet two requirements. First, it should be *strong* rather than *weak*. A validation method is considered ‘strong’ if it is able to assess how well a structure, or an ensemble of structures, predicts experimental data *not* used in the structure-determination process; otherwise it should be considered ‘weak’, since it is limited to reproducing the observed experimental data used in the determination of the protein models (Kleywegt, 2009). Second, it should be able to detect *fast* and *accurately*, at residue level, the existence of structural flaws. With these goals in mind a new server (*Che*Shift) has been developed recently to predict $^{13}C^{\alpha}$ chemical shifts of protein structures. It is based on a database of chemical shifts computed for 696,916 conformations as a function of the ϕ, ψ, ω, $\chi1$ and $\chi2$ torsional angles for *all* 20 naturally occurring amino acids. The $^{13}C^{\alpha}$ chemical shifts were computed at the DFT level of theory using the methodology described in section 2.1. Because of the large number of conformations, the computed shielding values were obtained using a small basis set (6-31G/3-21G) and later extrapolated to a large basis set [6-311+G(2d,p)/3-21G], as described in Methods section.

An analysis of the accuracy and sensitivity of the *Che*Shift predictions, in terms of the correlation coefficient *R* between the observed and predicted $^{13}C^{\alpha}$ chemical shifts, was carried out on 36 X-ray-derived protein structures solved at 2.3 Å, or better, resolution. Results indicate that for all the proteins the *R* values obtained using the *Che*Shift, SHIFTX (Neal et al., 2003), SPARTA (Shen and Bax, 2007), SHIFTS (Xu and Case, 2001, 2002), and PROSHIFT (Meiler, 2003) servers were comparable, although the *Che*Shift values were systematically lowest. This raises the following question: do these servers provide a more sensitive validation than *Che*Shift? To answer this question we choose protein 1RGE, solved at 1.15 Å resolution (Sevcik et al., 1996). The corresponding crystal structure of this protein contains two chemically identical but crystallographically independent molecules in the asymmetric unit, named here as A and B (Sevcik et al., 1996). The main structural difference between molecules A and B (with an all-heavy-atom rmsd of 1.1 Å) is due to differences in side chain conformations, especially those occupying different rotameric states. For this test, that do not require a comparison with the observed $^{13}C^{\alpha}$ chemical shifts, we computed the correlation coefficient *R* between the $^{13}C^{\alpha}$ chemical-shift predictions obtained for molecules A and B, respectively, by using five servers listed above. 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. Except for *Che*Shift (0.96) and SHIFTS (0.98), 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 (1.00), SPARTA (1.00), and PROSHIFT (1.00) servers indicate that molecules A and B are practically indistinguishable protein models. Therefore a lower *R* value between the predicted and observed $^{13}C^{\alpha}$ chemical shifts does not necessarily mean poorer accuracy but it could mean higher *sensitivity* to subtle structural differences. This conclusion can be confirmed by a similar analysis carried out at a higher level of accuracy, for example, by using a larger basis set and the actual geometry of chains A and B, i.e., without need for any torsional angle interpolations as with the *Che*Shift server. In this case, the *R* value (0.93) computed with the larger basis set was 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.

So far, we have shown that the QM basis of the *Che*Shift server enables us to predict the $^{13}C^{\alpha}$ chemical shifts with reasonable accuracy in seconds. Our results suggest that *Che*Shift can provide a standard with which to evaluate the quality of protein structures solved by either X-ray crystallography or NMR-spectroscopy, if the experimentally observed $^{13}C^{\alpha}$ chemical shifts are available.

4.3.2 *Che*Shift-2: A Picture is Worth a Thousand Words

Differences between the observed and *Che*Shift-predicted $^{13}C^{\alpha}$ chemical shifts can be used as a sensitive probe with which to detect possible local flaws in NMR-determined protein structures; hence, a graphical user interface has been added to the *Che*Shift-2 server (Martin et al., 2012) to render such flaws easily visible. *Che*Shift was originally developed to return a list of $^{13}C^{\alpha}$ predicted chemical-shift values, one for each amino acid in the sequence of a protein, except for the first and last residues (Vila et al., 2007a; 2009). The validation process, i.e., the comparison between the predicted and the observed $^{13}C^{\alpha}$ chemical-shift values, is left to the user of the server who can use the provided information to determine the quality of the NMR structure as a whole, e.g., by computing the *ca*-rmsd (Vila et al., 2007a). However, it is a highly desirable goal of any accurate validation method (Nabuurs et al., 2006; Vila and Scheraga, 2009) to identify the existence of local flaws in the sequence rather than only the global quality. Therefore, we added a graphical user interface (GUI) to the *Che*Shift server. As a result, it will be possible to facilitate the validation process by displaying the differences between the observed and computed $^{13}C^{\alpha}$ chemical shifts by using a three-color code mapped onto a 3D protein model. This graphic validation method, far from being only an aesthetic improvement, will enable users of *Che*Shift-2 to detect local flaws in proteins on a per-residue basis fast and accurately without the need for the user to carry out the extensive DFT calculations on which the server is based.

The *Che*Shfit-2 server (Martin et al., 2012) makes use of the following sequential steps: (*i*) for each amino acid residue *i* the average difference between the observed and predicted $^{13}C^{\alpha}$ chemical-shifts, Δ_i, is computed by using Equation (2); (*ii*) the Δ_i value is smoothed by averaging it over the values of the two nearest-neighbor residues ($< \Delta_i >$); (*iii*) the resulting nearest-neighbor averaged value, $< \Delta_i >$, is discretized, i.e., it is assigned an integer value of 1, 0 or -1, depending on the magnitude of $< \Delta_i >$; and (*iv*) these discrete values are mapped onto the 3D protein model and color coded as blue, white and red, respectively. This color-code assignment is based on the assumption that $< \Delta_i >$ values which are within ~1.7 ppm (blue), are considered as small; within ~3.4 ppm (white), as medium; and beyond 3.4 ppm (red), as large. Differences corresponding to blue and white colors are considered acceptable, while red color indicates possible flaws in the structure. In addition, the yellow color was adopted to specify the absence of observed or computed $^{13}C^{\alpha}$ chemical shifts (Martin et al., 2012).

When more than one protein model exists the averaged Δ_i values are computed considering *all* the deposited conformations, although the colored representation is illustrated by using only the first model. This situation is illustrated in Figure 9 for the 20 NMR-determined conformations (see Figure 9a) of *Bacillus Cereus*, a membrane associate protein, PDB id 2K5Q. The large dispersion of conformation in the loops and at the N- and C-termini shown in Figure 9a, rather than being poor representation of the protein, reflects the flexibility of these segments of the molecules in solution, as is clearly shown by the *Che*Shift-2 validation of 2K5Q (see Figure 9b).

4.3.3 Global Versus Local Validation of Proteins

The NMR-determined ensembles of *dynein light chain 2A* protein, PDB id 1TGQ and 2B95, respectively, show different fold, with one of them, namely 1TGQ (now obsolete) having a wrong fold; while the other one, 2B95 (that replaced the obsolete 1TGQ in the PDB), showing a correct fold. This difference is a result of the oligomeric state assumed during the protein-structure determination, namely a monomer for 1TGQ, and a homodimer for 2B95, as pointed out by Nabuurs *et al.* (2006).

Validation of both protein ensembles, as a whole, shows that 2B95 is a slightly better representation of the observed $^{13}C^{\alpha}$ chemical shifts, in terms of the *ca*-rmsd (Vila and Scheraga, 2009), than 1TGQ, viz., *ca*-rmsd = 2.08 and 2.35 ppm, for 2B95 and 1TGQ, respectively. However, the *ca*-rmsd difference between

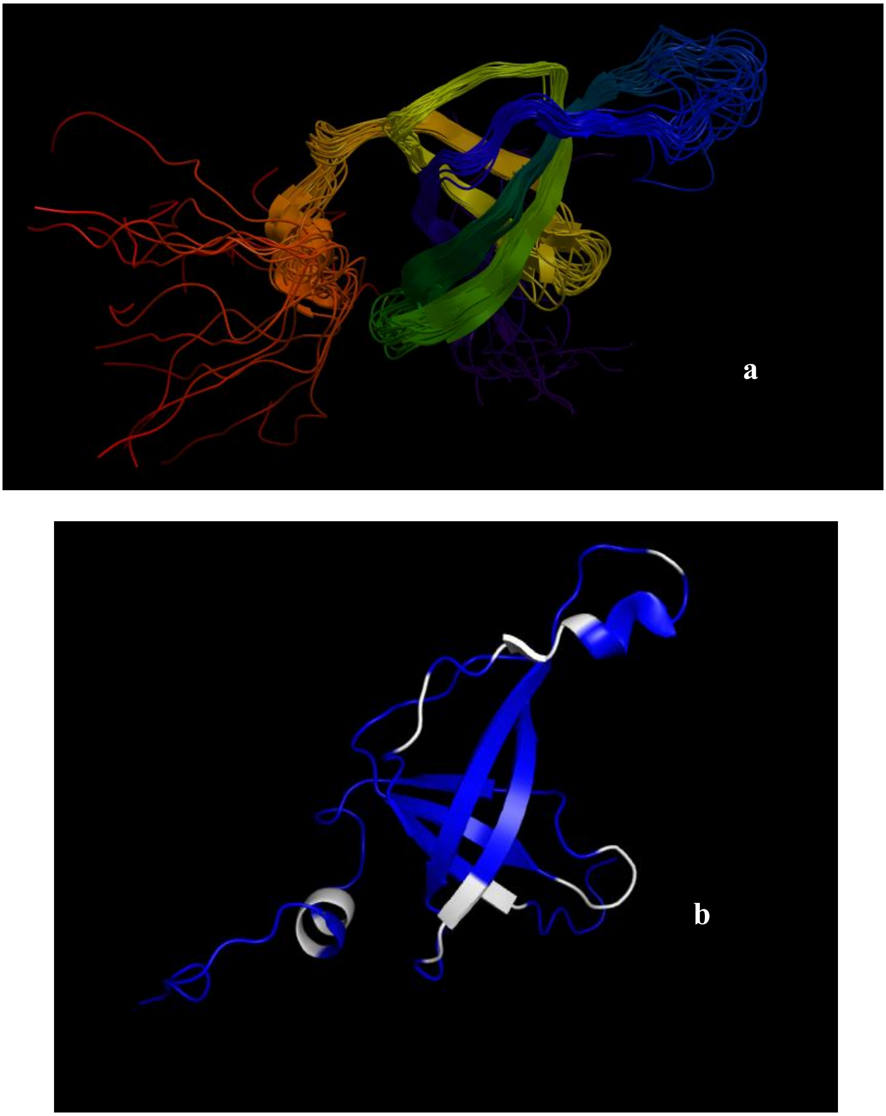

Fig. 9 (**a**) Superposition of 20 NMR-derived conformations of *Bacillus Cereus*, a membrane associate protein, PDB id 2K5Q; (**b**) Protein 2K5Q colored according to *Che*Shift-2. The BMRB accession number, from which the observed $^{13}C^{\alpha}$ chemical shifts were obtained, is 15846.

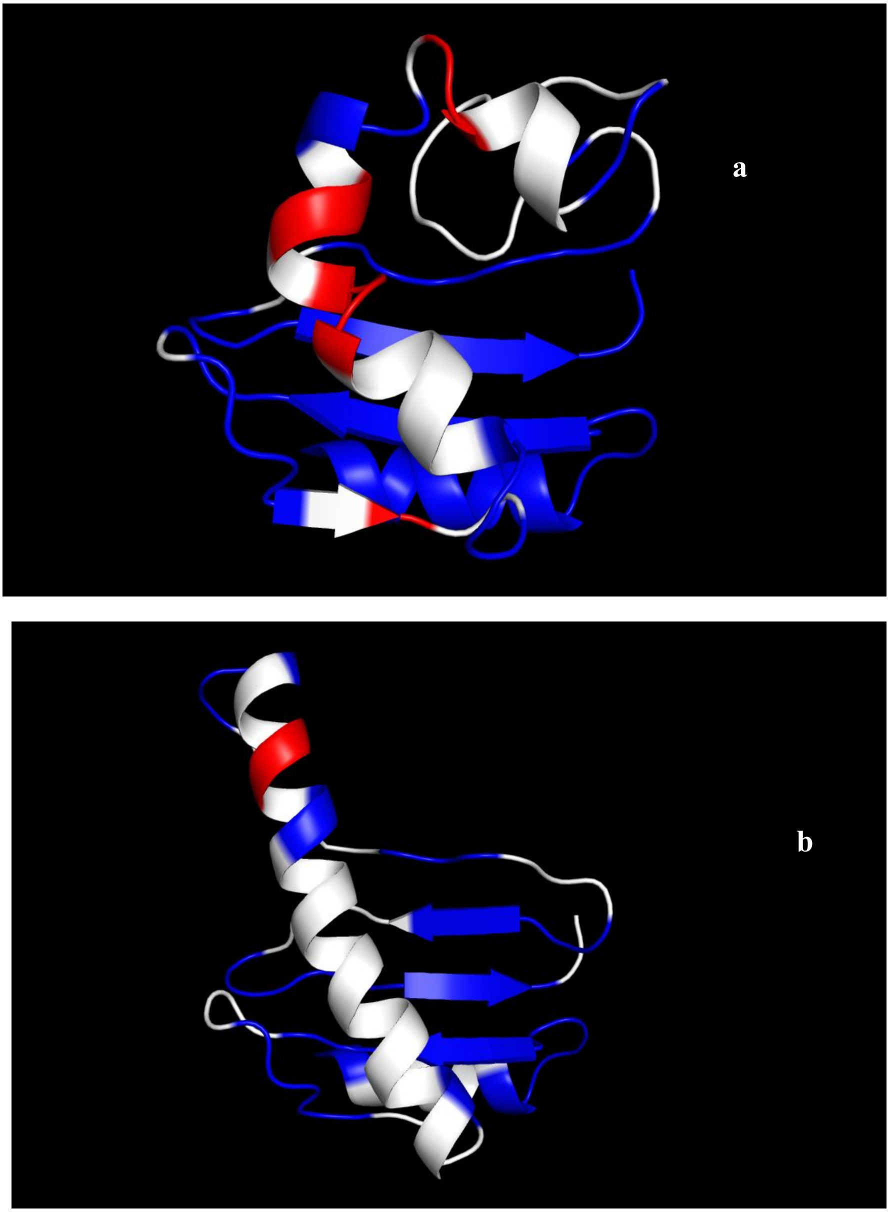

Fig. 10 Two models of the *dynein light chain 2A* protein: (**a**) 1TGQ (obsolete) and (**b**) 2B95 (successor). Both models are shown as ribbons and colored according to *Che*Shift-2. The BMRB accession number, from which the observed $^{13}C^{\alpha}$ chemical shifts were obtained, is 6527. Figure adapted from Martin et al., 2012 (with permission of Oxford University Press).

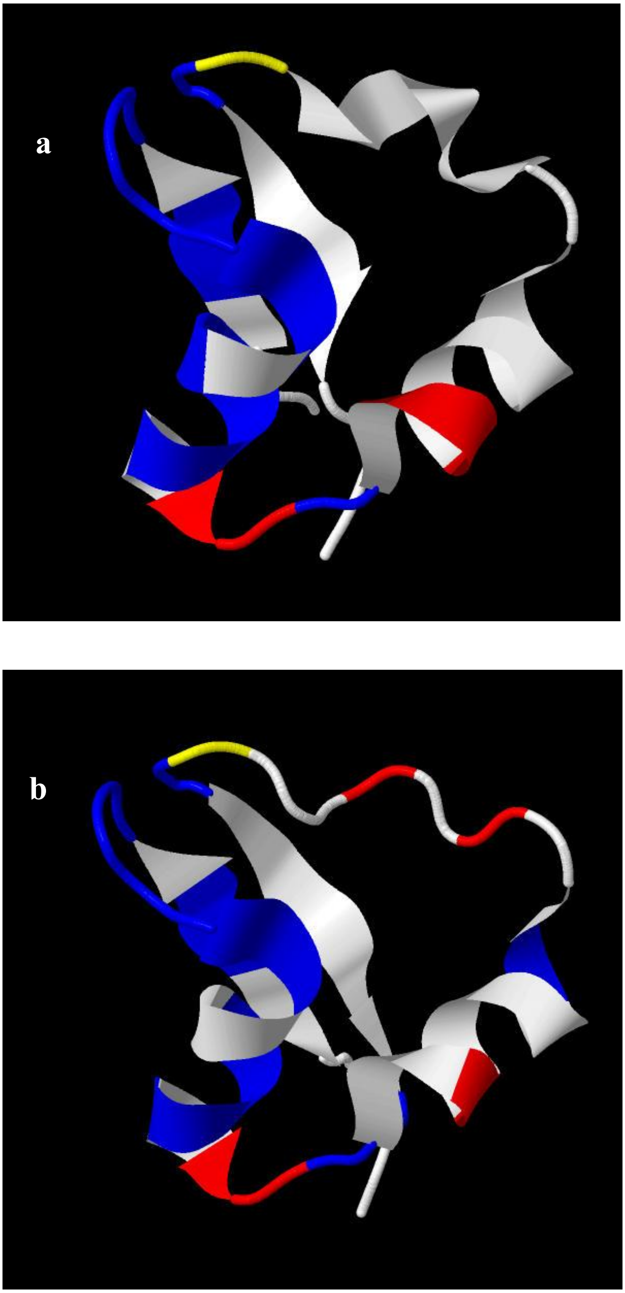

Fig. 11 Two models of *Membrane-bound Lytic Murein Transglycosylase D (fragment Lysm Domain)*: (**a**) PDB id 1E01 (obsolete) and (**b**) 1E0G (successor). The BMRB accession number, from which the observed $^{13}C^{\alpha}$ chemical shifts were obtained, is 4680. Figure adapted from Martin et al., 2012 (with permission of Oxford University Press).

these two ensembles (~0.30 ppm) is not large enough to assure, unambiguously, that the 1TGQ ensemble needs further refinement. In fact, a similar difference in terms of rmsd, i.e., within a range of ~0.30 ppm, was found among 5 new models of the protein ubiquitin (see gray bars in Figure 6), all of which fit X-ray diffraction data with R and R_{free} factors similar to those for the deposited X-ray structure, PDB id1UBQ, solved at 1.8 Å resolution (Arnautova et al., 2009). Certainly, these 5 new models can be considered to be of comparable structural quality. Consequently, variations of *ca*-rmsd ~0.30 ppm cannot be used as a universal criterion to unequivocally determine if a protein, such as 1TGQ, needs further refinement.

Analysis of *dynein light chain 2A* protein illustrates that validation of a protein as a whole (global validation), e.g., with the *ca*-rmsd, may not enable us to determine unambiguously whether one protein model is of better quality than another model of the same protein, while the validation at a per-residue basis (local validation), e.g., as with the *Che*Shift-2 server, does (see Figure 10). To further test the ability of *Che*Shift-2 server to detect small differences between protein models, a small set of 15 obsolete/successor pairs of proteins was also considered (see Supplementary Data of Martin et al., 2012). The results indicate that the *Che*Shift-2 server constitutes a fast and accurate validation tool with which to determine, at the per-residue basis, the existence of local flaws in protein models even for conformations that differ in small details, as for the obsolete and successor models of *Membrane-bound Lytic Murein Transglycosylase D (fragment Lysm Domain)* (see Figure 11).

In general, pairs of obsolete and successor proteins present in PDB can be used as a benchmark set with which to test validation methods. These ensembles of obsolete/successor pairs of proteins are very appealing because their members possess different topology and numbers of residues and a complete sets of ^{13}C$^{\alpha}$ chemical shifts are available for a large number of them from the Bio Magnetic Resonance Data Bank (BMRB) [Ulrich et al., 2008].

5 Conclusions and Future Directions

In this chapter we have illustrated how the information encoded in the ^{13}C chemical shifts can be used for an assorted number of applications, namely, from protein structure prediction to accurate detection of structural flaws, at a residue-level, in NMR-determined protein models.

The ability to detect and accurately characterize the mobility of the surface side chains by computing ^{13}C$^{\alpha}$ chemical shifts constitutes one of the strengths of the current methodology. Hence, we are planning to focus our research on the development of new *physics-based* algorithms for a fast and accurate determination and validation of side-chain conformations, with the goal to improve the quality of NMR-determined protein models. Since NMR spectroscopy provides chemical shifts for several other nuclei, besides ^{13}C$^{\alpha}$, feasibility of their DFT-computation and benefits of including the information encoded in these data in structure determination protocols is currently under investigation in our group. In general, new developments in the field of NMR spectroscopy are needed in order to develop protocols for high-throughput NMR determination of high-quality protein structures in solution.

References

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

Arnautova, Y.A., Jagielska, A., Scheraga, H.A.: A new force field (ECEPP05) for peptides proteins and organic molecules. J. Phys. Chem. B 110, 5025–5044 (2006)

Arnautova, Y.A., Vila, J.A., Martin, O.A., Scheraga, H.A.: What can we learn by computing 13Ca chemical shifts for X-ray protein models? Acta Crystallographica D65, 697–703 (2009)

Bachovchin, W.W.: 15N NMR spectroscopy of hydrogen-bonding interactions in the active site of serine proteases: Evidence for a moving histidine mechanism. Biochemistry 25, 7751–7759 (1986)

Barth, P., Alber, T., Harbury, P.B.: Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. Proc. Natl. Acad. Sci., USA 104, 4898–4903 (2007)

Berjanskii, M., Wishart, D.S.: A simple method to predict protein flexibility using secondary chemical shifts. J. Am. Chem. Soc. 127, 14970–14971 (2005)

Berjanskii, M., Wishart, D.S.: The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. Nucleic Acids Res. 35, W531–W537 (2007)

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Research 28, 235–242 (2000)

Bhattacharya, A., Tejero, R., Montelione, G.T.: Evaluating protein structures determined by structural genomics consortia. Proteins 66, 778–795 (2007)

Billeter, M., Wagner, G., Wüthrich, K.: Solution NMR structure determination of proteins revisited. J. Biomol. NMR 42, 155–158 (2008)

Bradbury, J.H., Scheraga, H.A.: Structural studies of ribonuclease. XXIV. The application of nuclear magnetic resonance spectroscopy to distinguish between the histidine residues of ribonuclease. J. Am. Chem. Soc. 88, 4240–4246 (1966)

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L.: Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D54, 905–921 (1998)

Brünger, A.T.: Version 1.2 of the Crystallography and NMR system. Nature Protocols 2, 2728–2733 (2007)

Case, D.A., Darden, T.A., Cheatham III, T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Merz, K.M., Wang, B., Pearlman, D.A., et al.: AMBER 8. University of California, San Francisco (2004)

Cavalli, A., Salvatella, X., Dobson, C.M., Vendruscolo, M.: Protein structure determination from NMR chemical shifts. Proc. Natl. Acad. Sci., USA 104, 9615–9620 (2007)

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

Cheng, F., Sun, H., Zhang, Y., Mukkamala, D., Oldfield, E.: A solid state 13C NMR, crystallographic, and quantum chemical investigation of chemical shifts and hydrogen bonding in histidine dipeptides. J. Am. Chem. Soc. 127, 12544–12554 (2005)

Chesnut, D.B., Moore, K.D.: Locally dense basis-sets for chemical-shift calculations. J. Comp. Chem. 10, 648–659 (1989)

Chothia, C., Levitt, M., Richardson, D.: Structure of proteins: packing of α-helices and β-sheets. Proc. Natl. Acad. Sci. USA 74, 4130–4134 (1977)

Chou, K.-C., Pottle, M., Némethy, G., Ueda, Y., Scheraga, H.: Structure of β sheets. Origin of the right handed twist and of the increased stability of antiparallel over parallel sheets. J. Mol. Biol. 162, 89–112 (1982)

Chou, K.-C., Scheraga, H.: Origin of the right handed twist of *β* sheets of poly(L Val) chains. Proc. Natl. Acad. Sci. USA 79, 7047–7051 (1982)

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

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

Creighton, T.E.: Proteins: Structure and Molecular Properties, p. 186, 223. W.E. Freeman and Company, New York (1984)

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall III, W.B., Snoeyink, J., Richardson, J.S., Richardson, D.C.: MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 35, W375–W383 (2007)

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

de Dios, A.C., Pearson, J.G., Oldfield, E.: Secondary and tertiary structural effects on protein NMR chemical shifts: An ab initio approach. Science 260, 1491–1496 (1993b)

DePristo, M.A., de Bakker, P.I.W., Blundell, T.L.: Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 12, 831–838 (2004)

Demchuk, E., Wade, R.C.: Improving the continuum dielectric approach to calculating pKas of ionizeable groups in proteins. J. Phys. Chem. 100, 17373–17387 (1996)

Dumbrack Jr., R.L., Karplus, M.: Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. J. Mol. Biol. 230, 543–574 (1993)

Emsley, P., Cowtan, K.: Coot: model-building tools for molecular graphics. Acta Cryst. D60, 2126–2132 (2004)

Farr-Jones, S., Wong, W.Y.L., Gutheil, W.G., Bachovchin, W.W.: Direct observation of the tautomeric forms of histidine in 15N NMR spectra at low temperatures. Comments on intramolecular hydrogen bonding on tautomeric equilibrium. J. Am. Chem. Soc. 115, 6813–6819 (1993)

Frank, A., Onila, I., Moller, H.M., Exner, T.E.: Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. Proteins 79, 2189–2202 (2011)

Frank, A., Möller, H.M., Exner, T.H.: Toward the quantum chemical calculation of NMR chemical shifts of proteins.2. Level of theory, basis set, and solvent model dependence. J. Chem. Theory Comput. 8, 1480–1492 (2012)

Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, J.A., Stratmann Jr., R.E., Burant, J.C., et al.: Gaussian 03, Revision E.01. Gaussian, Inc., Wallingford (2004)

Furnham, N., Blundell, T.L., DePristo, M.A., Terwilliger, T.C.: Is one solution good enough? Nature Struct. Mol. Biol. 13, 184–185 (2006)

Guerry, P., Herrmann, T.: Advances in automated NMR protein structure determination. Q. Rev. Biophys. 44, 257–309 (2011)

Güntert, P.: Structure calculation of biological macromolecules from NMR data. Q. Rev. Biophys. 31, 145–237 (1998)

Güntert, P.: Automated structure determination from NMR spectra. Eur. Biophys. J. 38, 129–143 (2009)

Güntert, P., Braun, W., Wüthrich, K.: Efficient computation of threedimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J. Mol. Biol. 217, 517–530 (1991)

Harbison, G., Herzfeld, J., Griffin, R.G.J.: Nitrogen-15 chemical shifts tensors in L-histidine hydrochloride monohydrate. J. Am. Chem. Soc. 103, 4752–4754 (1981)

Hass, M.A.S., Hansen, D.F., Christensen, H.E.M., Led, J.J., Kay, L.E.: Characterization of conformational exchange of a histidine side chain: protonation, rotamerization, and tautomerization of His61 plastocyanin from Anabaena variabilis. J. Am. Chem. Soc. 130, 8460–8470 (2008)

Hass, M.A.S., Yilmaz, A., Christensen, H.E.M., Led, J.J.: Histidine side-chain dynamics and protonation monitored by 13C CPMG NMR relaxation dispersion. J. Biomol. NMR 44, 225–233 (2009)

Havlin, R.H., Le, H., Laws, D.D., de Dios, A.C., Oldfield, E.: An ab initio quantum chemical investigation of carbon–13 NMR shielding tensors in glycine, alanine, valine, isoleucine, serine, and threonine: Comparisons between helical and sheet tensors, and effects of $\chi 1$ on shielding. J. Am. Chem. Soc. 119, 11951–11958 (1997)

Hu, F., Wenbin, L., Hong, M.: Mechanism of proton conduction and gating in influenza M2 proton channels from solid-state. NMR Science 330, 505–508 (2010)

Huang, Y.J., Powers, R., Montelione, G.T.: 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 (2005)

Huang, Y.J., Tejero, R., Powers, R., Montelione, G.T.: A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins 62, 587–603 (2006)

Höfinger, S., Almeida, B., Hansmann, U.H.E.: Parallel tempering molecular dynamics folding simulation of a signal peptide in explicit water. Proteins 68, 662–669 (2007)

Iwadate, M., Asakura, T., Williamson, M.P.: Cα and Cβ carbon-13 chemical shifts in proteins from an empirical database. J. Biomol. NMR 13, 199–211 (1999)

Jameson, A.K., Jameson, C.J.: Gas-phase 13C chemical shifts in the zero-pressure limit: Refinements to the absolute shielding scale for 13C. J. Chem. Phys. Lett. 134, 461–466 (1997)

Jang, S., Kim, E., Pak, Y.: Free energy surfaces of miniproteins with a beta beta alpha motif: Replica exchange molecular dynamics simulation with an implicit solvation model. Proteins 62, 663–671 (2006)

Jensen, M.R., Has, M.A.S., Hansen, D.F., Led, J.J.: Investigating metal-binding in proteins by nuclear magnetic resonance. Cell Mol. Life Sci. 64, 1085–1104 (2007)

Karplus, M.: Contact Electron-Spin Coupling of Nuclear Magnetic Moments. J. Chem. Phys. 30, 11–15 (1959)

Kawai, Y., Moriya, S., Ogasawara, N.: Identification of a protein YneA, responsible for cell division suppression during the SOS response in Bacillus subtilis. Mol. Microbiol. 47, 1113–1122 (2003)

Kleywegt, G.J.: On vital aid: the why, what and how of validation. Acta Cryst. D65, 134–139 (2009)

Korzhnev, D.M., Orekhov, V.Y., Arseniev, A.S.: Model-free approach beyond the borders of its applicability. J. Mag. Res. 127, 184–191 (1997)

Kruskal Jr., J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. American Math. Soc. 7, 48–50 (1956)

Kuszewski, J., Qin, J., Gronenborn, A.M., Clore, M.: The impact of direct refinement against 13Ca and 13Cb chemical shifts on protein structure determination by NMR. J. Magn. Reson. Ser. B 106, 92–96 (1995)

Kuzin, A.P., Su, M., Seetharaman, J., Janjua, H., Cunningham, K., Maglaqui, M., Owens, L.A., Zhao, L., Xiao, R., Baran, M.C., Acton, T.B., Rost, B., Montelione, G.T., Hunt, J.F., Tong, L.: Crystal structure of UPF0291 protein ynzC from Bacillus subtilis at resolution 2.0 A. Northeast Structural Genomics Consortium target SR384 (2008), doi:10.2210/pdb3bhp/pdb

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.: PROCHECK - a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26, 283–291 (1993)

Li, Z., Scheraga, H.A.: Monte Carlo minimization approach to the multiple minima problem in protein folding. Proc. Natl. Acad. Sci. USA 84, 6611–6615 (1987)

Li, Z., Scheraga, H.A.: Structure and free energy of complex thermodynamic systems. J. Molec. Str. (Theochem) 179, 333–352 (1998)

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

Lovell, S.C., Davis, I.W., Arendall III, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C.: Structure validation by Cα geometry: φ, ψ, and Cβ deviation. Proteins 50, 437–450 (2003)

Luginbühl, P., Szyperski, T., Wüthrich, K.: Statistical basis for the use of 13Cα chemical shift in protein structure determination. J. Magn. Reson. 109, 229–233 (1995)

Lüthy, R., Bowie, J.U., Eisenberg, D.: Assessment of protein models with three-dimensional profiles. Nature 356, 83–85 (1992)

Mandel, M.: Proton Magnetic resonance spectra of some proteins: I. Ribonuclease, oxidized ribonuclease, lysozyme, and cytochrome c. J. Biol. Chem. 240, 1586–1592 (1965)

Markley, J.L.: Observation of histidine residues in proteins by means of nuclear magnetic resonance spectroscopy. Acc. Chem. Res. 8, 70–80 (1974)

Martin, O.A., Vila, J.A., Scheraga, H.A.: CheShift-2: graphic validation of protein structures. Bioinformatics 28, 1538–1539 (2012)

Martin, O.A., Villegas, M.E., Vila, J.A., Scheraga, H.A.: Analysis of 13Cα and 13Cβ chemical shifts of cysteine and cystine residues in proteins: A quantum chemical approach. J. Biomol. NMR 46, 217–225 (2010)

Meadows, D.H., Jardetzky, O., Epand, R.M., Ruterjans, H.H., Scheraga, H.A.: Proc. Natl. Acad. Sci. USA 60, 766–772 (1968)

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

Mohanty, S., Hansmann, U.H.E.: Folding of proteins with diverse folds. Biophy. J. 91, 3573–3578 (2006)

Nabuurs, S.B., Spronk, C.A.E.M., Vuister, G.W., Vriend, G.: Tradional biomolecular structure determination by NMR spectroscopy allows for major errors. PLOS Comp. Biol. 2, 71–79 (2006)

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

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

Palmer III, A.G.: NMR characterization of the dynamics of biomacromolecules. Chem. Rev. 104, 3623–3640 (2004)

Parr, R.G., Yang, W.: Density Functional Theory of Atoms and Molecules. Oxford University Press, New York (1989)

Pearson, J.G., Le, H., Sanders, L.K., Godbout, N., Havlin, R.H., Oldfield, E.J.: Predicting chemical shifts in proteins: Structure refinement of valine residues by using ab initio and empirical geometry optimizations. J. Am. Chem. Soc. 119, 11951–11958 (1997)

Pelton, J.G., Torchia, D.A., Meadow, N.D., Roseman, S.: Tautomeric states of the active-site histidine of phosphorylated and unphosphorylated IIIGlc, a signal-transducing protein from Escherichia coli, using two-dimensional heternoculear NMR techniques. Prot. Sci. 2, 543–558 (1993)

Quirt, A.R., Lyerla Jr., J.R., Peat, I.R., Cohen, J.S., Reynolds, W.F., Freedman, M.H.: Carbon-13 nuclear magnetic resonance titration shifts in amino acids. J. Am. Chem. Soc. 96, 570–574 (1974)

Rabenstein, D.L., Sayer, T.L.: Carbon-13 shifts parameters for amines, carboxylic acids and amino acids. J. Magn. Res. 24, 27–39 (1976)

Reynolds, W.F., Peat, I.R., Freedman, M.H., Lyerla Jr., J.R.: Determination of the tautomeric form of the imidazole ring of L-Histidine in basic solution by carbon-13 magnetic resonance spectroscopy. J. Am. Chem. Soc. 95, 328–331 (1973)

Ringe, D., Petsko, G.A.: Study of protein dynamics by X-ray diffraction. Methods in Emzymology 131, 389–433 (1986)

Ripoll, D.R., Vorobjev, Y.N., Liwo, A., Vila, J.A., Scheraga, H.A.: Coupling between folding and ionization equilibria: Effects of pH on the conformational preferences of polypeptides. J. Mol. Biol. 264, 770–783 (1996)

Ripoll, D.R., Ni, F.: Refinement of the thrombin-bound structure of a hirudin peptide by a restrained Electrostatically Driven Monte-Carlo Method. Biopolymers 32, 359–365 (1992)

Rosato, A., Aramini, J.M., Arrowsmith, C., Bagaria, A., Baker, D., Cavalli, A., Doreleijers, J.F., Eletsky, A., Giachetti, A., Guerry, P., et al.: Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20, 227–236 (2012)

Rosato, A., Bagaria, A., Baker, D., Bardiaux, B., Cavalli, A., Doreleijers, J.F., Giachetti, A., Guerry, P., Guntert, P., Herrmann, T., et al.: CASDNMR: critical assessment of automated structure determination by NMR. Nat. Methods 6, 625–626 (2009)

Santiveri, C.M., Santoro, J., Rico, M., Jiménez, M.A.: Factors involved in the stability of isolated beta-sheets: turn sequence, beta-sheet twisting, and hydrophobic surface burial. Prot. Sci. 13, 1134–1147 (2004)

Sayer, T.L., Rabenstein, D.L.: Nuclear magnetic resonance studies of the acid-base chemistry of amino acids and peptides. III. Determination of the microscopic and macroscopic acid dissociation constants of α,ω-diaminocarboxylic acids. Can. J. Chem. 54, 3392–3400 (1976)

Schuster, I.I., Roberts, J.D.: Nitrogen-15 nuclear magnetic resonance spectroscopy. Effects of hydrogen bonding and protonation on nitrogen chemical shifts in imidazoles. J. Org. Chem. 44, 3864–3867 (1979)

Schwarzinger, S., Kroon, G.J.A., Foss, T.R., Chung, J., Wright, P.E., Dyson, H.J.: Sequence-dependent correction of random coil NMR chemical shifts. J. Am. Chem. Soc. 123, 2970–2978 (2001)

Serrano, P., Johnson, M.A., Chatterjee, A., Neuman, B., Joseph, J.S., Buchmeier, M.J., Kuhn, P., Wüthrich, K.: NMR structure of the nucleic acid-binding domain of the SARS coronavirus nonstructural protein 3. J. Virol. 83, 12998–13008 (2009)

Sevcik, J., Dauter, Z., Lamzin, V.S., Wilson, K.S.: Ribonuclease from streptomyces aureofaciens at atomic resolution. Acta Cryst. D52, 327–344 (1996)

Shen, Y., Bax, A.: Protein backbone chemical shifts predicted from searching a database for torsional angle and sequence homology. J. Biomol. NMR 38, 289–302 (2007)

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., et al.: Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. USA 105, 4685–4690 (2008)

Shimba, N., Serber, Z., Lewidge, R., Miller, S.M., Craik, C.S., Dotsch, V.: Quantitative identification of the protonation state of histidine in vitro and in vivo. Biochem. 42, 9227–9234 (2003)

Shimba, N., Takahashi, H., Sakakura, M., Fuji, I., Shimada, I.: Determination of protonation and deprotonation forms and tautomeric states of histidine residues in large proteins using nitrogen-carbon J couplings in imidazole ring. J. Am. Chem. Soc. 120, 10988–10989 (1998)

Sippl, M.J.: Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355–362 (1993)

Sitkoff, D., Sharp, K.A., Honig, B.: Accurate calculation of hydration free energies using macroscopic solvent models. J. Phys. Chem. 98, 1978–1988 (1994)

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

Steiner, T.: L-Histidyl-L-alanine dehydrate. Acta Cryst. C 52, 2554–2556 (1996)

Steiner, T., Koellner, G.: Coexistence of both histidines tautomers in the solid state and stabilization of the unfavorable Nd-H form by intramolecular hydrogen bonding: rystalline L-His-Gly hemihydrates. Chem. Commun. 13, 1207–1208 (1997)

Strohmeier, M., Stueber, D., Grant, D.M.: Accurate 13C and 15N chemical shift and 14N quadrupolar coupling constant calculations in amino acid crystals: Zwitterionic, hydrogen-bonded systems. J. Phys. Chem. A 107, 7629–7642 (2003)

Sudmeier, J.L., Bradshaw, E.M., Coffman Haddad, K.E., Day, R.M., Thalhauser, C.J., Bullock, P.A., Bachovchin, W.W.: Identification of histidine tautomers in proteins by 2D 1H/13Cδ2 one-bond correlated NMR. J. Am. Chem. Soc. 125, 8430–8431 (2003)

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

Surprenant, H.L., Sarneski, J.E., Key, R.R., Byrd, J.T., Reilley, C.N.: Carbon-13 studies of amino acids: chemical shifts, protonation shifts, microscopic protonation behavior. J. Magn. Res. 40, 231–243 (1980)

Terwilliger, T.C., Berendzen, J.: Automated MAD and MIR structure solution. Acta Cryst. D55, 849–861 (1999)

Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C.F., Tolmie, D.E., Wenger, R.K., Yao, H., Markley, J.L.: BioMagResBank. Nucleic Acids Res. 36, D402–D408 (2008)

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

Vila, J., Williams, R.L., Vásquez, M., Scheraga, H.A.: Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. Proteins: Structure, Function, and Genetics 10, 199–218 (1991)

Vila, J.A., Arnautova, Y.A., Martin, O.A., Scheraga, H.A.: Quantum-Mechanics-Derived 13Cα Chemical Shift Server (CheShift) for Protein Structure Validation. Proc. Natl. Acad. Sci. USA 106, 16972–16977 (2009)

Vila, J.A., Arnautova, Y.A., Scheraga, H.A.: Use of 13Cα chemical shifts for accurate determination of β-sheet structures in solution. Proc. Natl. Acad. Sci. USA 105, 1891–1896 (2008a)

Vila, J.A., Baldoni, H.A., Scheraga, H.A.: performance of density functional models to reproduce observed 13C*α* chemical shifts of proteins in solution. J. Comp. Chem. 38, 884–892 (2008b)

Vila, J.A., Aramini, J.M., Rossi, P., Kuzin, A., Su, M., Seetharaman, J., Xiao, R., Tong, L., Montelione, G.T., Scheraga, H.A.: Quantum Chemical 13Cα Chemical Shift Calculations for Protein NMR Structure Determination, Refinement, and Validation. Proc. Natl. Acad. Sci. USA 105, 14389–14394 (2008c)

Vila, J.A., Arnautova, Y.A., Vorobjev, Y., Scheraga, H.A.: Assessing the fractions of tautomeric forms of the imidazole ring of histidine in proteins as a function of pH. Proc. Natl. Acad. Sci. USA 108, 5602–5607 (2011)

Vila, J.A., Baldoni, H.A., Ripoll, D.R., Ghosh, A., Scheraga, H.A.: Polyproline II helix conformation in a proline-rich enviroment: A theoretical Study. Biophysical Journal 86, 731–742 (2004)

Vila, J.A., Baldoni, H.A., Ripoll, D.R., Scheraga, H.A.: Unblocked Statistical-Coil Tetrapeptides in Aqueous Solution: Quantum-Chemical Computation of the Carbon-13 NMR Chemical Shifts. Journal of Biomolecular NMR 26, 113–130 (2003)

Vila, J.A., Ripoll, D.R., Arnaturova, Y.A., Vorobjev, Y.N., Scheraga, H.A.: Coupling between conformation and proton binding in proteins. Proteins 61, 56–68 (2005)

Vila, J.A., Villegas, M.E., Baldoni, H.A., Scheraga, H.A.: Predicting 13Cα chemical shifts for validation of protein structures. J. Biomol. NMR 38, 221–235 (2007a)

Vila, J.A., Ripoll, D.R., Scheraga, H.A.: Use of 13Cα chemical shifts in protein structure determination. J. Phys. Chem. B 111, 6577–6585 (2007b)

Vila, J.A., Scheraga, H.A.: Factors affecting the use of 13Cα chemical shifts to determine, refine, and validate protein structures. Proteins: Structure, Function, and Bioinformatics 71, 641–654 (2008)

Vila, J.A., Scheraga, H.A.: Assessing the accuracy of protein structures by quantum mechanical computations of 13Cα chemical shifts. Acc. Chem. Res. 42, 1545–1553 (2009)

Vila, J.A., Serrano, P., Wüthrich, K., Scheraga, H.A.: Sequential nearest-neighbor effects on computed 13Cα chemical shifts. Journal of Biomolecular NMR 48, 23–30 (2010)

Villegas, M.E., Vila, J.A., Scheraga, H.A.: Effects of side-chain orientation on the 13C chemical shifts of antiparallel β-sheet model peptides. J. Biomol. NMR 37, 137–146 (2007)

Vorobjev, Y.N., Scheraga, H.A.: A Fast Adaptive Multigrid Boundary Element Method for macromolecule electrostatic computations in solvent. J. Comp. Chem. 18, 569–583 (1997)

Vorobjev, Y.N., Vila, J.A., Scheraga, H.A.: FAMBE-pH: a Fast and Accurate Method to Compute the Total Solvation Free Energies of Proteins. Journal of Physical Chemistry B 112, 11122–11136 (2008)

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

Vásquez, M., Scheraga, H.A.: Variable-Target-Function and buildup procedures for the calculation of protein conformation - application to bovine pancreatic trypsin-inhibitor using limited simulated Nuclear Magnetic-Resonance data. J. Biomol. Struct. Dyn. 5, 757–784 (1988)

Wang, Y., Jardetzky, O.: Investigation of the neighboring residue effects on protein chemical shifts. J. Am. Chem. Soc. 12, 14075–14084 (2002a)

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

Williamson, M.P., Craven, C.J.: Automated protein structure calculation from NMR data. J. Biomol. NMR 43, 131–143 (2009)

Williamson, M.P., Kikuchi, J., Asajura, T.: Application of 1H-NMR chemical-shifts to measure the quality of protein structures. J. Mol. Biol. 247, 541–546 (1995)

Wishart, D., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H., Oldfield, E., Markley, J., Sykes, B.: 1H, 13C and 15N chemical shift referencing in biomolecular NMR. J. Biomol. NMR 6, 135–140 (1995a)

Wishart, D., Bigam, C.G., Holm, A., Hodges, R.S., Sykes, B.D.: 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigation of nearest-neigbor effects. J. Biomol. NMR 5, 67–81 (1995b)

Wüthrich, K.: NMR in Biological Research: Peptides and Proteins. North-Holland, Amsterdam (1976)

Wüthrich, K.: NMR of Proteins and Nucleic Acids. Wiley, New York (1986)

Xu, X.-P., Case, D.A.: Probing multiple effects on 15N, 13Cα, 13Cβ and 13C′ chemical shifts in peptides using density functional theory. Biopolymers 65, 408–423 (2002)

Xu, X.-P., Case, D.A.: Automated prediction of 15N, 13Cα, 13Cβ and 13C′ chemical shifts in proteins using a density functional database. J. Biomol. NMR 21, 321–333 (2001)

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

Zhou, R.: Free energy landscape of protein folding in water: Explicit vs. implicit solvent. Proteins 53, 148–161 (2003)

Zhou, Y., Vitkup, D., Karplus, M.: Native proteins are surface-molten solids: Application of the Lindemann criterion for the solid versus liquid state. J. Mol. Biol. 285, 1371–1375 (1999)