ESCASA: Analytical Estimation of Atomic Coordinates from Coarse-Grained Geometry for NMR-Assisted Protein Structure Modeling. I. Backbone and H^{β} protons.

Emilia A. Lubecka*, Adam Liwo†

May 13, 2021

Abstract

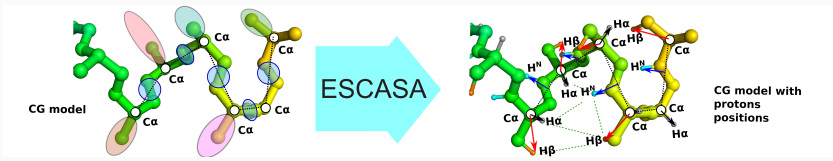
A method for the estimation of coordinates of atoms in proteins from coarse-grained geometry by simple analytical formulas (ESCASA), for use in nuclear-magnetic-resonance (NMR) data-assisted coarse-grained simulations of proteins is proposed. In this paper, the formulas for the backbone H^{α} and amide (H^{N}) protons, and the side-chain H^{β} protons, given the C^{α}-trace, have been derived and parameterized, by using the interproton distances calculated from a set of 140 high-resolution non-homologous protein structures. The mean standard deviation over all types of proton pairs in the set was 0.44 Å after fitting. Validation against a set of 41 proteins with NMR-determined structures, which were not considered in parameterization, resulted in average standard deviation from average proton-proton distances of the NMR-determined structures of 0.25 Å, compared to 0.21 Å obtained with the PULCHRA all-atom-chain reconstruction

*Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, G. Narutowicza 11/12, 80-233 Gdańsk, Poland
†Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland

algorithm and to the 0.12 Å standard deviation of the average-structure proton-proton distance of NMR-determined ensembles. The formulas provide analytical forces and can, therefore, be used in coarse-grained molecular dynamics.

Keywords: coarse graining, data-assisted molecular modeling, nuclear magnetic resonance, proteins

■



Analytical expressions for approximate positions of the backbone H^{α} and amide protons and of the side-chain H^{β}-protons in terms of α-carbon-trace coordinates were proposed and parameterized for use in coarse-grained NMR-data-assisted protein-structure modeling. The expressions provide analytical forces and are thus suitable for coarse-grained molecular dynamics.

INTRODUCTION

Coarse-grained (CG) models are nowadays increasingly used in the modeling of protein structure and dynamics[1–7], because they provide the extension of time- and size-scale of simulations by several orders of magnitude, compared to all-atom models. This feature results, in turn, from averaging out the secondary degrees of freedom, which are not included in the model[8,9]. Due to the inevitable inaccuracy of the force fields, the coarse-grained models are often used in bioinformatics- or data-assisted mode[5]. Examples of the first one include incorporating predicted contacts[10,11], distance distribution or template geometries[12,13] extracted from the Protein Data Bank (PDB)[14]. The experimental information comes from nuclear magnetic resonance (NMR)[15,16], small angle X-ray scattering (SAXS) or small-angle neutron scattering (SANS) spectroscopy[17], chemical cross-link mass-spectroscopy (XLMS)[18–20] or fluorescence resonance energy transfer (FRET) measurements[21,22].

NMR has been used since 1980 for protein-structure determination[23]. It is one of the principal techniques used to determine 3D structures of biomolecules at the atomic precision and to analyze their dynamics in solution under near-physiological conditions. In contrast to single-crystal diffraction, it does not require protein crystallization, which still presents a major bottleneck. The NMR spectroscopy usually provides the information of proton-proton distances, which largely define the spatial structure of a protein, as well as chemical shifts and coupling constants that can be used to determine the local structure. The distances between the carbon and nitrogen atoms can also be estimated if the sample is enriched in carbon-13 and nitrogen-15. The distance and local-structure information is converted into the distance- and dihedral-angle restraints that are added to the energy function in simulations[24].

Use of NMR-derived restraints in all-atom simulations is straightforward. However, only few coarse-grained models keep some of the atomic details. For example, protein backbone is represented at the all-atom level in the Rosetta[25] and AWSEM[26] coarse-grained models of proteins. However, most of coarse-grained protein models[5] do not keep explicit proton positions, which are necessary for direct implementation of NMR-generated distance restraints. Therefore, at present, the main approach to use restraints from NMR-data with fully coarse-grained simulations is the reconstruction of all-atom structures from the CG representation

and then evaluation of the restraints; this approach is used with the Cα-Cβ-Side group (CABS) statistical force field[27,28], which is used together with a dynamic Monte Carlo conformational search. However, reconstruction of all-atom chains generates additional cost and requires interrupting simulations from time to time. Additionally, it is difficult to use with molecular dynamics, because it would be difficult to generate the forces due to restraints. A plausible alternative is expressing the coordinates of hydrogens and other atoms active in NMR directly from coarse-grained geometry as continuous functions of coarse-grained coordinates, i.e., without having to go through the laborious process of reconstruction of the whole all-atom chain. In our earlier work[29], we used a "naïve" NMR-data-assisted approach with the UNited RESidue (UNRES) force field developed in our laboratory[30–32], in which we added 2 Å to the distances involving C^{α} atoms or the sidechain pseudoatoms to obtain an estimate of the corresponding proton-proton distance. This simple approach had some success in the 13th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP13), improving noticeably the models obtained with plain UNRES and with contact-assisted UNRES[29].

In this paper, we propose an analytical approach to the estimation of coordinates of atoms in proteins from coarse-grained geometry. We apply the method to the calculation of the H^{α}, the amide (H^{N}) and the H^{β} proton coordinates given the C^{α}-trace geometry. Because the derived formulas can be used to calculate the analytical forces due to NMR-penalty terms, this method enables us to incorporate NMR-derived distance restraints directly in coarse-grained molecular dynamics simulations. The approach is applicable to any coarse-grained models that keep the α-carbon positions, which is the case of most of the coarse-grained protein models[5], including UNRES[31].

METHODOLOGY

Analytical formulas for approximate atomic positions

Given the C^{α}-trace, the protein backbone can be reconstructed to a very good accuracy by using the knowledge-based algorithms such as, e.g., BBQ[33] and PULCHRA[34]. This feature

results from the fact that the low-energy regions constitute a small part of the Ramachandran maps and, consequently, there is little room for the peptide planes to accommodate given a C^α-trace geometry[35]. In our previous work[35] we derived the analytical formulas for the distances between the atoms of two polymer units, given the orientation of their virtual-bond axes, and the angles λ for rotation about these axes. The λ angles are Boltzmann-averaged but, because of the Ramachandran-map restrictions, the averaging is confined to narrow regions. Therefore, the formulas for approximate positions of the backbone (H^α and H^N) and of the H^β protons and, consequently, for the respective interproton distances, can be derived by using the theory described in Ref. 35. The treatment could also be extended to side-chain protons, as well as to the non-hydrogen atoms; we leave these extensions to our future work. We coin the name ESCASA to the new approach (EStimation of Coordinates of Atoms in proteins from coarse-grained geometry by Simple Analytical formulas). The advantage of our approach is that the forces due to NMR penalty term are expressed analytically, thus enabling us to carry out NMR-data-assisted molecular dynamics simulations.

H^α protons. We define the local coordinate system centered at residue i with origin at C_i^α as the Frenet frame of this residue[36], the x being the normal, the the y axis the tangent, and the z axis the binormal (Figure 1). For the Reader's convenience the expressions for the unit vectors of these axes are given by eqs. (S1 – S3) of the Supplementary Material. Defining $\alpha_i^{H^\alpha}$ as the zenith angle of the H^α proton of the ith residue, the coordinates of this proton in the local coordinate systems are expressed by eqs. (1 – 3). The azimuth angle $\beta_i^{H^\alpha}$ is assumed to be -90°; fitting this angle did not result in any remarkable improvement of the predicted distances. Because the glycine residue contains two H^α protons, which are indistinguishable in NMR measurements, in this case eqs. (1 – 3) express the average coordinates of the protons, corresponding to the Q^α pseudo-atom.

$$x_i^{H^\alpha} = d_i^{H^\alpha C^\alpha} \cos \alpha_i^{H^\alpha} \quad (1)$$

$$y_i^{H^\alpha} = 0 \quad (2)$$

$$z_i^{H^\alpha} = -d_i^{H^\alpha C^\alpha} \sin \alpha_i^{H^\alpha} \quad (3)$$

where the virtual-bond lengths, $d_i^{\mathrm{H}^\alpha C^\alpha}$, are adjustable parameters dependent on residue type (glycine, proline, and other). The cosine of the zenith angle is expressed as a 6-order power series in $\cos\theta_i$, in which $\cos\alpha_{i0}^{H^\alpha}$ and $a_{1i}^{H^\alpha} - a_{6i}^{H^\alpha}$ are adjustable parameters [eq. (4)].

$$\cos\alpha_i^{H^\alpha C^\alpha} = \cos\alpha_{0i}^{H^\alpha} + \sum_{k=1}^{6} a_{ki}^{H^\alpha}(\cos\theta_i)^k, \quad 0 \le \alpha_i^{H^\alpha} \le 180^\circ \tag{4}$$

The parameters of eqs. (1 – 4) depend on residue type. Because the local-interaction pattern depends mostly on the immediate neighborhood of the backbone, three residue types are defined, namely glycine, proline, and other. These residue types are identical to the residue types with regard to local interactions defined in the UNRES model of polypeptide chains[30,31,35].

H$^\beta$ protons. As shown in Figure 1, both the zenith ($\alpha_i^{\mathrm{H}^\beta}$) and the azimuth ($\beta_i^{\mathrm{H}^\beta}$) angles are used to define the approximate position of a H$^\beta$ proton. These angles are defined in the Figure and in its legend. As for the H$^\alpha$ protons, the positions of H$^\beta$ should be considered as average positions if there are more than one H$^\beta$ atom attached to C$^\beta$. The coordinates are expressed in the local coordinate system shown in Figure 1.

$$x_i^{H^\beta} = d_i^{H^\beta C^\alpha}\cos\alpha_i^{H^\beta} \tag{5}$$

$$y_i^{H^\beta} = d_i^{H^\beta C^\alpha}\sin\alpha_i^{H^\beta}\cos\beta_i^{H^\beta} \tag{6}$$

$$z_i^{H^\beta} = d_i^{H^\beta C^\alpha}\sin\alpha_i^{H^\beta}\sin\beta_i^{H^\beta} \tag{7}$$

where the $d_i^{H^\beta C^\alpha}$ bond length is an adjustable parameter, while the cosine of $\alpha_i^{\mathrm{H}^\beta}$, and the cosine and the sine of $\beta_i^{\mathrm{H}^\beta}$ depend on the virtual-bond-angle θ_i and are expressed by eqs. (8) – (12), respectively.

$$\cos\alpha_i^{H^\beta} = \cos\alpha_{0i}^{H^\beta} + \sum_{k=1}^{6} a_{ki}^{H^\beta}(\cos\theta_i)^k, \quad 0 \le \alpha_i^{H^\beta} \le 180^\circ \tag{8}$$

$$\cos\beta_i^{H^\beta} = \frac{C_i^{H^\beta}}{\sqrt{(C_i^{H^\beta})^2 + (S_i^{H^\beta})^2}} \tag{9}$$

$$\sin\beta_i^{H^\beta} = \frac{S_i^{H^\beta}}{\sqrt{(C_i^{H^\beta})^2 + (S_i^{H^\beta})^2}} \tag{10}$$

$$C_i^{H^\beta} = c_{0i}^{H^\beta} + \sum_{k=1}^{6} c_{ki}^{H^\beta}(\cos\theta_i)^k \tag{11}$$

$$S_i^{H^\beta} = s_{0i}^{H^\beta} + \sum_{k=1}^{6} s_{ki}^{H^\beta}(\cos\theta_i)^k \tag{12}$$

The parameters of eqs. (5 – 12) do not depend on residue type (proline or other, because glycine does not have H^β protons).

H^N protons. In our earlier work[35,37], we found that the optimal orientation of a peptide plane in four-C^α-atom frame can be obtained as the vector sum of two fictitious dipoles, one ($\boldsymbol{\mu}_{2,i}$) determined by the geometry of the $C^\alpha_{i-1}\cdots C^\alpha_i\cdots C^\alpha_{i+1}$ and the other one ($\boldsymbol{\mu}_{1,i+1}$) by that of the the $C^\alpha_i\cdots C^\alpha_{i+1}\cdots C^\alpha_{i+2}$ frame, as illustrated in Figure 2, in which also the local-coordinate systems of the two frames are also shown. The unit vectors of the axes of the coordinate systems are expressed by eqs. (4 –9) of the Supplementary Material. The vectors $\boldsymbol{\mu}_{2,i}$ and $\boldsymbol{\mu}_{1,i+1}$ are defined in their respective coordinate systems by eqs. (13) and (14), respectively.

$$\boldsymbol{\mu}_{2,i} = \begin{pmatrix} 0 \\ \sin\theta_i[a_{21i}^y + a_{22i}^y\cos\theta_i + a_{23i}^y(\cos\theta_i)^2] \\ \sin\theta_i[a_{21i}^z + a_{22i}^z\cos\theta_i + a_{23i}^z(\cos\theta_i)^2] \end{pmatrix} \tag{13}$$

$$\boldsymbol{\mu}_{1,i+1} = \begin{pmatrix} 0 \\ \sin\theta_i[a_{11,i+1}^y + a_{12,i+1}^y\cos\theta_i + a_{13,i+1}^y(\cos\theta_i)^2] \\ \sin\theta_i[a_{11,i+1}^z + a_{12,i+1}^z\cos\theta_i + a_{13,i+1}^z(\cos\theta_i)^2] \end{pmatrix} \tag{14}$$

where the as are adjustable parameters. These parameters do not depend on residue type (glycine or other, because proline does not have an H^N proton). The vector $\boldsymbol{\mu}_i$ is expressed in the local coordinate system of residue i by eq. (15).

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_{2,i} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma_i & \sin\gamma_i \\ 0 & -\sin\gamma_i & \cos\gamma_i \end{pmatrix} \boldsymbol{\mu}_{1,i+1} \tag{15}$$

Finally, the approximate coordinates of the H^N_{i+1} atom are expressed by eqs. (16 – 18).

$$x^{H^N}_{i+1} = \delta \tag{16}$$

$$y^{H^N}_{i+1} = \varepsilon \frac{\mu_{yi}}{\sqrt{\mu^2_{yi} + \mu^2_{zi}}} \tag{17}$$

$$z^{H^N}_{i+1} = \varepsilon \frac{\mu_{zi}}{\sqrt{\mu^2_{yi} + \mu^2_{zi}}} \tag{18}$$

where δ is the distance between C^α_i atom and the perpendicular projection of H^N on the $C^\alpha_i \cdots C^\alpha_{i+1}$ virtual bond and ε is the distance between H^N atom and $C^\alpha_i \cdots C^\alpha_{i+1}$ virtual bond (Figure 2). Both δ and ε are adjustable parameters.

Parameterization of the formulas

To determine the parameters of eqs. (1 – 18), we used the interproton distances calculated from a set of 140 high-resolution non-homologous protein structures selected in our earlier work[38] to derive side-chain-potential parameters. The list of these proteins, along with their PDB codes, the numbers of amino-acid residues, and the numbers of protons of a given kind, are shown in Table S1 of the Supplementary Material. For the structures in which the proton coordinates were missing, we calculated them by using the LEAP program form the AMBER14 package[39]. Overall, 24458, 22750 and 21185 H^α, H^N and H^β proton coordinate sets, respectively, were obtained. From each protein, interproton distances were calculated. Because only short ($\leqslant$ 5 Å[40]) interproton distances are detectable by NMR and because the distances shorter than 2.5 Å (the sum of van der Waals radii of hydrogens bound to heavy

atoms) are physical, only the distances from 2.5 Å to 8 Å were considered. The total number of distances was over 1,300,000. The numbers of distances corresponding to all the types of protons considered are detailed in Table 1.

We used the Levenberg-Marquardt nonlinear least-squares algorithm[41,42]. The target function is expressed by eq. (19).

$$\Phi(\boldsymbol{\xi}) = \sum_{p=1}^{140} \sum_{r,s=1}^{3} \sum_{k,l=1}^{3} \sum_{i=1}^{N_{klrsp}} \left[d_{iklrsp}^{PDB} - d_{iklrsp}^{calc}(\boldsymbol{\xi}) \right]^2 \tag{19}$$

where $\boldsymbol{\xi}$ denotes the vector of all adjustable parameters of eqs. (1 – 18), d_{iklrsp}^{PDB} is the ith interproton distance for the proton of kinds k and l (H^{α}, H^{β} or H^{N}) and residue of types r and s (glycine, proline or other) calculated from the PDB structure of protein p, $d_{iklrsp}^{calc}(\boldsymbol{\xi})$ is the corresponding distance calculated from approximate proton coordinates that are, in turn, calculated using (1 – 18), and N_{klrsp} is the number of proton-proton distances for protons of kind k and l of residues of kinds r and s.

Because the total number of parameters for all proton types is large, fitting was carried out in two stages. In stage 1, only the parameters for a given proton type were fitted (e.g., H^{α} of glycine), using the pertinent subset of data. In this stage, the necessary number of parameters was also established by gradually increasing the number of parameters in the respective formula(s), setting the remaining ones at the default values (0 to all except for those which have the sense of angles). The statistical significance of adding new parameters was assessed by means of the F-test[43]. In stage 2, all parameters were fitted together using the complete set of data, and starting from the values obtained in stage 1.

Validation of the method

The method was validated with a set of 41 proteins of the NMR/X-Ray pairs of structures database[44]. The PDB IDs of these proteins are listed in Table S2 of the Supplementary Material. The NMR-structure ensemble of each protein consisted of 20 models, except for the 1PQX entry, which contains only 10 models. For each of the protein and for each structure of the NMR ensemble, we extracted the C^{α} coordinates and rebuilt the all-atom chain by using the PULCHRA software[45]. Because PULCHRA does not define proton coordinates,

these were determined from the PULCHRA structures by using the LEAP program of the AMBER14 package[39]. Subsequently, we used ESCASA to estimate proton coordinates from the C^{α}-traces of the structures of the NMR ensembles of each of the 41 proteins. For each protein, the $H^{\alpha} \cdots H^{\alpha}$, $H^{\beta} \cdots H^{\beta}$ and $H^{N} \cdots H^{N}$ proton pairs that were within the 8 Å cut-off distance were subsequently identified and the averages over all 20 (or 10 for 1PQX) models were computed for the original NMR ensemble as well as for the ensembles obtained with ESCASA and with PULCHRA. The model-averaged proton-proton distances from both methods were subsequently compared with those from NMR structures.

RESULTS AND DISCUSSION

Results of fitting

The parameters of eqs. (1 – 4), for the calculation of the positions of H^{α}s, of eqs. (5 – 12), for the calculation of the positions of H^{β}s, and of eqs. (13 – 18) for calculating the positions of H^{N}s are summarized in Tables 2, 3, and 4, respectively. The plots of the actual vs. the estimated interproton distances for the three kinds of protons are shown in Figure 3. The standard deviations of the estimated from the actual positions over the pairs of proton types and averaged over the types of one or both protons in a pair are collected in Table 5. It can be seen that the distances between the H^{α} protons have the smallest (0.22 Å) and those between the H^{β} protons have the largest (0.63 Å) standard deviation. The standard deviation over all proton types is 0.44 Å.

To illustrate the quality of the proton positions estimated by using ESCASA, with the parameter fitted as discussed above, the fragments of the De novo Designed Protein Foldit3 (PDB: 6msp)[46] with hydrogen-atom positions of all three types from the average NMR structure and those calculated by using ESCASA are shown in Figure 5. The mean standard deviations of the approximate positions of the H^{α}, H^{β}, and H^{N} atoms, excluding those of the 17-residue disordered N-terminal section, from the respective experimental positions were 0.16 Å, 0.59 Å, and 0.40 Å, respectively. The experimental coordinates of the H^{β} atoms corresponding to the methylene and methyl groups were averaged to compare with

the estimated positions.

The results of fitting for all three proton types are discussed below.

H^α protons. As can be seen from Table 2, for glycine the only fitted parameter of eqs. (1 – 4) was the $C^\alpha \cdots H^\alpha$ distance, $d^{H^\alpha C^\alpha}$. The parameter $\alpha_0^{H^\alpha}$ of the expression for the zenith angle α^{H^α} of the Q^α (the "average" H^α) atom from the virtual-bond-angle bisector (Figure 1), was set at 180°, which means that the Q^α atom of glycine lies approximately on the bisector and points outwards the $C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1}$ frame. The standard deviation of the estimated interproton distances was 0.202 Å, regardless of whether $d^{H^\alpha C^\alpha}$ was the only adjustable parameters or more parameters were considered (Table 6). The statistical significance of extending the set of adjustable parameters beyond $d^{H^\alpha C^\alpha}$ is less than 90 %, as assessed by the F-test (90 % confidence level). Therefore, eqs. (1 – 3) with $\alpha = 180°$ are sufficient to define the approximate local coordinates of glycine H^α atoms.

For the proline and the other residue types, the full-blown expansion of the $\cos\alpha^{H^\alpha}$ must be used to obtain good accuracy of fitting. As can be seen from Table 6, with only the $d^{H^\alpha C^\alpha}$ parameter, the standard deviations are 0.66 and 0.87 Å for the proline and the other residue types, respectively, while using the 6-order expansion of the $\cos\alpha^{H^\alpha}$ resulted in decreasing the standard deviations to 0.26 Å and 0.25 Å, respectively, the decrease being significant by F-test at the 99 % confidence level for non-proline residues. For proline, the significance of including the last group of parameters was 95 % due to the fact that there were only 120 pairs of proline H^α protons within the 8 Å cut-off in the data set.

To determine whether there are any systematic deviations of the calculated from the experimental $H^\alpha \cdots H^\alpha$ distances, we made a scatter plot of these differences in the virtual-bond angle θ (Figure 4A). As can be seen from the Figure, the differences are distributed symmetrically, regardless of θ.

H^β protons. As mentioned in *Methodology*, the parameters of eqs. (5 – 12) for the approximate positions of H^β protons were assumed to be independent of residue type (proline or other, because glycine does not have H^β protons). Splitting the parameters into the parameters for the proline and the other residue types did not improve the fit of the estimated

distances to those calculated from the PDB. However, in contrast to estimating the positions of the H^{α} atoms, both the zenith angle $\alpha^{H^{\beta}}$ and the azimuth angle $\beta^{H^{\beta}}$ (Figure 1) had to be made variable. The standard deviation of the estimated distances from those calculated from the PDB structures was 1.26 Å, when the only adjustable parameter was $d^{H^{\beta}C^{\alpha}}$ and $\alpha^{H^{\beta}}$ was set at 180°, decreasing to 0.71 Å when $\alpha_{\circ}^{H^{\beta}}$ was additionally selected as an adjustable parameter, to 0.66 Å when the full-blown 6th order expansion for $\cos\alpha^{H^{\beta}}$ was used and, finally to 0.63 Å with adding the azimuth angle $\beta^{H^{\beta}}$ (Table 7). The improvement of fitting following the introduction of each group of parameters is significant at the 99 % confidence level.

The standard deviation of the estimated from the calculated $H^{\beta}\cdots H^{\beta}$ distances is remarkably greater than that for the $H^{\alpha}\cdots H^{\alpha}$ distances (Table 5), which results from the fact that the positions of the H^{β} protons are more uncertain than those of the H^{α} protons given only the C^{α}-trace geometry. This greater uncertainty is manifested by greater deviations of the estimated from the experimental H^{β} positions in the fragments of the 6msp protein shown in Figure 5C and D.

As for the H^{α} protons, we checked if there were any systematic deviations of the estimated H^{β} positions from those calculated from the PDB. The respective plot is shown in Figure 4B. As shown in the Figure, the differences are symmetrically distributed regardless of the virtual-bond angle θ.

H^N (amide) protons. The standard deviation of the estimated $H^N\cdots H^N$ distances from those calculated from the experimental structures is 0.34 Å (Table 5), this suggesting that the approximate formulas [eqs. (13 – 18)] with 14 adjustable parameters (collected in Table 4) are a good approximation to the actual proton positions. This observation is confirmed by the scatter plot of the distances calculated from the estimated and actual proton coordinates shown in Figure 3C and by Figure 5E and F, in which the estimated and actual positions of the H^N protons are compared for an α-helical and β-sheet section of 6msp. The standard deviation of the estimated from the experimental positions of the H^N protons of 6msp is 0.40 Å. As can be seen from Table 8, the full-blown expression for approximate proton coordinates with all 14 adjustable parameters is necessary to obtain good agreement; introducing

new groups of parameters results in remarkable decrease of the standard deviation and is significant at the 99 % confidence level, as assessed by the F-test (Table 8).

Validation

As mentioned in *Methodology*, we used the set of 41 NMR structures[44] (see Table S2 of the Supplementary Material), none of which was used in parameterization, to validate our approach, as described in the *Validation* section of *Methodology*. The plots of the deviations from the mean $H^{\alpha} \cdots H^{\alpha}$, $H^{\beta} \cdots H^{\beta}$, and the mean $H^{N} \cdots H^{N}$ distances calculated by using the proton positions estimated by using PULCHRA[34] or ESCASA, respectively, together with error bars which indicate the standard deviations of these differences, are shown in Figure 6A–C. The average values and the standard deviations shown in the Figure (each corresponding to a given benchmark protein and a given proton pair) are defined by eqs. (20) and eq. (21), respectively.

$$\overline{\Delta d}_c = \overline{d}_c - \overline{d}_{NMR} \tag{20}$$

$$\sigma_{\overline{\Delta d}_c} = \sqrt{\frac{1}{N_s}\left(\sigma_c^2 + \sigma_{NMR}^2\right)} \tag{21}$$

where $\overline{d}_c$ is the average proton-proton distance calculated by using PULCHRA or ESCASA, respectively, $\overline{d}_{NMR}$ is the respective average proton-proton distance calculated from NMR structures, σ_c and σ_{NMR} are the standard deviations from the above averages, calculated over the conformations of the NMR ensemble, $\overline{\Delta d}_c$ is the difference of the average proton-proton distance calculated by using PULCHRA or ESCASA, respectively, and that from the NMR structures, $\sigma_{\overline{\Delta d}_c}$ is the estimated standard deviation of this difference, and N_s is the number of conformations in the NMR ensemble.

The values of the mean standard deviations of the ensemble-average proton-proton distances calculated with PULCHRA and with ESCASA from the corresponding ensemble-averaged distances calculated from the NMR structures, averaged over all proton pairs of of a given benchmark protein, are summarized in Table S3 of the Supplementary Material. The deviations from the mean proton-proton distances calculated from the NMR structures

are slightly greater for ESCASA (0.17 Å, 0.48 Å, and 0.36 Å on average for H^{α}, H^{β}, and H^{N}, respectively) compared to PULCHRA (0.11 Å, 0.38 Å, and 0.32 Åon average for H^{α}, H^{β}, and H^{N}, respectively). This is understandable, because our method does not explicitly include valence-geometry constraints. To determine how the above standard deviations compare with the uncertainty of proton-proton distances from the NMR structures, we estimated the standard deviations of the mean proton-proton distances from the NMR-ensemble data from eq. (22).

$$\overline{\sigma^{NMR}_{H^X\cdots H^X}} = \sqrt{\frac{1}{N_{H^XH^X}} \sum_{i=1}^{N_{H^X\cdots H^X}} \frac{1}{N_s(N_s-1)} \sum_{j=1}^{N_s} \left(d_{ji}^{H^X\cdots H^X} - \overline{d_i^{H^X\cdots H^X}} \right)^2} \tag{22}$$

where $d_{ji}^{H^X\cdots H^X}$ is the distance of the ith proton pair of type X in jth conformation of the corresponding NMR ensemble and $\overline{d_i^{H^X\cdots H^X}}$ is the corresponding distance averaged over the conformations of the NMR ensemble, $N_{H^X\cdots H^X}$ is the number of proton-proton pairs of type X in the whole set of 41 proteins, and N_s is the number of conformations in each NMR ensemble. As shown in Table S3, these standard deviations are 0.12 Å, 0.15 Å, and 0.10 Å for the H^{α}, H^{β}, and H^{N} protons, respectively. The standard deviations averaged over all proton types are 0.25 Å, 0.21 Å, and 0.12 Å for ESCASA, PULCHRA, and NMR structures, respectively. Thus, for both ESCASA and PULCHRA the error of the method is of the order of the error of NMR structure determination for H^{α} and is about 2.5–3 times greater for H^{β} and H^{N} and about 2 times greater when averaged over all proton types. It is worth noting that, for H^{β} and H^{N}, detailed all-atom chain reconstruction using PULCHRA does not bring the standard deviation of proton-proton distance remarkably closer to that resulting from the uncertainty of NMR-determined structures. This suggests that the value of about 0.4–0.5 Å is the limit of the accuracy of reproducing the distances involving H^{β} and H^{N} protons. Nevertheless, it should be noted that the 0.5 Å mean distance error is small, given the fact that the NMR restraints are upper distance limits, which are assigned quite arbitrarily given the presence of a respective NOE signal (usually 5 Å or 6 Å). Consequently, ESCASA provides good estimates of the backbone and H^{β} proton distances from the C^{α}-trace geometry.

CONCLUSIONS

We have proposed and parameterized analytical formulas for the positions of H^α, H^N and H^β protons of proteins, which are based on C^α-trace geometry alone. The parameters of the formulas for the H^α protons depend on reduced residue type (glycine, proline, and other), while those of the H^β and H^N protons do not (it should be noted, though, that there are no H^β protons for glycine and there is no H^N proton for proline). For residue i, the positions of H^α and H^β in the local coordinate system of this residue depend on the $C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1}$ virtual-bond angle θ_i [eqs. (1 – 12) and Figure 1], while that of the proton of the peptide group between residue i and $i+1$ depends on the virtual-bond angles θ_i and θ_{i+1} and on the $C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1} \cdots C^\alpha$ virtual-bond dihedral angle γ_i [eqs. (13 – 18) and Figure 2].

The standard deviation of the estimated interproton distances over all proton and all residue types, calculated over all 140 proteins used in parameterization, is 0.44 Å (Table 5). The distances between the H^α protons are fitted most accurately, with the standard deviation of 0.22 Å, while those between amide protons give not so good fit, with the standard deviation of 0.34 Å. The goodness of fit decreases even more H^β protons, the standard deviation being 0.63 Å, which probably results from the fact that only the C^α-trace geometry is used to estimate their positions. As shown in Figure 5 with the example of the 6msp protein, the proton positions estimated with ESCASA match those in the NMR-determined structures well.

Subsequent validation of the method with the set of 41 benchmark proteins[44], for which both X-ray and NMR structures are available and none of which was used in parameterization, resulted in the mean standard deviations between the ensemble-averaged proton-proton distances calculated by the method developed in this work and those calculated from NMR structures of 0.17 Å, 0.48 Å, and 0.36 Å for the H^α, H^β, and H^N proton distances, respectively, which are only slightly greater than those calculated from the C^α-traces converted to all-atom representation by using PULCHRA[34] (0.11 Å, 0.38 Å, and 0.32 Å, respectively). Except for those of the mean H^α distances, the standard deviations computed with ESCASA are remarkably greater than the estimated standard deviation of the mean proton-proton distances of the NMR-determined structures (0.12 Å, 0.15 Å, and 0.10 Å, respectively). How-

ever, detailed all-atom-chain reconstruction using PULCHRA does not result in a significant improvement, this suggesting that the accuracy limit of the calculation of the H^{β} and H^{N} distances given C^{α}-trace has been reached. Moreover, the mean error of up to 0.5 Å does not seem to be big in view of the fact that the NMR distance restraints are upper distance boundaries, which are set given the presence of NOE or related signals.

The derived formulas enable us to calculate the distances between the H^{α}, H^{β}, and H^{N} protons and the respective forces in coarse-grained coordinates. Therefore, they can be used in NMR-data-assisted coarse-grained molecular dynamics simulations of protein structures, without having to convert the coarse-grained structures to all-atom structures. The formulas developed in this work can be used with any coarse-grained model that keeps α-carbon-atom coordinates, which includes the CABS model[28] and the UNRES model[30–32]. Given the fact that, because of averaging out the fine-grain degrees of freedom, coarse-grained approaches enable us to extend the time scale of simulations by at least 3 orders of magnitude[9], use of ESCASA will enable us to run NMR-data-assisted simulations of large proteins.

The formulas have been implemented in UNRES and the results of tests of the full-blown NMR-assisted structure modeling with UNRES will be reported in our next paper. Further development of ESCASA includes sidechain protons further to H^{β}, developing the formulas for backbone and sidechain carbon, oxygen, and nitrogen atoms, and integrating the method with the algorithms for computation of Residual Dipolar Coupling (RDC)[47] intensities and chemical shifts[48,49] given atomic coordinates, which will enable us to use the full set of observables available from NMR measurements in data-assisted protein-structure modeling with UNRES.

DATA AVAILABILITY

The source code of the version of UNRES with the NMR-assisted-simulation feature, which uses the ESCASA algorithm, and ESCASA parameters are available from the Downloads section of the UNRES package page (https://unres.pl/downloads, files unres-src-HCD-5D_-nmr-May-5-2021.tar.gz and PARAM-May-5-2021.tar.gz, respectively).

The experimental structures of the proteins used to parameterize the method and their basic characteristics (oligomeric state, chain length, number of protons of particular types)

are summarized in Tables S1 and S2 of the Supplementary Material. Condensed results of the statistical analysis of the goodness of fit of the calculated interproton distances and those resulting from proton coordinates carried out for 41 proteins of the validation set are summarized in Table S3 of the Supplementary Material. Raw data and the program for nonlinear least-squares fitting can be obtained from the authors upon request; please email to adam.liwo@ug.edu.pl.

ACKNOWLEDGMENTS

We thank Prof. Gaetano T. Montelione, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute for providing high-quality NMR structures of 41 proteins using for validation of our method. This work was supported by grant No. UMO-2017/25/B/ST4/01026 from the National Science Center of Poland (Narodowe Centrum Nauki). Computational resources were provided by (a) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) the University of Warsaw under grant No. GA76-11, (b) the Centre of Informatics - Tricity Academic Supercomputer & network (CI TASK) in Gdańsk, (c) the Academic Computer Centre Cyfronet AGH in Krakow under grants: asunres18 and unres19, and (d) 796-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

References

[1] A. Kolinski and J. Skolnick, Polymer **45**, 511 (2004).

[2] A. Kolinski and J. Skolnick, Proteins: Struct., Funct., Bioinf. **32**, 475 (1998).

[3] G. Voth, *Coarse-Graining of Condensed Phase and Biomolecular Systems* (CRC Press, Taylor & Francis Group, 2008), 1st ed.

[4] S. J. Marrink and D. P. Tieleman, Chem. Soc. Rev. **42**, 6801 (2013).

[5] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, Chem. Rev. **116**, 7898 (2016).

[6] N. Singh and W. Li, Int. J. Mol. Sci. **20**, 3774 (2019).

[7] A. Liwo, C. Czaplewski, A. K. Sieradzan, E. A. Lubecka, A. G. Lipska, Ł. Golon, A. Karczyńska, P. Krupa, M. A. Mozolewska, M. Makowski, et al., in *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*, edited by B. Strodel and B. Barz (Academic Press, 2020), vol. 170 of *Progress in Molecular Biology and Translational Science*, pp. 73 – 122.

[8] M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H. A. Scheraga, J. Phys. Chem. B **109**, 13785 (2005).

[9] M. Khalili, A. Liwo, A. Jagielska, and H. A. Scheraga, J. Phys. Chem. B **109**, 13798 (2005).

[10] H. Kamisetty, S. Ovchinnikov, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **110**, 15674 (2013).

[11] E. A. Lubecka and A. Liwo, J. Comput. Chem. **40**, 2164 (2019).

[12] Y. Zhang and J. Skolnick, Proteins: Struct. Funct. Bioinf. **57**, 702 (2004).

[13] K. Joo, I. Joung, S. Y. Lee, J. Y. Kim, Q. Cheng, B. Manavalan, J. Y. Joung, S. Heo, J. Lee, M. Nam, et al., Proteins: Struct., Funct., Bioinf. **84**, 221 (2015).

[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucl. Acid Res. **28**, 235 (2000).

[15] D. A. Case, H. J. Dyson, and P. E. Wright, in *NMR in Proteins*, edited by G. Clore and A. Gronenborn (MacMillan, New York, 1993), pp. 53–91.

[16] K. Joo, I. Joung, J. Lee, J. Lee, W. Lee, B. Brooks, S. J. Lee, and J. Lee, Proteins: Struct., Funct., Bioinf. **83**, 2251 (2015).

[17] D. Kimanius, I. Pettersson, G. Schluckebier, E. Lindahl, and M. Andersson, J. Chem. Theory Comput. **11**, 3491 (2015).

[18] A. Leitner, L. A. Joachimiak, P. Unverdorben, T. Walzthoeni, J. Frydman, F. Förster, and R. Aebersold, Proc. Natl. Acad. Sci. U.S.A. **111**, 9455 (2014).

[19] M. Grimm, T. Zimniak, F. Herzog, and A. Kahraman, Nucl. Acids Res. **43**, W362 (2015).

[20] J. Fajardo, R. Shrestha, N. Gil, A. Belsom, S. Crivelli, C. Czaplewski, K. Fidelis, S. Grudinin, M. Karasikov, A. Karczyńska, et al., Proteins: Struct., Funct., and Bioinf. **87**, 1283 (2019).

[21] M. Dimura, T. O. Peulen, C. A. Hanke, A. Prakash, H. Gohlke, and C. A. Seidel, Curr. Opinion Struct. Biol. **40**, 163 (2016).

[22] B. Hellenkamp, S. Schmid, O. Doroshenko, O. Opanasyuk, R. Kühnemuth, S. Rezaei Adariani, B. Ambrose, M. Aznauryan, A. Barth, V. Birkedal, et al., Nature Meth. **15**, 669 (2018).

[23] M. P. Williamson, T. F. Havel, and K. Wüthrich, J. Mol. Biol. **182**, 295 (1985).

[24] J. Cavanagh, W. J. Fairbrother, A. G. Palmer, M. Rance, and N. J. Skelton, eds., *Protein NMR Spectroscopy* (Academic Press, Burlington, 2007), 2nd ed.

[25] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, in *Numerical Computer Methods, Part D* (Academic Press, 2004), vol. 383 of *Methods in Enzymology*, pp. 66 – 93.

[26] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, J. Phys. Chem. B **116**, 8494 (2012).

[27] D. Latek and A. Koliński, J. Comput. Chem. **32**, 536 (2011).

[28] A. Kolinski, Acta Biochim. Polym. **51**, 349 (2004).

[29] E. A. Lubecka, A. S. Karczyńska, A. G. Lipska, A. K. Sieradzan, K. Zięba, C. Sikorska, U. Uciechowska, S. A. Samsonov, P. Krupa, M. A. Mozolewska, et al., J. Molec. Graph. Model. **92**, 154 (2019).

[30] A. Liwo, C. Czaplewski, S. Ołdziej, A. V. Rojas, R. Kaźmierkiewicz, M. Makowski, R. K. Murarka, and H. A. Scheraga, in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. Voth (CRC Press, 2008), chap. 8, pp. 1391–1411.

[31] A. Liwo, M. Baranowski, C. Czaplewski, E. Gołaś, Y. He, D. Jagieła, P. Krupa, M. Maciejczyk, M. Makowski, M. A. Mozolewska, et al., J. Mol. Model. **20**, 2306 (2014).

[32] A. Liwo, A. K. Sieradzan, A. G. Lipska, C. Czaplewski, I. Joung, W. Żmudzińska, A. Hałabis, and S. Ołdziej, J. Chem. Phys. **150**, 155104 (2019).

[33] D. Gront, S. Kmiecik, and A. Kolinski, J. Comput. Chem. **28**, 1593 (2007).

[34] P. Rotkiewicz and J. Skolnick, J. Comput. Chem. **29**, 1460 (2008).

[35] A. K. Sieradzan, M. Makowski, A. Augustynowicz, and A. Liwo, J. Chem. Phys. **146**, 124106 (2017).

[36] S. Hu, A. Krokhotin, A. J. Niemi, and X. Peng, Phys. Rev. E **83**, 041907 (2011).

[37] A. Liwo, S. Ołdziej, C. Czaplewski, U. Kozłowska, and H. A. Scheraga, J. Phys. Chem. B **108**, 9421 (2004).

[38] A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, J. Comput. Chem. **18**, 849 (1997).

[39] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, H. Gohlke, et al., *Amber 14* (2014), University of California: San Francisco.

[40] W. Braun and N. Gō, J. Mol. Biol. **186**, 611 (1985).

[41] K. Levenberg, Q. Appl. Math. **2**, 164 (1944).

[42] D. W. Marquardt, J. Soc. Indust. Appl. Math. **11**, 431 (1963).

[43] G. A. Seber and C. J. Wild, *Nonlinear regression* (Wiley, New York, 1989).

[44] J. K. Everett, R. Tejero, S. B. K. Murthy, T. B. Acton, J. M. Aramini, M. C. Baran, J. Benach, J. R. Cort, A. Eletsky, F. Forouhar, et al., Prot. Sci. **25**, 30 (2016).

[45] P. Rotkiewicz and J. Skolnick, J. Comput. Chem. **29**, 1460 (2008).

[46] B. Koepnick, J. Flatten, T. Husain, A. Ford, D.-A. Silva, M. J. Bick, A. Bauer, G. Liu, Y. Ishida, A. Boykov, et al., Nature **570**, 390 (2019).

[47] J.-R. Huang and S. Grzesiek, J. Am. Chem. Soc. **132**, 694 (2010).

[48] B. Han, Y. Liu, S. W. Ginzinger, and D. S. Wishart, J. Biomol. NMR **50**, 43 (2011).

[49] O. A. Martin, Y. A. Arnautova, A. A. Icazatti, H. A. Scheraga, and J. A. Vila, Proc. Natl. Acad. Sci. USA **110**, 16826 (2013).

Figure 1: Illustration of the definition of the positions of H^α and H^β atoms of the ith amino-acid residue given the geometry of the $C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1}$ α-carbon-trace geometry. The C^α atoms are shown as small white spheres and the $C^\alpha \cdots C^\alpha$ virtual bonds are shown as thick gray lines. The x axis of the local coordinate system is the bisector of the $C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1}$ (θ_i) virtual-bond angle, the y axis lies in the plane of the three C^αs and runs in the direction of the chain, the z axis forms a right-handed coordinate system with the x and y axes. The H^α atom is located in the xz plane, forming the zenith angle $\alpha_i^{H^\alpha}$ with the x axis, which depends on the virtual-bond angle θ [eq. (4)], and is attached to C^α_i with the $d_i^{H^\alpha C^\alpha}$ bond length. It should be noted that for glycine the H^α position is the average position of the two H^α protons (the respective pseudo-atom is usually denoted by Q^α). The (average) position of H^β_i is defined by the zenith angle $\alpha_i^{H^\beta}$ and the azimuth angle $\beta_i^{H^\beta}$ of counter-clockwise rotation from the xy plane, which are depend on angle θ_i [eqs. (8) and (9)] and this atom is attached to C^α_i with the $d_i^{H^\beta C^\alpha}$ virtual-bond length (the virtual bond being shown as a dashed line). For illustration, the C^β atom (not used in the calculations) and the attached virtual bonds are shown in light-gray.

Figure 2: Illustration of definition of the position of the amide proton of residue i (located between C^α_i and C^α_{i+1}. The vector $\boldsymbol{\mu}_i$ that defines the orientation of the peptide group located between C^α_i and C^α_{i+1} is a sum of component from residue i ($\boldsymbol{\mu}_{2,i}$) and residue $i+1$ ($\boldsymbol{\mu}_{1,i+1}$). The position of the proton is obtained by drawing a segment with length ε_i following the direction of $\boldsymbol{\mu}_i$ (i.e., perpendicular to the $C^\alpha_i \cdots C^\alpha_{i+1}$ virtual bond) at a distance δ_i from C^α_i. The N, C', and O atoms of the peptide group are shown in light-gray symbols and the respective bonds are shown in right-gray lines for better illustration. The x, y_i, and z_i axes of th coordinate system of residue i (in which $\boldsymbol{\mu}_{2,i}$ is defined) and those of the coordinate system of residue $i+1$ (in which $\boldsymbol{\mu}_{1,i+1}$ is defined) are also shown. These two systems share the x axis, while the y and z axes of the system of residue i is rotated about the x axis by the $C^\alpha_{i-2} \cdots C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1}$ (γ_i) dihedral angle with respect to that of residue $i+1$. The y_i axis of the coordinate system of residue i is in the $C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1}$ plane and runs from C^α_i to C^α_{i-1}, while the z_i axis forms a right-handed coordinate system together with the x and y_i axis. The y_{i+1} axis of the coordinate system of residue $i+i$ is in the $C^\alpha_i \cdots C^\alpha_{i+1} \cdots C^\alpha_{i+2}$ plane and runs from C^α_{i+1} to C^α_{i+2}, while the z_{i+1} axis forms a right-handed orthonormal coordinate system together with the x and z_{i+1} axes.

Figure 3: Plots of the actual interproton distances calculated from the 140 structures of non-homologous proteins listed in Table S1 of the Supplementary Material vs. the corresponding distances calculated from approximate proton positions estimated from C^α-trace geometry [eqs. (1 – 18)]. (A): $H^\alpha \cdots H^\alpha$ distances, (B): $H^\beta \cdots H^\beta$ distances, (C) $H^N \cdots H^N$ distances, (D) all distances. The diagonal line in each panel is shown in red color.

Figure 4: A scatter plot of the differences between the differences between the interproton distances calculated from the experimental structures of 140 non-homologous proteins listed in Table S1 of the Supplementary Material and the corresponding distances calculated from the positions of the protons calculated from the C^{α}-trace geometry vs. the virtual-bond angle θ. (A): $H^{\alpha} \cdots H^{\alpha}$ distances, (B): $H^{\beta} \cdots H^{\beta}$ distances.

Figure 5: Estimated (orange, marked "Calc") and experimental (white, marked "Exp") positions of H^{α} [(A) and (B)], H^{β} [(C) and (D)], and H^{N} [(E) and (F)] protons in selected α-helix [(A), (C), and (E)] and β-sheet [(B), (D), and (F)] fragments of the 6msp protein. The distances between the experimental and the calculated H^{α} positions are (A) 0.349, 0.356 and 0.300 Å, for Leu11, Gln10 and Arg7, respectively, and (B) 0.218, 0.092 and 0.095 Å, for Val24, Glu25 and Val26, respectively. The distances between the experimental and the calculated H^{β} positions are (C) 0.656, 0.398 and 0.314 Å, for Glu6, Arg7 and Leu8, respectively, and (D) 0.460, 0.524 and 0.639 Å, for Val23, Val24 and Glu25, respectively. For amino-acid residues with more than one β hydrogen, the experimental positions are averaged over the constituent protons. The distances between the experimental and the calculated H^{N} positions are (E) 0.287, 0.191 and 0.177 Å, for Leu11, Gln10 and Leu8, respectively, and (F) 0.450, 0.149 and 0.141 Å, for His27, Val26 and Val24, respectively.

Figure 6: Plots of the differences of the proton-proton distances calculated from the proton coordinates calculated by using ESCASA (purple) and by PULCHRA[34] (green) vs. the mean distances from the NMR structures of the set of 41 proteins for which both X-ray and NMR structures are available[44] for the $H^{\alpha} \cdots H^{\alpha}$ (A), $H^{\beta} \cdots H^{\beta}$ (B), and $H^{N} \cdots H^{N}$ (C) proton pairs. The error bars of the differences of the ensemble-averaged NMR-structure and calculated values are shown as vertical lines. For clarity, only the distances up to 5 Å are shown.

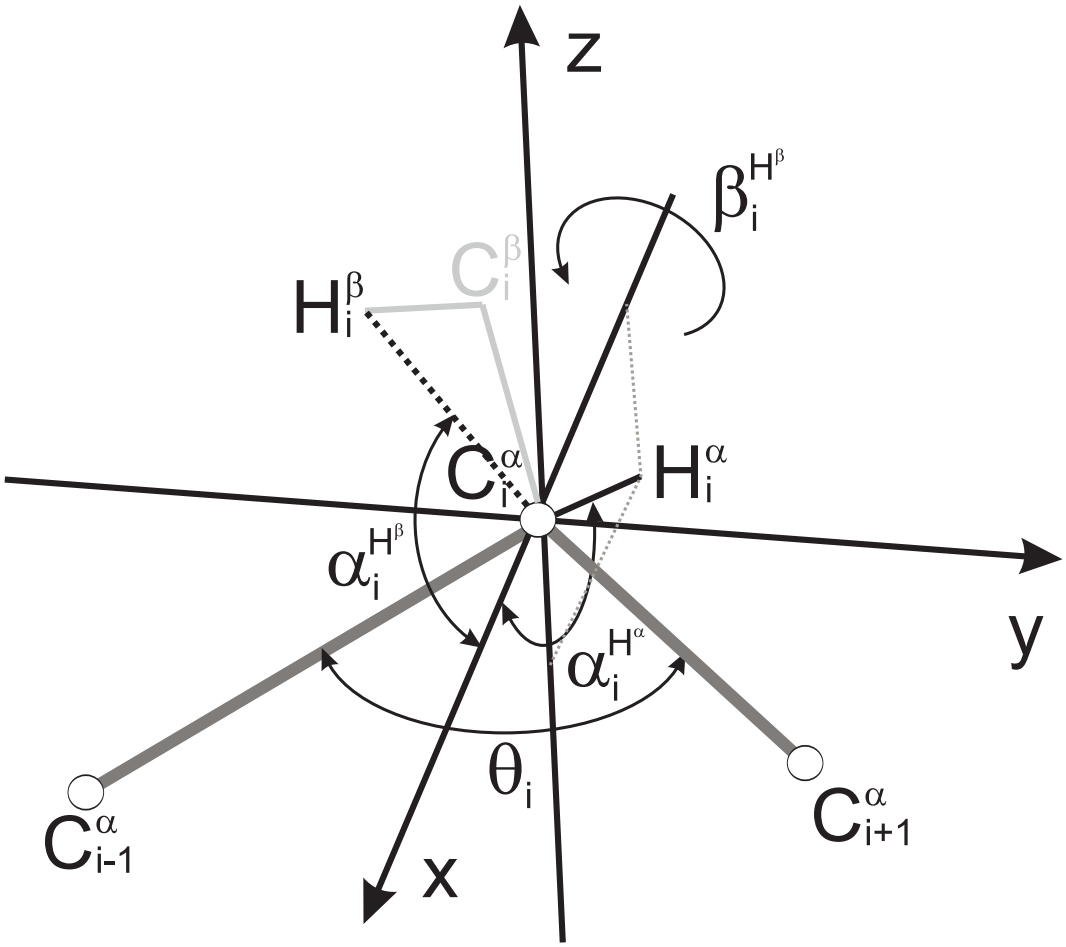


Figure 1
E.A. Lubecka, A. Liwo
J. Comput. Chem.

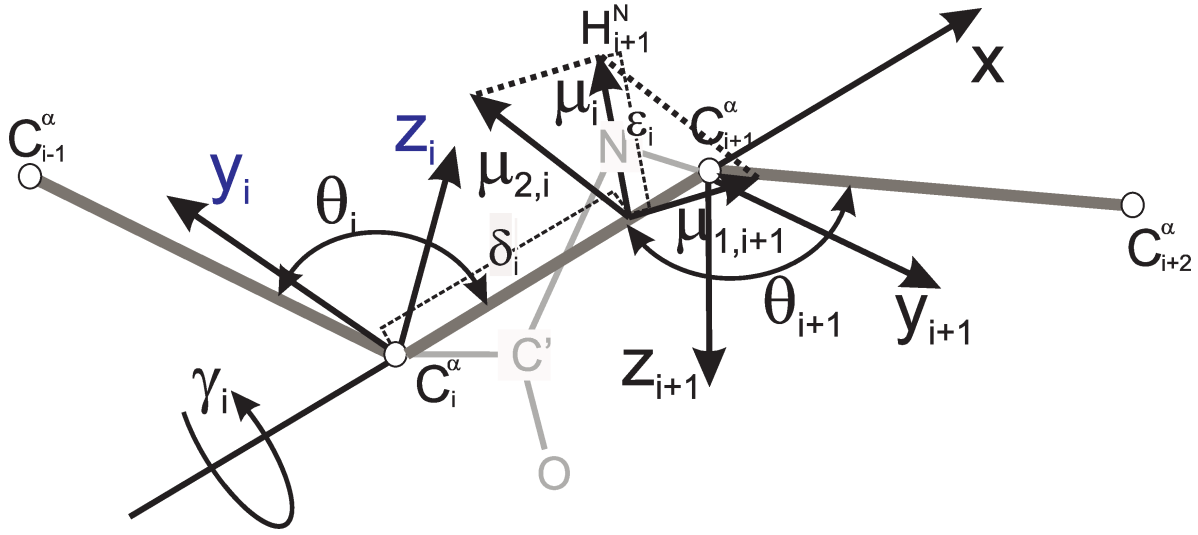


Figure 2
E.A. Lubecka, A. Liwo
J. Comput. Chem.

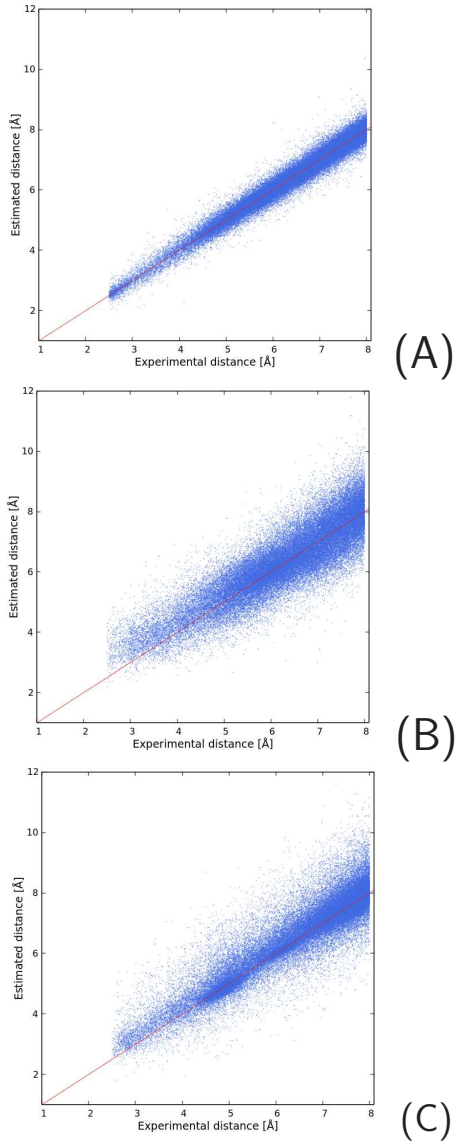


Figure 3
E.A. Lubecka, A. Liwo
J. Comput. Chem.

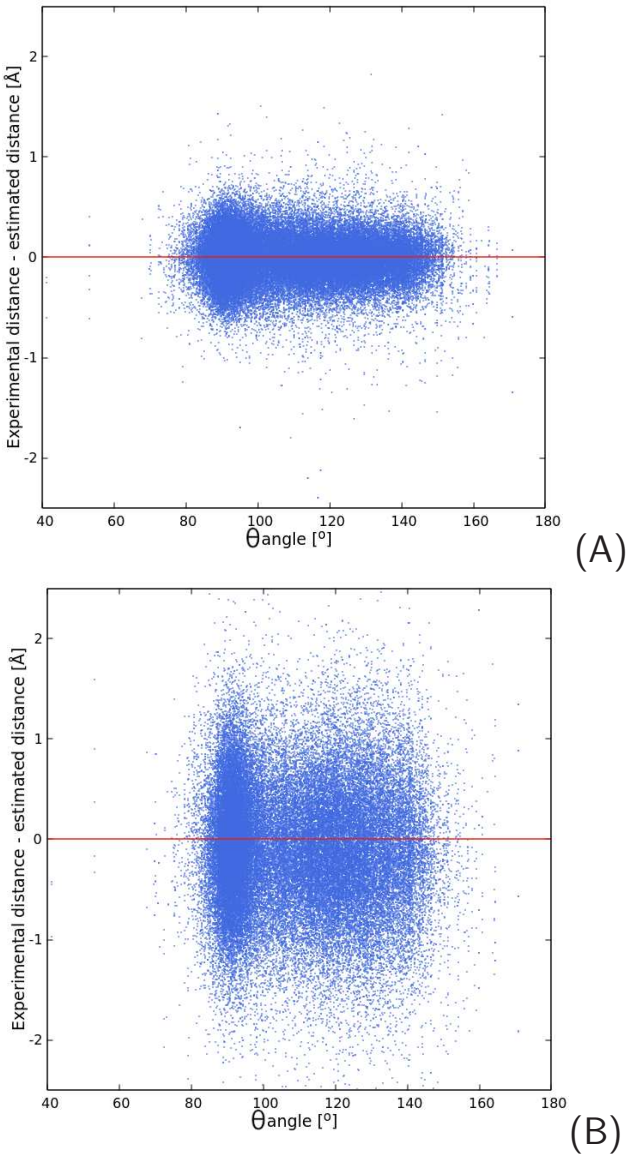


Figure 4
E.A. Lubecka, A. Liwo
J. Comput. Chem.

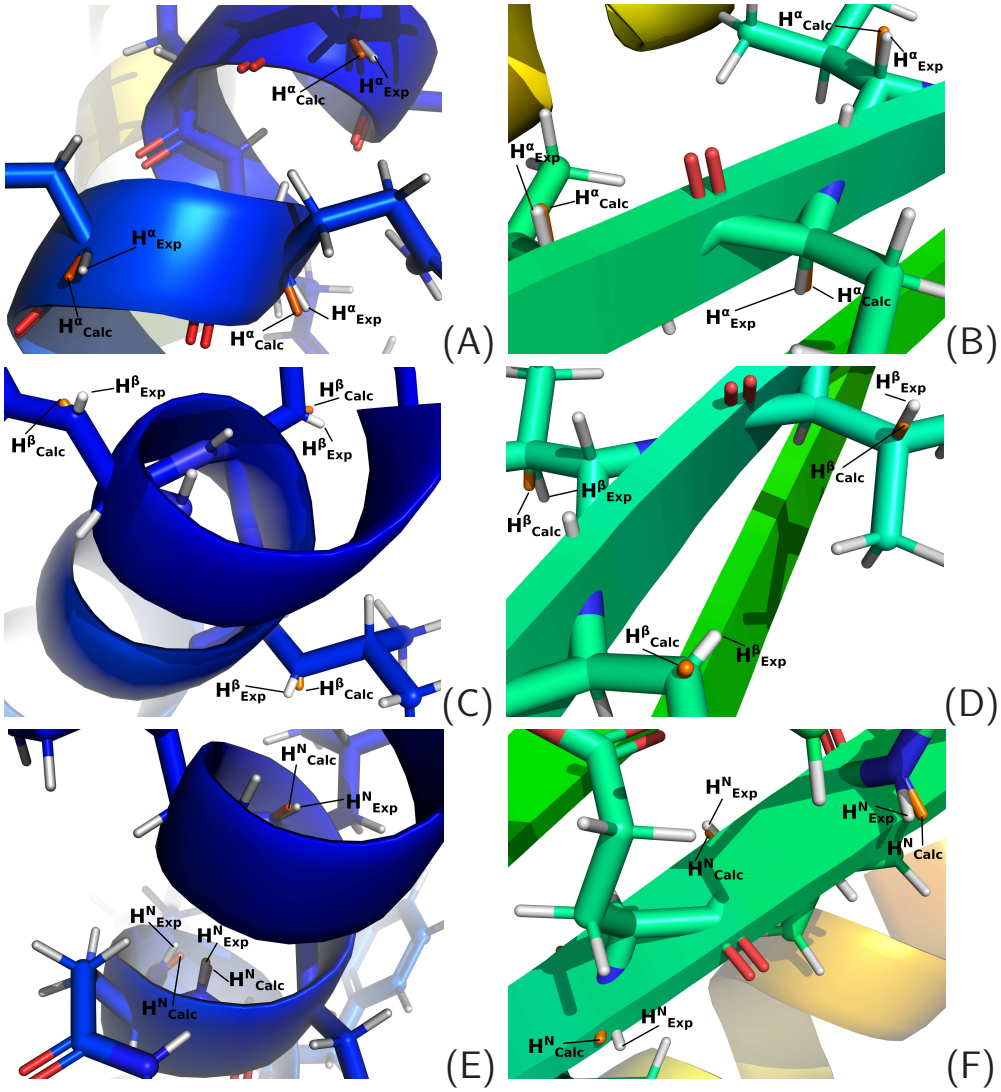


Figure 5
E.A. Lubecka, A. Liwo
J. Comput. Chem.

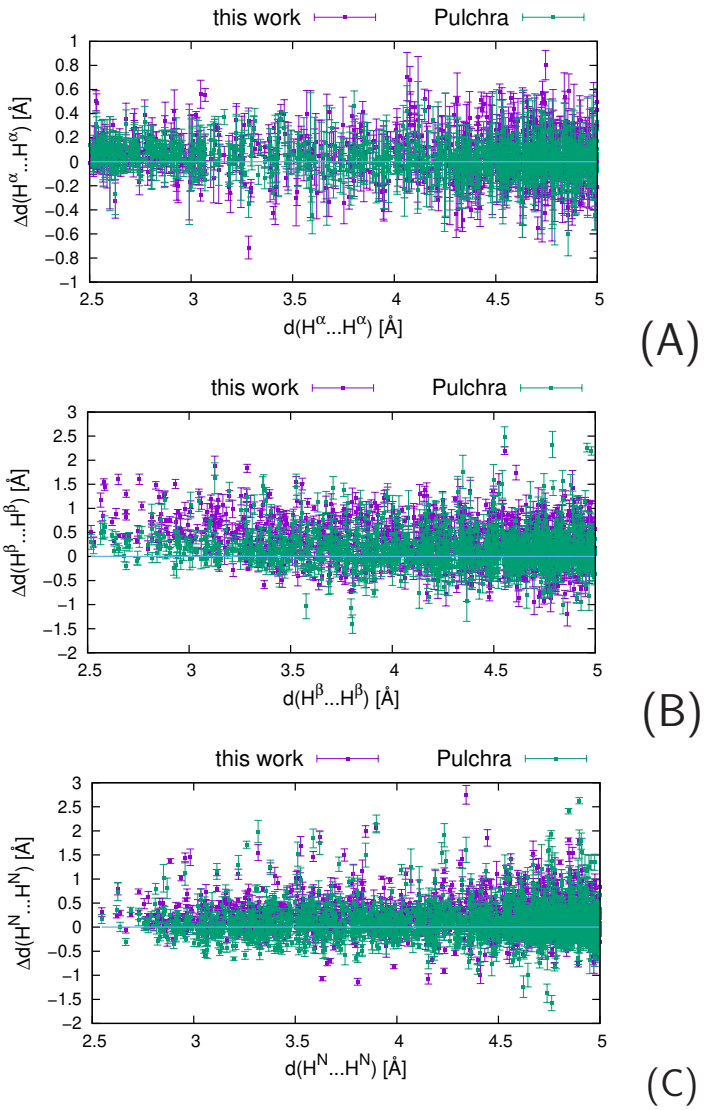


Figure 6
E.A. Lubecka, A. Liwo
J. Comput. Chem.

Table 1: Numbers of distances between protons of different kinds calculated from the set of 140 non-homologous protein structures (Table S1 of the Supplementary Material) used for parameterization of the approximate expressions of proton positions [eqs. (3 – 18)].

	Number of distances			
Proton	H^{α}	H^{N}	H^{β}	All
H^{α}	75863	146662	133336	635859
H^{N}	146662	66066	133191	625772
H^{β}	133336	133191	60259	593313
All	635859	625772	593313	1308419

Table 2: Parameters of the formulas for approximate local coordinates of the H^{α} [eqs. (1 – 4)]

Parameter	Value		
	Gly	Pro	other
$d^{H^{\alpha}C^{\alpha}}$ [Å]	0.571	1.079	1.033
α_0 [deg]	180.000	110.560	112.401
a_1	-	-1.985	-0.236
a_2	-	0.185	-0.082
a_3	-	-1.011	-0.096
a_4	-	0.270	-0.166
a_5	-	-0.273	-0.092
a_6	-	0.290	0.067

"-" - not present/omitted

Table 3: Parameters of the formulas for approximate local coordinates of the H^{β} protons [eqs. (5 – 12)].

Parameter					
Name	Value	Name	Value	Name	Value
$d^{H^{\beta}C^{\alpha}}$ [Å]	1.812				
α_0 [deg]	45.827	β_0 [deg]	-78.187		
a_1	-0.273	s_1	0.323	c_1	0.200
a_2	-0.361	s_2	0.090	c_2	0.226
a_3	-0.591	s_3	0.515	c_3	0.405
a_4	-0.659	s_4	0.279	c_4	0.210
a_5	-0.395	s_5	-0.082	c_5	0.316
a_6	-0.141	s_6	0.022	c_6	-0.043

Table 4: Parameters of the formulas for the approximate local coordinates of the H^N protons [eqs. (13 – 18)].

Parameter		
name	**value [Å]**	
δ	2.139	
ε	1.146	
	y	z
$a_{21,i}$	0.528	-0.199
$a_{22,i}$	1.329	-1.632
$a_{23,i}$	-0.154	-1.070
$a_{11,i+1}$	-0.259	-0.116
$a_{12,i+1}$	-0.153	0.440
$a_{13,i+1}$	0.718	0.043

Table 5: Standard deviations of the estimated interproton distances from the distances calculated from the experimental structures of 140 non-homologous proteins used to parameterize the formulas for the approximate proton positions obtained with the final sets of parameters [eqs. (1 – 18)].

	σ [Å]			
Proton	H^{α}	H^{N}	H^{β}	All
H^{α}	0.224	0.425	0.488	0.418
H^{N}	0.425	0.339	0.579	0.513
H^{β}	0.488	0.579	0.628	0.555
All	0.418	0.513	0.555	0.442

Table 6: Standard deviations (σ [Å])[a] of the estimated $H^{\alpha} \cdots H^{\alpha}$ distances from the distances in the all-atom structures obtained with increasing number of adjustable parameters, calculated using the experimental structures of 140 non-homologous proteins used to parameterize the formulas for the approximate H^{α} proton positions [eqs. (1 – 4)] and significance of including subsequent groups of parameters (P) determined by means of the F-test[43].

Parameters		Gly		Pro		other	
Number	Names	σ	P	σ	P	σ	P
1	$d^{H^{\alpha}C^{\alpha}}$	**0.202**	–	0.656	–	0.867	–
2	$d^{H^{\alpha}C^{\alpha}}$,$\alpha_{\circ}$	0.202	$< 90\%$	0.301	$> 99\%$	0.264	$> 99\%$
8	$d^{H^{\alpha}C^{\alpha}}$,$\alpha_{\circ}$,a_1-a_6	0.202	$< 90\%$	**0.266**	95%	**0.248**	$> 99\%$

[a]The standard deviations corresponding to the finally accepted set of parameters are in boldface font.

Table 7: Standard deviations (σ [Å]) of the estimated $H^{\beta} \cdots H^{\beta}$ distances from the distances in the all-atom structures obtained with increasing number of adjustable parameters, calculated using the experimental structures of 140 non-homologous proteins used to parameterize the formulas for the approximate H^{α} proton positions [eqs. (5 – 12)] and significance of including subsequent groups of parameters (P) determined by means of the F-test[43].

Parameters		σ	P
Number	Names		
1	$d^{H^{\beta}C^{\alpha}}$	1.263	–
2	$d^{H^{\beta}C^{\alpha}}$,α_0	0.715	$> 99\%$
3	$d^{H^{\beta}C^{\alpha}}$,α_0,β_0	0.685	$> 99\%$
8	$d^{H^{\beta}C^{\alpha}}$,α_0,a_1-a_6	0.661	$> 99\%$
9	$d^{H^{\beta}C^{\alpha}}$,α_0,a_1-a_6,β_0	0.633	$> 99\%$
15	$d^{H^{\beta}C^{\alpha}}$,α_0,a_1-a_6,β_0,s_1-s_6	0.630	$> 99\%$
21	$d^{H^{\beta}C^{\alpha}}$,α_0,a_1-a_6,β_0,s_1-s_6,c_1-c_6	0.628	$> 99\%$

Table 8: Standard deviations (σ [Å]) of the estimated $H^N \cdots H^N$ distances obtained using the different sets of parameters from those calculated from the experimental structures of 140 non-homologous proteins used to parameterize the formulas for the approximate H^N proton positions [eqs. (13 – 18)] and significance of including subsequent groups of parameters (P) determined by means of the F-test[43].

Parameters		σ	P
Number	Names		
1	δ	0.824	
2	δ,ϵ	0.637	$> 99\%$
5	δ,ϵ,$a^y_{21,i} - a^y_{23,i}$	0.555	$> 99\%$
8	δ,ϵ,$a^y_{21,i} - a^y_{23,i}$,$a^z_{21,i} - a^z_{23,i}$,	0.374	$> 99\%$
11	δ,ϵ,$a^y_{21,i} - a^y_{23,i}$,$a^z_{21,i} - a^z_{23,i}$, $a^y_{11,i+1}$-$a^y_{13,i+1}$	0.345	$> 99\%$
14	δ,ϵ,$a^y_{21,i} - a^y_{23,i}$,$a^z_{21,i} - a^z_{23,i}$ $a^y_{11,i+1}$-$a^y_{13,i+1}$,$a^z_{11,i+1} - a^z_{13,i+1}$	0.339	$> 99\%$