Published in final edited form as:

Comput Theor Chem. 2017 January 1; 1099: 152–166. doi:10.1016/j.comptc.2016.11.025.

Error assessment in molecular dynamics trajectories using computed NMR chemical shifts

Dr. David R. Koes, PhD and Dr. John K. Vries, MD

University of Pittsburgh

Abstract

Accurate chemical shifts for the atoms in molecular mechanics (MD) trajectories can be obtained from quantum mechanical (QM) calculations that depend solely on the coordinates of the atoms in the localized regions surrounding atoms of interest. If these coordinates are correct and the sample size is adequate, the ensemble average of these chemical shifts should be equal to the chemical shifts obtained from NMR spectroscopy. If this is not the case, the coordinates must be incorrect. We have utilized this fact to quantify the errors associated with the backbone atoms in MD simulations of proteins. A library of regional conformers containing 169,499 members was constructed from 6 model proteins. The chemical shifts associated with the backbone atoms in each of these conformers was obtained from QM calculations using density functional theory at the B3LYP level with a 6-311+G(2d,p) basis set. Chemical shifts were assigned to each backbone atom in each MD simulation frame using a template matching approach. The ensemble average of these chemical shifts was compared to chemical shifts from NMR spectroscopy. A large systematic error was identified that affected the ^1H atoms of the peptide bonds involved in hydrogen bonding with water molecules or peptide backbone atoms. This error was highly sensitive to changes in electrostatic parameters. Smaller errors affecting the $^{13}\text{C}^\alpha$ and ^{15}N atoms were also detected. We believe these errors could be useful as metrics for comparing the force-fields and parameter sets used in MD simulation because they are directly tied to errors in atomic coordinates.

Keywords

Molecular dynamics; Chemical shift; Peptide bonds; Force fields

1. Introduction

Molecular dynamics simulations (MD) are essential tools for understanding the structure and function of proteins at the atomic level [1]. Advances in computer technology enabling MD simulations on millisecond time scales have further enhanced the biological relevance of these simulations [2]. The ability to reproduce experimentally accessible properties has also shown steady improvement with enhancements to the empirical force fields supporting MD

simulations [3]. Unfortunately, MD simulations still have shortcomings. The most widely used force fields are not polarizable [4], cutoffs and higher order relationships are not handled in a consistent fashion [5], aromatic-aromatic interactions may be inaccurate without special optimizations [6], and a water model must be selected from a bewildering array of possibilities with different bulk properties [7-10]. These difficulties are compounded by algorithmic differences between the major MD simulation packages [11]. Parameter optimization for force fields based on gas phase quantum mechanical (QM) studies is another problematic area [12]. The final parameters may not be a good choice for a specific protein or a different chemical environment. Many parameters require additional manipulation to accurately reproduce experimental properties in the condensed phase [13-17]. Further improvements are necessary to achieve the full potential of this technology.

New simulation methodologies must be validated by comparing properties calculated from simulation trajectories with experimental values. These properties are a function of the atoms in each frame and their associated coordinates. Properties may also reflect the time dependencies or ensemble statistical distributions of these atomic positions. In all cases, errors in calculated properties are ultimately related to errors in atomic coordinates. A large number of properties are available for this purpose, ranging from simple RMSD structural comparisons to complex thermodynamic variables. Recently, several laboratories have investigated the possibility of using NMR observables for error assessment [18-25]. These observables include order parameters, residual dipolar couplings, scalar J couplings and NMR chemical shifts. The latter are particularly attractive because they are widely available through the BioMagResBank (BMRB) [26]. A variety of empirical programs based on sequence and/or structural similarity are currently available for estimating chemical shifts in proteins. They provide accurate values for ensemble averages in most cases. These programs include SHIFTY, SHIFTX, SHIFTX2, SPARTA+, SHIFTCALC, SHIFTS, and PROSHIFT [26-33]. In 2012, Beauchamp et al successfully used empirically derived scalar J couplings and chemical shifts to evaluate 11 different force fields and 5 different water models [34]. Recently, Li and Brüschweiler developed a protocol for optimizing force field constants for individual proteins based on empirically derived chemical shifts for backbone carbon atoms [35]. Despite their utility, empirical methods also have limitations. A recent QM study by Sumowski et al [36] compared the accuracy of empirically derived chemical shifts with shifts calculated at high levels of QM theory. They showed that empirical programs were accurate for predicting ensemble averages, but were insensitive to local structural variations. These variations are obscured because empirical programs augment the information derived from structural coordinates with statistical information extracted from external sources.

In 2009 Vila et al showed that accurate $^{13}\text{C}^{\alpha}$ chemical shifts could be obtained for ubiquitin from QM calculations if the bond lengths and angles of the ubiquitin conformers were adjusted to a rigid ECEPP geometry [37,38]. They used this approach to develop a database of backbone conformers with known $^{13}\text{C}^{\alpha}$ chemical shifts. They subsequently developed a utility called CheShift [39] that was capable of assigning accurate $^{13}\text{C}^{\alpha}$ chemical shifts to the atoms in individual MD simulation frames using a template matching approach. In the present study, we show that it is feasible to extend this approach to include the ^1H and ^{15}N atoms of the peptide backbone. In contrast to CheShift, we generate our database with uncorrected atomic coordinates. We do this because our primary interest lies in the error

associated with flawed atomic coordinates. The chemical shift of an atom obtained by template matching will be incorrect if the coordinates of the atoms in its local surround are incorrect. Incorrect coordinates are in turn caused by defects in the force fields, algorithms or parameter sets underlying MD simulation. Comparison of the ensemble average of these chemical shifts with chemical shifts from NMR spectroscopy provides an error value that bares a direct relationship to these defects. The workflow associated with chemical shift assignment is illustrated in Figure 1. In the following sections we will address the problem of chemical shift baselines, QM related errors, combinatoric complexity, conformational sampling and template matching errors. Finally, we will demonstrate the utility of this approach by calculating the error values for the ^1H , ^{15}N , and $^{13}\text{C}^{\alpha}$ atoms in the peptide bonds of 6 model proteins. A significant error value associated with the peptide protons involved in hydrogen bonding with backbone oxygens and water molecules was found in all trajectories. These error values were much greater than the maximum error estimates for the methodology.

2. Level of QM Theory and Chemical Shift Baselines

Magnetic resonance shielding tensors (MRST) for ^1H , ^{13}C and ^{15}N atoms were calculated with the B3LYP [40-48] functional using the gauge independent atomic orbital (GIAO) approach [49]. Reference values for chemical shifts for different basis sets and estimates of the associated errors were obtained in vacuo and in the polarizable continuum model (PCM) with water as the solvent [50]. All QM calculations were carried out using the Gaussian G09 suite of programs [51]. Regression analysis on observed values from organic compounds obtained from the Spectral Database for Organic Compounds (SDBS) [52] and the database maintained by the University of Wisconsin Department of Chemistry (WISC) [53] was employed to calculate reference values for ^1H and ^{13}C chemical shifts. A similar process using data from Dokaklik et al [54] was employed for organic compounds in CDCL3 and DMSO to calculate reference values for ^{15}N chemical shifts. The 6-311+G(2d,p) basis set was ultimately selected for the studies presented in this paper. The details associated with these reference studies including an analysis of the estimated errors are presented in Section S1 of the Supplement. The reference values adopted for ^1H , ^{13}C and ^{15}N were 32.023, 182.629 and 233.133, respectively. The respective error estimates for ^1H , ^{15}N and ^{13}C values with the selected basis set were 3.2%, 1.5% and 1.9%. The regression data for the carbon and hydrogen atoms are shown in Figure 2A and 2B. A comparison of the calculated and observed values for ^1H and ^{13}C using the combined SDBS and WISC databases is shown in Figures 2C and 2D.

3. Defining the size of regions of interest

Chemical shifts for the atoms in peptide bonds must be calculated from protein fragments because QM calculations on proteins with more than 30-40 residues are not feasible from a computational standpoint [55]. Fortunately, magnetic interactions have a limited range. Bonded effects such as spin correlations drop to zero at distances greater than three chemical bonds while the spatial interactions seen in the Nuclear Overhauser effect are negligible at distances greater than 6 Å [56]. It has been shown that the capped peptide sequence Ac-GLY-X-GLY-Me where X is an arbitrary amino acid residue is sufficient to calculate

accurate $^{13}\text{C}^{\alpha}$ values [37]. It has also been shown that the fragmentation approach can be extended to proteins of arbitrary size [55]. This led us to conduct benchmark studies to determine the optimal fragment size for calculating chemical shifts for the $^{13}\text{C}^{\alpha}$, ^1H , and ^{15}N atoms in peptide bonds. For descriptive purposes, the optimal fragment will be referred to as the region of interest (ROI). An ROI has two components: (1) a capped chain of peptide backbone atoms containing the target atoms for chemical shift calculations; and (2) all non-backbone and non-bonded atoms that are close enough to influence the chemical shifts of these atoms. The second component also includes the atoms that are necessary for completing realistic chemical structures.

Our benchmarking studies began with an optimization of the Ac-GLY-GLY-GLY-Me sequence. The target atoms for chemical shift calculations were the $^{13}\text{C}^{\alpha}$, ^1H , and ^{15}N atoms of the peptide bond of the middle residue. An important feature of this sequence was the presence of a peptide bond above and below the bond of interest. We introduced the terms NDOME and ODOM to describe the zones containing the non-backbone and non-bonded atoms that were close enough to influence the chemical shifts of the target atoms. NDOME referred to a hemisphere centered on the hydrogen atom of the target peptide bond in which the plane bisecting the full sphere was normal to the vector connecting the hydrogen atom to the nitrogen atom. ODOM was defined in an analogous fashion using the oxygen atom of the target peptide bond as the center of the hemisphere with the normalization vector defined by the oxygen atom and its carbonyl carbon. These relationships are illustrated in Figure 3. Since aliphatic atoms that are more than three chemical bonds away from target atoms have little effect on magnetic properties, our interest was focused on the polar and aromatic ring atoms within the hemisphere. Aromatic ring atoms were also expected to influence magnetic properties at a longer range than polar atoms [56].

The cutoff radius for magnetic influence in the NDOME was determined by walking a set of small molecule probes down the trajectory of the vector defined by the peptide nitrogen and hydrogen. A similar process was carried out for the ODOM using the trajectory of the vector defined by the peptide oxygen and its carbonyl carbon. The small molecule probes were water, N-methyl-acetamide, benzene and methane. The stepping began at a distance of 10 Å and continued until the van der Waals limit was reached. Details of the stepping process are presented in section S2 of the Supplement. We concluded that the optimal radius for polar atoms in the NDOME was 2.5 Å. The optimal distance for the aromatic atoms was 5.0 Å. A methane probe was stepped in the NDOME to confirm that aliphatic atoms had negligible influence. The stepping process also showed that the ODOM patterns only contributed a small fraction (on the order of 1-5%) to the chemical shift. In the baseline studies described in section 7, the ensemble averages were not affected by pattern atoms more than 3.9 Å from the peptide oxygen. A single cutoff radius of 3.9 Å was subsequently adopted for the ODOM to simplify computation.

4. Model proteins and molecular dynamics simulations

The 6 proteins listed in Table 1 were chosen as model proteins for chemical shift calculations. These proteins represented a broad cross-section of structural motifs and amino

acid distributions. Experimental chemical shift data was available for the ^1H , $^{13}\text{C}^\alpha$ and ^{15}N atoms of each protein through the BMRB or from experimental collaborators [26,57].

MD simulations were carried out on each of these proteins using the GPU accelerated AMBER version 14 with the ff14sb force field [58-62]. Each simulation was run for 100 ns with coordinates saved at picosecond intervals. All protein structures were solvated to form an octahedral TIP3P water box that extended 12 Å beyond the protein. Each system was neutralized with Na^+ or Cl^- ions as needed. Each system underwent two rounds of minimization with the first round minimizing the solvent and the second round minimizing the full system. A 100 ps constant volume equilibration was performed with weak positional restraints on the protein during which the temperature was warmed from 0 K to 300 K. Simulations were performed using the particle mesh Ewald (PME) method with a non-bonded cutoff of 10 Å. For production runs, pressure periodic boundaries and isotropic position scaling were used to maintain a pressure of 1 atm, and Langevin dynamics were used for temperature control.

The MD simulation for ubiquitin was repeated 20 times to check for consistency. Each of the first 10 runs was started from the crystal structure of 1UBQ with a different random seed. The second 10 runs were started from the NMR models of ubiquitin extracted from 1D3Z.

5. ROI Generation, Patterns and Conformational Matching

A Java program called NmrCore was developed for generating ROI's from the PDB files extracted from the trajectories of the MD simulations [63]. Each ROI was centered on a capped peptide chain that was four residues in length. The target peptide was always in the third position and all four of the sidechains attached to the alpha carbons were retained. This was similar to the Ac-X-X-X-Me sequence with the Ac cap replaced by the alpha carbon of the first residue and its sidechain. All chains contained three complete peptide bonds with the target peptide in the middle position. If the chain started or ended with a proline, it was extended until a non-proline residue was encountered. All heavy polar and aromatic ring atoms within 5.0 Å of the origin of the NDOME or ODOME were identified. If any of these atoms belonged to water molecules, complete water molecules were inserted by referencing the associated PDB file. A similar process was applied to atoms belonging to the protein structure subject to the following rules: (1) Peptide chains not connected to the four residue chain were capped; (2) No chain could be less than three residues in length; (3) Chains starting or ending with proline residues were extended; (4) All chains with gaps of three residues or less were merged into a single chain; and (5) residues with no polar or aromatic sidechain atoms within the 5.0 Å limit were converted to glycines. The result of this process was coverage of the target peptide in the NDOME or ODOME to an effective radius of 7-10 Å. This insured a realistic magnetic environment for the $^{13}\text{C}^\alpha$, ^1H , and ^{15}N atoms of the target peptides. The NmrCore program also generated the input files required for the Gaussian G09 suite of programs [51]. All chemical shifts were calculated using the B3LYP functional with a 6-311+G(2d,p) basis set.

5.1. Combinatoric Complexity

It was not feasible from a computational standpoint to calculate the NDOME and ODOM chemical shifts for every residue in every frame of these MD simulation. A protein with 100 residues and 100,000 simulation frames would generate 20 million ROI's. This led us to develop a template matching approach using a representative database of conformations. Each conformer was characterized by a pattern of letters related to the polar and aromatic atoms within the NDOME or ODOM cutoff radius. Each polar atom was assigned a letter code based on its atom type as defined in the CHARMM22 force field [61]. A single letter was assigned to all aromatic ring atoms. The letter assignments are provided in the Supplement in table S3.1. The assigned letters were always written in the same order and atoms with the same letter assignment were sorted in order based on ascending distance from the base atom of the NDOME or ODOM. Patterns were then clustered into residue specific groups using the sequence of letters as a lookup key. Patterns were only considered for matching if they involved the same residue and the same number and types of atoms. Three examples of NDOME patterns include ILE:OC, GLY:W and SER:RRRR. The first pattern involves the interaction of the peptide proton of isoleucine with a remote peptide oxygen and its carbonyl carbon. The second pattern involves the interaction of the peptide proton of glycine with the oxygen of a water molecule. The third pattern involves the interaction of the peptide proton of a serine with four atoms of a remote aromatic ring.

The distribution of residue specific NDOME patterns extracted from the molecular dynamics simulations of the 6 model proteins is shown in Figure 4. The total number of frames was 600,000 with a pattern count of 41,887,942. The total number of unique patterns, however, was only 1189. The figure also shows that the first 25 patterns rank ordered by frequency were sufficient to cover 50% of the conformations. Coverage could be extended to 95% with only 200 patterns. The patterns with the highest frequency were associated with the hydrogen bonds in alpha helices and beta sheets and the interactions between coil regions and water molecules.

A list of the 100 NDOME patterns with the highest frequency is provided in section S3 of the Supplement.

5.2. Comparison Metric

Template matching was confined to NDOME and ODOM conformations with identical residues and patterns. Identical patterns implied that each conformer had the same number and types of atoms. A series of distance and angular measurements were explored involving the pattern atoms and the peptide backbone atoms to find a satisfactory metric for comparing templates. These metrics were tested against sets of patterns with calculated chemical shifts. In each test, the distribution of chemical shifts as a function of the metric was examined. The object was to find a metric that produced a Gaussian distribution of chemical shifts with a zero mean as the difference in the metric approached zero. The metric that emerged with the best performance was based on the distance measurements shown in Figure 5. For each atom (X) in the NDOME or ODOM, the distance to the origin of the hemisphere is determined along with the distances to the nitrogen and oxygen atoms in the peptide bond above and below the target peptide. Conformations were compared by calculating the average absolute

difference between these distance measurements for all atoms in the pattern. This metric was called the matching parameter (MP). Measurement comparisons always involved corresponding atoms and distances in corresponding residues and patterns. A zero (Z) pattern was also introduced to handle the case where no polar or ring atoms were within the cutoff radii of the NDOME or ODOM. In this situation, the peptide hydrogen or oxygen took the place of the pattern atom.

Conformers that were identical had an MP of 0.000 Å. Conformers that were near neighbors had chemical shift errors that were linear functions of the MP value. As the MP decreased, these errors approached Gaussian distributions with mean values of zero. Conformers that were not near neighbors had large MP values and chemical shift errors that approached the values seen in random pairwise selections from the pattern space.

6. NmrSelect and the database of conformations

A program called NmrSelect was developed for selecting the smallest subset of conformations that covers the conformational space represented in an MD simulation for a given matching parameter (MP) threshold. A selected conformation represents, or “covers”, another conformation if the MP value of the two conformations is below the specified threshold. For small thresholds, these conformers will exhibit nearly identical chemical shifts, and the calculated value for one can be substituted for the other.

NmrSelect represents the selection problem as an instance of the dominating set problem. For each pattern, the ROI conformations of the MD simulation are nodes in a graph G that are connected by an edge if and only if the MP value between the frames is less than the specified threshold. The goal of the dominating set problem is to find the smallest subset of nodes D such that every node of G is either in D or is adjacent to a node in D . Although this problem is NP-hard, a simple greedy algorithm provides a logarithmic approximation ratio [64]. The greedy algorithm iteratively selects the node with the highest degree, removes this node and its neighbors from the graph, and adds this node to the dominating set. In NmrSelect high degree nodes correspond to ROI conformations that are representative of frequently sampled conformations.

NmrSelect can also take as input a database of “preselected” ROI conformers. These are conformers that have already had their chemical shifts calculated. Any conformers in the provided MD simulation that are already covered by this set, meaning the MP value is less than the specified threshold, are not included in the dominating set calculation so that only conformations that provide new chemical shift information are selected. The output of NmrSelect is a rank ordered list of ROI conformations from the provided MD simulation data where the most informative conformations, those that cover the greatest percentage of the simulation, are listed first.

Using NmrSelect with an error threshold of 0.20 Å, a representative database of conformers was constructed from the MD simulation trajectories of the 6 model proteins. The chemical shifts for the $^{13}\text{C}^\alpha$, ^1H , and ^{15}N atoms of the target residues in each conformer were calculated using the B3LYP functional and the 6-311+G(2d,p) basis set. Separate

contributions for the NDOME and ODOMES components were obtained by processing each conformer with and without the ODOMES atoms. The NDOME contribution was the difference between the chemical shifts from the complete conformer and the corresponding shifts without the ODOMES atoms. The ODOMES contribution was obtained by subtracting the NDOME contribution from the contribution for the complete conformer. The results for the NDOME and ODOMES contributions along with the distances characterizing these conformations were stored in a database of conformers. The results were indexed by residue and pattern specific keys. Template matching involved extracting the NDOME and ODOMES patterns for each residue and frame of an MD trajectory and finding the corresponding pattern in the conformer database with the smallest value for the matching parameter.

The database of conformers was used for template matching in the MD simulations of the 6 model proteins. The conformations associated with the NDOME and the ODOMES for each residue in each frame of these simulations were extracted. These conformations were compared to conformers in the database having the same residue and pattern. The chemical shifts for the NDOME and ODOMES conformers with the lowest MP values were assigned to the target atoms of the trajectory conformations. The NDOME and ODOMES contributions were added to obtain the chemical shifts for each residue. The chemical shifts for the full trajectory were calculated by taking the ensemble averages. The average matching parameter value for each residue over the complete trajectory was also determined. This average was called the conformational matching error (CME). The chemical shift error (CSE) was calculated by taking the average of the difference in parts per million (PPM) between the experimentally observed and calculated chemical shifts for each residue over the complete trajectory.

The number of conformers required in the conformer database to obtain low conformational matching errors was determined by a series of bootstrapping operations. In each cycle, a value was assigned to the MP parameter in NmrSelect and a level of coverage was selected. The list of conformers was extracted for chemical shift calculations. When processing was complete, the conformers were added to the conformer database and the chemical shifts for the trajectories were recalculated. The CME and the CSE were examined on each cycle. If the CSE improved, the coverage by NmrSelect was increased by decreasing the MP parameter and the process was repeated. The process terminated when no further improvements in the CSE were obtained. The final result was a conformer database with 169,499 entries that provided 99% coverage for the 6 model proteins using an MP parameter of 0.20 Å. This implied that the average matching error would be in the 0.10-0.20 Å range.

7. Conformational Matching Errors

When the conformer database was fully populated, studies were undertaken to quantify the relationship between the CSE, the CME and the MP. This was important, because imperfect matches could introduce additional errors into chemical shift measurements. This information was also useful for optimizing the parameters used by NmrSelect. The first studies involved pairwise comparison of conformations from the fully populated database using the approach employed to evaluate comparison metrics. A representative sample was extracted from the conformer database for each common pattern. The MP and the CSE for

all target atoms were determined on a pairwise basis against all other members of that pattern space. CSE distributions were examined as a function of the MP value to determine the point where the CSE approached a Gaussian distribution with a mean of zero. A plot of the CSE as a function of the MP parameter is shown in Figure 6. The plot involved the OC pattern for glycine and the chemical shift for the ^1H atom. The upper panel shows a scatter plot of the CSE over the MP range from 0.000 – 0.180 Å. In this sample, 100 representative members of the OC pattern space underwent pairwise comparison with 2,022 OC patterns for glycine from the conformer database. The lower panel shows that the CSE distribution approaches a Gaussian distribution with a zero mean as the MP decreases. This implies that ensemble averages of chemical shifts for individual conformers will converge to the correct values if the average MP values (CME for the ensemble) are below a critical level.

A closer view of the relationship between the MP and the CSE for four common patterns is shown in Figures 7A-D. In each panel, the best Gaussian fits for the CSE associated with the ^1H , $^{13}\text{C}^\alpha$ and ^{15}N atoms is presented as a function of four MP levels. Figure 7A shows the OC pattern which was the most common pattern in the database. It occurred in alpha helices, beta sheets and hairpin turns. In this case the critical MP level for a Gaussian distribution with zero mean was 0.150 Å for ^1H and ^{15}N , but 0.075-0.100 Å for $^{13}\text{C}^\alpha$. A similar result for the W pattern is shown in Figure 7B. The W pattern referred to the oxygen atom in a single water molecule which was the second most common pattern. Again the critical values for the ^1H and ^{15}N atoms were higher than for the $^{13}\text{C}^\alpha$ atoms. The pattern involving 9 ring atoms from a tryptophan sidechain is shown in Figure 7C. Here the critical value lies in the 0.350 Å range with little difference between the atom types. Finally, the Z pattern is displayed in Figure 7D. This pattern was the least robust of all the patterns requiring MP's of 0.050 Å or less to achieve a symmetrical Gaussian distribution for $^{13}\text{C}^\alpha$ and ^{15}N atoms and questionable results for ^1H values. Fortunately, this was also a rare pattern. In all other cases, if the MP was below a critical level, the error over the ensemble average should be negligible.

We also determined the error associated with removing rare conformations from the conformer database. This was accomplished by reprocessing the database with NmrSelect using a progressively lower level of coverage. This resulted in a progressive loss of the conformations with the lowest frequency. For each setting, we calculated the chemical shift error for a 100,000 frame simulation of ubiquitin using the 99% coverage level as the baseline. The results are shown in Figure 8.

When the coverage dropped to 90%, an error of ~0.5% was introduced. At 80% coverage, this error rose to ~1.0%. This error never rose above 4%, however, as long as the conformations with the highest frequency remained in the database.

Finally, the average NDOME contribution for the target atoms in the conformer database was compared to the contribution from the ODOM. The ODOM contribution averaged 3.72% for ^{15}N , 1.04% for ^1H and 0.04% for $^{13}\text{C}^\alpha$. This meant that the ODOM could tolerate much larger matching errors than the NDOME without affecting ensemble chemical shifts.

7. Baseline studies for ubiquitin

Chemical shifts were calculated for the target peptide atoms in 20 different ubiquitin trajectories. The chemical shifts for individual residue:pattern combinations (keys) were assigned from the conformer database which had 99% coverage for ubiquitin at MP = 0.20 Å. Each trajectory was 100 ns in length sampled at 1 ps intervals (100,000 frames/trajectory). The first 10 trajectories were started from different random seeds using the 1UBQ crystal structure. The second 10 trajectories were started from the 10 different NMR models from 1D3Z. The average values for the chemical shifts and the associated CSE and CME parameters for each trajectory are shown in Table 2. The standard deviations show that there was little variation in the parameter averages for the 20 trajectories. This is confirmed by the superposition of the chemical shifts for the $^{13}\text{C}^\alpha$, ^1H and ^{15}N atoms from the 20 trajectories shown in Figure 9. Examination of the trajectories with VMD [65] on a frame by frame basis showed that the residue positions with differences were usually associated with random coil secondary structure. These differences were usually attributable to differences in rotameric conformations. In Figure 10, the mean calculated chemical shifts for the $^{13}\text{C}^\alpha$, ^1H and ^{15}N atoms over the 20 trajectories were compared to experimental observations (BMRB 6457). The mean difference for $^{13}\text{C}^\alpha$ was -0.036 ppm. The error percentage defined as the average of the absolute difference between observed and calculated values divided by the observed value was 2.1%. The average difference for ^1H was -1.388 ppm and the corresponding error percentage was 16.8%. The average difference for ^{15}N was -2.680 ppm and the corresponding error percentage was 3.7%.

The error percentage for ^1H was much larger than the error percentage for the other target atoms. The ^1H error also appeared to have a large systematic component. Inspection of the trajectories with VMD showed that most of the non-systematic component of the ^1H error could be attributed to differences in local rotameric conformations. It should be noted that the $^{13}\text{C}^\alpha$ values are highly dependent on the bond lengths and angles in the immediate vicinity of the target carbon and relatively insensitive to the atoms in the NDOM and ODOM.³⁹ The ^1H values, on the other hand, are highly influenced by the atoms in the NDOM with interactions that are largely dependent on the electrical environment. The ^{15}N values are influenced by both factors because of the polarizability of the nitrogen atom.

8. Chemical shift calculations for the 6 model proteins

The observations for ubiquitin were confirmed for all 6 of the model proteins. In all cases, the coverage in the conformer database was above the 99% level. The average errors for the chemical shifts of the target atoms are presented in Table 3. The error associated with each residue was calculated as the absolute difference between calculated and observed values divided by the observed value. It was expressed as a percentage. The RMSD error expressed in PPM was also calculated. The values in the table were the averages over all the residues in each protein. For each protein, errors of similar magnitude were recorded. The error rate for the ^{15}N and $^{13}\text{C}^\alpha$ atoms was in the 2.5-4.0% range. This was double the maximum error estimate for the methodology if one considers the worst case scenario for errors in experimental observations, errors related to the QM level of theory and conformational matching errors. See section S1 in the Supplement for additional details related to errors in

experimental observations and errors related to the level of QM theory. For the ^1H chemical shifts, the error rate was in the 18% range. This was three times the maximum error estimate for the methodology.

A residue by residue comparison of the ^{13}C , ^1H and ^{15}N chemical shifts for 3OBL are shown in Figure 11. The complete plots for all 6 model proteins are presented in section S4 of the Supplement. In all 6 proteins, the error for the ^{13}C and ^{15}N atoms is in the 2.5-4.0% range and is manifest primarily in the coil regions. The ^1H error by contrast is in the 16% range with a strong systematic component. Non-systematic components of the ^1H error also tend to be associated with coil regions and appear to be a function of local rotameric conformations. The systematic components of the ^1H error are largely associated with hydrogen bonds involving water molecules or peptide bond oxygens (alpha helices and beta sheets). Analysis of interatomic distances suggests that the distance between the peptide protons and the hydrogen bond acceptors is too large. This is confirmed by a series of MD simulations carried out using dielectric constants varying from 0.5 to 1.5. As the electrostatic interaction between the proton and the hydrogen bond acceptor increased, the average interatomic distance decreased and the ^1H error in chemical shifts was significantly reduced. A table with the results of the dielectric manipulations is presented in Section S5 of the Supplement. In the case of 3OBL, it is interesting to note that the ^1H chemical shifts associated with positions 26 and 93 have no significant error. Both of these NDOME regions involve the relationship between a peptide hydrogen that is oriented at 90° to the ring of a tryptophan residue. This suggests that the ^1H error may also have a pattern specific component. The analysis of patterns common to all 6 model proteins presented in section S5 of the Supplement supports this assumption.

9. Application of Method to New Proteins

The ability to efficiently calculate chemical shifts for a new simulation depends on the number of novel template conformations that require QM calculations. This will be highly variable depending on the nature of the protein simulation and the extensiveness of the existing database. However, we have observed that a significant number of ROI conformations are already covered by our existing database. For example, the 1IGD protein in the model set was processed after the other 5 proteins. The initial coverage of 1IGD from the processed proteins was almost 75%. On the flip side, even if a simulation of a protein has already been processed to full coverage, a new simulation is likely to explore novel ROI conformations. For the ubiquitin simulations we typically saw 95-98% initial coverage of the conformations when processing a new simulation. If a different force field is used, novel conformations are more likely and we observed initial coverage percentages of 80-90% depending on the degree of divergence from the force field used to populate the conformer database. Combined with the error analysis shown in Figure 8, this suggests that in many cases a new simulation will be able to be analyzed without further QM calculation at the expense of a 1% error. Of course, this error will drop and the initial coverage will increase as the database of conformers is expanded.

10. Comparison to Empirical Methods

A number of empirical methods for predicting NMR observables have been developed [26-33]. These methods are designed to make predictions based on a single structure that is representative of the conformational ensemble. Although such empirical methods have been used to evaluate force fields [34], the conclusion of this evaluation was that these methods lack the necessary precision to discriminate between most force fields. This lack of precision is likely due to the necessary insensitivity of single-structure empirical methods to local conformational changes [36]. This effect is illustrated in Figure 12 where the behavior of the GLY59 residue of 3OBL is compared to its predicted chemical shift. This residue is located at the transition between a beta sheet and a loop region and the backbone hydrogen bond to GLN85 that forms the beta sheet opens and closes during the course of a 100 ns simulation. These fluctuations, which significantly alter the electrostatic environment of the hydrogen, are reflected in the ^1H chemical shifts calculated using our method (QM) but not in those calculated with an empirical method (SHIFTX2, without SHIFTY homolog information) as can be seen by a comparison with the hydrogen bond distance. Furthermore, in addition to being inverse correlated with the hydrogen bond distance, our calculated chemical shifts are sensitive to other modifications of the electromagnetic environment. For example, from 60-70 nanoseconds the hydrogen bond is open, but instead of being bridged by a water molecule, as it is elsewhere in the simulation, the loop backbone desolvates the bond as GLN59 interacts with the backbone oxygen of GLY77. This different conformational state is clearly represented in the computed ^1H chemical shift.

The illustrative example of Figure 12 is typical of residues that sample different conformational states during the molecular dynamics simulation. Overall, the chemical shifts computed by our method have a larger variance across the simulation compared to empirical calculations (see Supplemental Figure S7.1). The residues with the highest variance in computed chemical shift across the simulation are those in loop regions, as should be expected given the wide range of conformational space sampled by these regions. The sensitivity of the method to conformational change make it an appropriate tool for assessing simulation force fields.

A significant amount of the observed error with our technique appears to be systematic (e.g., the ^1H chemical shift tends to under-predict by $\sim 1.5\text{PPM}$). Consequently, these errors can be compensated for with regression techniques to achieve results comparable to empirical predictors (see section S8 in the Supplement). However, doing so defeats the purpose of calculating the actual chemical shift inherent in the simulation coordinates, although we note that even these empirically modified chemical shifts remain sensitive to conformational changes.

11. Conclusions

Every atom in every MD simulation step has a chemical shift that is solely a function of the coordinates of the atoms in its local neighborhood. If the conformational space is adequately sampled, the ensemble averages of the chemical shifts for these atoms should equal the observed experimental chemical shifts for those atoms. If this is not the case, the coordinates

are not correct. Since the coordinates are a function of the force field, algorithms and parameter sets underlying the MD simulation, the difference between calculated and observed chemical shifts becomes a direct measure of the efficacy of these combinations.

We have shown that it is feasible to obtain the chemical shifts for the ^1H , $^{13}\text{C}^\alpha$ and ^{15}N atoms in every peptide bond in every frame of an MD simulation using a template matching approach and a library of conformers with chemical shifts obtained from *de novo* QM calculations. This approach has been applied to 6 model proteins with 100 ns trajectories sampled at picosecond intervals. In all cases, a large systematic error associated with the ^1H chemical shifts was identified. A smaller systematic error associated with the $^{13}\text{C}^\alpha$ and ^{15}N chemical shifts was also identified. The magnitude of these errors was compared to worst case estimates of the errors associated with the template matching methodology. This included an estimate of the errors associated with the experimental values, the errors associated with the level of QM theory and the errors associated with the template matching algorithm. The ^1H systematic errors were three times greater than the worst case error for the methodology. The $^{13}\text{C}^\alpha$ and ^{15}N errors were two times greater than the worst case scenario.

Chemical shifts for $^{13}\text{C}^\alpha$ atoms are not significantly influenced by the surrounding electrostatic environment [37], while electrostatics are critical for ^1H values. This suggested that the electrostatic components of the force field might be involved in this error. This possibility was explored by varying the dielectric constant in a series of MD simulations. These studies showed that increasing the strength of the interaction between the hydrogen of the peptide bond and the nearest pattern atom in the NDOME decreased the distance between these atoms and significantly lowered the systematic error. Studies are currently underway to evaluate polarizing force fields by comparing the chemical shift errors in MD simulations of the 6 model proteins with the simulations carried out with the AMBER ff14sb force field [66,67].

It is easy to extend this approach to new proteins by confining the analysis to those residues that have adequate coverage in the current conformational database, or by adding new conformations using the NmrSelect utility. The *de novo* nature of the approach means that the calculated chemical shifts are directly determined by the coordinates of the molecular dynamics simulation, without any bias introduced by empirical fitting, and so provide a means to correlate simulation with experimental observables.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Dr. Leonard Koharudin, Dr. Antonio Ferreira, Karl Debiec and Dr. Angela Gronenborn for useful discussion and for providing the NMR chemical shift data for the OAA protein. We are also grateful to Dawei Li for providing valuable feedback on the manuscript. This work was supported in part by grants LM007994 from the National Library of Medicine and R01GM108340 from the National Institute of General Medical Sciences, and the Center for Simulation and Modeling at the University of Pittsburgh. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

1. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol*. 2002; 9:646–52. [PubMed: 12198485]
2. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-level characterization of the structural dynamics of proteins. *Science*. 2010; 330:341–6. [PubMed: 20947758]
3. Mackerell AD Jr. Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem*. 2004; 25:1584–604. [PubMed: 15264253]
4. Jiang W, Hardy DJ, Phillips JC, Mackerell AD Jr, Schulten K, Roux B. High-performance scalable molecular dynamics simulations of a polarizable force field based on classical Drude oscillators in NAMD. *J Phys Chem Lett*. 2011; 2:87–92. [PubMed: 21572567]
5. Piana S, Lindorff-Larsen K, Dirks RM, Salmon JK, Dror RO, Shaw DE. Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations. *PLoS One*. 2012; 7:e39918. [PubMed: 22768169]
6. Macias AT, Mackerell AD Jr. CH/ π interactions involving aromatic amino acids: refinement of the CHARMM tryptophan force field. *J Comput Chem*. 2005; 26:1452–63. [PubMed: 16088926]
7. Guillot B. A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids*. 2002; 101:219–260.
8. Jorgensen WL, Tirado-Rives J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci U S A*. 2005; 102:6665–70. [PubMed: 15870211]
9. Lamoureux G, Harder E, Vorobyov IV, Roux B, MacKerell AD Jr. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem Phys Lett*. 2006; 418:245–249.
10. Yu W, Lopes PE, Roux B, MacKerell AD Jr. Six-site polarizable model of water based on the classical Drude oscillator. *J Chem Phys*. 2013; 138:034508. [PubMed: 23343286]
11. Ponder JW, Case DA. Force Fields for Protein Simulation. *Adv Prot Chem*. 2003; 66:27–85.
12. Best RB, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, Mackerell AD Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput*. 2012; 8:3257–3273. [PubMed: 23341755]
13. Shirts MR, Pitera JW, Swope WC, Pande VS. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J Chem Phys*. 2003; 119:5740.
14. Best RB, Buchete NV, Hummer G. Are current molecular dynamics force fields too helical? *Biophys J*. 2008; 95:L07–9. [PubMed: 18456823]
15. Lange OF, van der Spoel D, de Groot BL. Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. *Biophys J*. 2010; 99:647–55. [PubMed: 20643085]
16. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic validation of protein force fields against experimental data. *PLoS One*. 2012; 7:e32131. [PubMed: 22384157]
17. Cino EA, Choy WY, Karttunen M. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J Chem Theory Comput*. 2012; 8:2725–2740. [PubMed: 22904695]
18. Showalter SA, Brüschweiler R. Validation of molecular dynamics simulation of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *J Chem Theory Comput*. 2007; 3:961–975. [PubMed: 26627416]
19. Maragakis P, Lindorff-Larsen K, Eastwood MP, Dror RO, Klepeis JL, Arkin IT, Jensen MØ, Xu H, Trbovic N, Friesner RA, Palmer AG Iii, Shaw DE. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *J Phys Chem B*. 2008; 112:6155–8. [PubMed: 18311962]
20. Showalter SA, Brüschweiler R. Quantitative molecular ensemble interpretation of NMR dipolar couplings without restraints. *J Am Chem Soc*. 2007; 129:4158–9. [PubMed: 17367145]

21. Markwick PR, Bouvignies G, Blackledge M. Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. *J Am Chem Soc.* 2007; 129:4724–30. [PubMed: 17375925]
22. Wickstrom L, Okur A, Simmerling C. Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. *Biophys J.* 2009; 97:853–6. [PubMed: 19651043]
23. Markwick PR, Showalter SA, Bouvignies G, Brüschweiler R, Blackledge M. Structural dynamics of protein backbone phi angles: extended molecular dynamics simulations versus experimental (3) J scalar couplings. *J Biomol NMR.* 2009; 45:17–21. [PubMed: 19629714]
24. Best RB, Hummer G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B.* 2009; 113:9004–15. [PubMed: 19514729]
25. He X, Wang B, Merz KM Jr. Protein Nmr chemical shift calculations based on the automated fragmentation QM/MM approach. *J Phys Chem B.* 2009; 113:10380–10388. [PubMed: 19575540]
26. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent WR, Yao H, Markley JL. BioMagResBank. *Nucleic Acids Res.* 2008; 36:D402–8. [PubMed: 17984079]
27. Wishart DS, Watson MS, Boyko RF, Sykes BD. Automated ¹H and ¹³C chemical shift prediction using the BioMagResBank. *J Biomol NMR.* 1997; 10:329–36. [PubMed: 9460240]
28. Neal S, Nip AM, Zhang H, Wishart DS. Rapid and accurate calculation of protein ¹H, ¹³C and ¹⁵N chemical shifts. *J Biomol NMR.* 2003; 26:215–40. [PubMed: 12766419]
29. Han B, Liu Y, Ginzinger SW, Wishart DS. SHIFTX2: significantly improved protein chemical shift prediction. *J Biomol NMR.* 2011; 50:43–57. [PubMed: 21448735]
30. Shen Y, Ba A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR.* 2010; 48:13–22. [PubMed: 20628786]
31. Iwadate M, Asakur T, Williamson MP. C alpha and C beta carbon-13 chemical shifts in proteins from an empirical database. *J Biomol NMR.* 1999; 13:199–211. [PubMed: 10212983]
32. Moon S, Case DA. A new model for chemical shifts of amide hydrogens in proteins. *J Biomol NMR.* 2007; 38:139–50. [PubMed: 17457516]
33. Meile J. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J Biomol NMR.* 2003; 26:25–37. [PubMed: 12766400]
34. Beauchamp KA, Lin YS, Das R, Pande VS. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J Chem Theory Comput.* 2012; 8:1409–1414. [PubMed: 22754404]
35. Li DW, Brüschweiler R. Iterative Optimization of Molecular Mechanics Force Fields from NMR Data of Full-Length Proteins. *J Chem Theor Comput.* 2011; 7:1773–1782.
36. Sumowski CV, Hanni M, Schweizer S, Ochsenfeld C. Sensitivity of Ab-Initio vs. Empirical Methods in Computing Structural Effects on NMR Chemical Shifts for the Example of Peptides. *J Chem Theory Comput.* 2014; 10:122–133. [PubMed: 26579896]
37. Vila JA, Scheraga HA. Assessing the accuracy of protein structures by quantum mechanical computations of ¹³C^α chemical shifts. *Acc Chem Res.* 2009; 42:1545–1553. [PubMed: 19572703]
38. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J Phys Chem.* 1992; 96:6472–6484.
39. Vila JA, Arnautova YA, Martin OA, Scheraga HA. Quantum mechanics-derived ¹³C^α chemical shift server (CheShift) for protein structure validation. *Proc Natl Acad Sci U S A.* 2009; 106:16972–7. [PubMed: 19805131]
40. Hehre, WJ., Radom, L., Schleyer, PV., Pople, J. AB INITIO Molecular Orbital Theory. Wiley; New York: 1986.
41. Cramer, CJ. Essentials of Computational Chemistry. Vol. 189. John Wiley & Sons, Ltd; Chichester: 2002.
42. Becke AD. Density functional exchange-energy approximation with correct asymptotic-behavior. *Phys Rev A.* 1988; 38:3098–100.

43. Becke AD. Density functional thermochemistry III The role of exact change. *J Chem Phys.* 1993; 98:5648–5652.
44. Hohenberg P, Kohn W. Inhomogenous electron gas. *Phys Rev.* 1964; 136:B864–B871.
45. Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. *Phys Rev.* 1964; 140:A1133–A1138.
46. Lee C, Yang W, Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys Rev B.* 1988; 37:785–89.
47. Miehlich B, Savin A, Stoll H, Preuss H. Results obtained with the correlation-energy density functionals of Becke and Lee, Yang and Parr. *Chem Phys Lett.* 1989; 157:200–06.
48. St-Amant, A. The Study of Biomolecules by Density Functional Methods. In: Lipkowitz, KB., Boyd, DB., editors. *Reviews in Computational Chemistry*. Vol. 7. VCH Publishers; New York: 1996.
49. Cheeseman JR, Trucks GW, Keith TA, Frisch MJ. A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J Chem Phys.* 1996; 104:5497–5510.
50. Scalmani G, Frisch MJ. Continuous surface charge polarizable continuum models of solvation. I. General formalism. *J Chem Phys.* 2010; 132:114110. [PubMed: 20331284]
51. Frisch, MJ., Trucks, GW., Schlegel, HB., Scuseria, GE., Robb, MA., Cheeseman, JR., Montgomery, JA., Jr, Vreven, T., Kudin, KN., Burant, JC., Millam, JM., Iyengar, SS., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, GA., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, JE., Hratchian, HP., Cross, JB., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, RE., Yazyev, O., Austin, AJ., Cammi, R., Pomelli, C., Ochterski, JW., Ayala, PY., Morokuma, K., Voth, G., Salvador, P., Dannenberg, JJ., Zakrzewski, VG., Dapprich, S., Daniels, AD., Strain, MC., Farkas, O., Malick, DK., Rabuck, AD., Raghavachari, K., Foresman, JB., Ortiz, JV., Cui, Q., Baboul, AG., Clifford, S., Cioslowski, J., Stefanov, BB., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, RL., Fox, DJ., Keith, T., Al-Laham, MA., Peng, CY., Nanayakkara, A., Challacombe, M., Gill, PM., Johnson, B., Chen, W., Wong, MW., Gonzalez, C., Pople, JA. Gaussian, Inc; Wallingford CT: 2004.
52. http://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi
53. <http://www.chem.wisc.edu/areas/reich/chem605/index.htm>
54. Dokalik A, Kalchauer H, Mikenda W, Schweng G. NMR spectra of nitrogen-containing compounds. Correlations between experimental and GIAO calculated data. *Magn Reson Chem.* 1999; 37:895–902.
55. Frank A, Möller HM, Exner TE. Toward the quantum chemical calculation of NMR chemical shifts of proteins. 2. Level of theory, basis set and solvents model dependence. *J Chem Theory Comput.* 2012; 8:1480–1492. [PubMed: 26596758]
56. Keeler, J. *Understanding NMR Spectroscopy*. 2. John Wiley and Sons, Ltd; New York: 2010.
57. Gronenborn, AM. Personal communication. Department of Structural Biology, University of Pittsburgh; 2012.
58. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods R. The Amber biomolecular simulation programs. *J Computat Chem.* 2005; 26:1668–1688.
59. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. *J Comput Chem.* 2005; 26:1781–1802. [PubMed: 16222654]
60. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, Hess B, Lindahl E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics.* 2013; 29:845–54. [PubMed: 23407358]
61. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL, Evanseck J, Field MJ, Fischer S, Gao SJ, Guo H, Ha S, Joseph D, Kuchnir L, Kucsera K, Lau FTK, Mattos C, Michnick CS, Ngo T, Nguyen DT, Prodhom B, Reiher IWE, Roux B, Schlenkrich M, Smith J, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kucsera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B.* 1998; 102:3586–3616. [PubMed: 24889800]

62. Case, DA., Babin, V., Berryman, JT., Betz, RM., Cai, Q., Cerutti, DS., Cheatham, TE., III, Darden, TA., Duke, RE., Gohlke, H., Goetz, AW., Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, TS., LeGrand, S., Luchko, T., Luo, R., Madej, B., Merz, KM., Paesani, F., Roe, DR., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, CL., Smith, W., Swails, J., Walker, RC., Wang, J., Wolf, RM., Wu, X., Kollman, PA. AMBER 14. University of California; San Francisco: 2014.
63. Gosling J, Joy W, Steele G, Bracha G, Buckley A. The Java® Language Specification (Java SE 8 ed). 2014
64. <http://dx.doi.org/10.1016/j.ic.2008.07.003>
65. Humphrey W, Dalke A, Schulten K. VMD - Visual Molecular Dynamics. J Molec Graphics. 1996; 14:33–38.
66. Lopes PEM, Huang J, Shim J, Luo Y, Li H, Roux B, MacKerrell AD. Polarizable force field for peptides and proteins based on the classical Drude oscillator. J Chem Theory Comput. 2013; 9:5430–5449. [PubMed: 24459460]
67. Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder JW, Ren P. Polarizable atomic multipole-based AMOEBA force field for Proteins. J Chem Theory Comput. 2013; 9:4046–4063. [PubMed: 24163642]

Fig 1.

The general workflow of the method. Protein dynamics are sampled with molecular dynamics. Regions of interest are defined around each backbone N and O atom for each frame that encapsulate the dominant magnetic interactions. ROIs that are conformationally representative of the protein dynamics are selected and chemical shifts are calculated and stored in a template database. An ensemble value is approximated by identifying the closest matching ROIs with calculated chemical shifts resulting in a calculated value with a reasonable empirical error.

Fig 2.

(A) Linear regression of MRST vs chemical shift for carbon with the 6-311+G(2d,p) basis set. The MRST reference for ^{13}C is 182.629. (B) Linear regression of MRST vs chemical shift for hydrogen with the 6-311+G(2d,p) basis set. The MRST reference for ^1H is 32.023. (C) Superposition of calculated and observed chemical shifts for ^{13}C after bias correction. (D) Superposition of calculated and observed chemical shifts for ^1H after bias correction.

Fig 3.

The NDOME and the ODOME are hemispheres centered on the hydrogen and oxygen atoms of a target peptide bond. The plane bisecting the full sphere of the NDOME is normal to the vector connecting the nitrogen and hydrogen atoms. The plane bisecting the ODOME is normal to the vector connecting the oxygen and carbonyl carbon atoms. Polar and aromatic atoms within a limiting radius are close enough to influence the chemical shifts of the ^{13}C , ^1H , and ^{15}N atoms in the peptide bond.

Fig 4.

Cumulative coverage for NDOME patterns in 6 model proteins. A total of 1189 unique patterns were identified in a sample of 41,887,942 patterns drawn from the MD trajectories of these proteins. A smaller subset of these patterns accounted for the bulk of the coverage. Half the coverage was accounted for by only 25 patterns. The coverage reached the 95% level with only 200 patterns. The patterns with the highest frequency were associated with the hydrogen bonding sites in alpha helices or beta sheets. Next in frequency were the interactions between coil regions and water molecules.

Fig 5.

Distances used in the metric for comparing conformations. X represents a polar or aromatic ring atom in the NDOME or ODOMES. The NDOME and ODOMES patterns for a conformer are formed from an ordered list of all the atoms within the cutoff limits of their respective hemispheres. NDOME and ODOMES patterns are compared by computing the average absolute difference between corresponding distances for all atoms in the pattern. In all comparisons, the patterns must be a perfect match in terms of residue type as well as number and type of polar or aromatic ring atoms. All NDOME and ODOMES pattern atoms have distance measurements related to the polar atoms in the peptide bond proximal (N₁,O₁) and distal (N₂,O₂) to the target peptide bond as well as to the hydrogen or oxygen (H₀,O₀) of the target peptide.

Fig 6.

Relationship of the chemical shift error (CSE) to the matching parameter (MP) for the OC pattern in glycine. A random sample of 100 conformations from the database was matched against every glycine OC pattern (N=2022). The difference in chemical shift for ^1H was plotted as a function of the MP parameter (upper panel). The MP parameter varied from 0.047 to 1.800 when identity matches were excluded. As the MP decreased, the error distribution approached a Gaussian distribution with a zero mean and a decreasing standard deviation (lower panel). This implied that the chemical shift would converge to the correct value during ensemble averaging if the MP were below a critical level. Similar results were obtained for all common residue and pattern combinations for all three chemical shifts.

Fig 7. Distribution of chemical shift errors for ^1H , $^{13}\text{C}^\alpha$ and ^{15}N for different patterns as a function of the MP parameter. Black arrows indicate the zero mean. (A) In the OC pattern, the ^1H and ^{15}N errors maintain a zero mean to an MP of 0.150 Å. A negative bias appears in the $^{13}\text{C}^\alpha$ distribution as the MP exceeds 0.075 Å. (B) The W pattern shows similar findings, but the negative bias for $^{13}\text{C}^\alpha$ is less pronounced. (C) The RRRRRRRRRR pattern (R9) related to an orthogonal tryptophan side chain maintains a zero mean to an MP level of 0.350 Å for all chemical shifts. (D) The Z pattern picks up a negative bias at MP levels in the 0.50 to 0.075 Å range. Approximately 95% of all patterns fall into the OC, W or R (ring) category.

Fig 8.

Error caused by the progressive removal of rare conformations. At the lowest level of coverage, only the conformations with the highest frequency remained. As coverage decreased, calculated chemical shifts picked up an incremental error that reached the 1% level for coverage in the 80% range. This error never rose above 4%, however, as long as the conformations with high frequency remained in the database.

Fig 9.

Superposition of the $^{13}\text{C}^\alpha$, ^1H and ^{15}N chemical shift values for 20 ubiquitin trajectories. The chemical shift values were highly reproducible in regions associated with alpha helix or beta sheet secondary structure. The variability in coil regions was usually associated with differences in rotameric conformations.

Fig 10.

Comparison of the calculated chemical shifts for $^{13}\text{C}\alpha$, ^1H and ^{15}N atoms with experimental observations (BMRB 6457). The chemical shifts were averaged from 20 separate ubiquitin simulations. Each simulation was 100 ns in length and contained 100,000 samples. The differences between calculated and observed values were small for $^{13}\text{C}\alpha$ and ^{15}N atoms. The calculated values for ^1H were consistently low by an average of 1.388 ppm. There appears to be a large systematic component to the ^1H error. There were also localized non-systematic differences that may be related to variations in rotameric conformations.

Fig 11.

Comparison of observed and calculated chemical shifts for 3OBL. The error for the $^{13}\text{C}_\alpha$ and ^{15}N atoms (2.6% and 3.5%) is relatively small and manifest primarily in coil regions. The ^1H error is much larger (15.6%) and contains a significant systematic component. It is also interesting to note that residue positions 26 and 93 which involve the relationship of a peptide proton with a tryptophan ring have no error. This suggests that the ^1H error may also be pattern specific.

Fig 12.

Sensitivity of computed chemical shifts to conformational changes. (Top Panel) The backbone hydrogen bond distance of the GLY59 residue of 3OBL with GLN75 shown with corresponding conformations at each stage of the 100 ns simulation. (Bottom Panel) The computed ^1H chemical shifts using an empirical (SHIFTX2) method and our *de novo* method (QM). Plotted values are the trailing average of the previous 1000 picoseconds.

Table 1

Model Proteins

PDB ID	Name	Class	Architecture	Function	Residues
1ENH	Engrailed homeodomain	Mostly Alpha	Orthogonal bundle	DNA binding	54
1IGD	Protein G	Alpha Beta	Roll	Cell wall component	61
1HIK	Interleukine-4	Mainly Alpha	Up-down bundle	Cytokine activity	129
3OBL	Cyanobacterial lectin	Mostly Beta	Beta barrel	Carbohydrate binding	132
1QZM	ATP dependent protease	Mainly alpha	Orthogonal bundle	ATP dependent protease	94
1UBQ	Ubiquitin	Alpha Beta	Roll	Regulatory protein	76

Table 2

Variation in Calculated Chemical Shifts for 20 Ubiquitin Trajectories

RUN	N-CALC	H-CALC	C-CALC	N-CSE	H-CSE	C-CSE	ND-CME	OD-CME
1	116.375	6.965	56.661	-2.624	-1.386	-0.031	0.085	0.249
2	116.231	6.980	56.696	-2.768	-1.371	0.004	0.091	0.260
3	116.231	6.980	56.696	-2.768	-1.371	0.004	0.091	0.260
4	116.385	6.976	56.646	-2.615	-1.374	-0.047	0.085	0.252
5	116.373	6.954	56.662	-2.627	-1.397	-0.030	0.086	0.256
6	116.306	6.955	56.624	-2.693	-1.395	-0.068	0.086	0.250
7	116.270	6.950	56.686	-2.730	-1.400	-0.007	0.088	0.257
8	116.386	6.958	56.675	-2.613	-1.392	-0.017	0.085	0.250
9	116.328	6.960	56.624	-2.672	-1.390	-0.068	0.085	0.243
10	116.195	6.966	56.718	-2.805	-1.385	0.026	0.088	0.265
11	116.148	6.952	56.689	-2.852	-1.398	-0.004	0.085	0.256
12	116.350	6.964	56.623	-2.649	-1.387	-0.069	0.084	0.248
13	116.373	6.960	56.640	-2.626	-1.390	-0.053	0.086	0.250
14	116.281	6.972	56.592	-2.718	-1.378	-0.100	0.086	0.250
15	116.457	6.969	56.640	-2.543	-1.381	-0.052	0.085	0.246
16	116.337	6.966	56.600	-2.663	-1.385	-0.092	0.086	0.245
17	116.351	6.954	56.671	-2.649	-1.396	-0.021	0.085	0.251
18	116.320	6.958	56.622	-2.679	-1.393	-0.070	0.085	0.246
19	116.320	6.954	56.701	-2.679	-1.397	0.008	0.085	0.247
20	116.372	6.961	56.659	-2.628	-1.389	-0.033	0.085	0.254
MEAN	116.319	6.963	56.656	-2.680	-1.388	-0.036	0.086	0.252
STD	0.075	0.009	0.035	0.075	0.009	0.035	0.002	0.006

Table 3

Error rate and RMSD difference for calculated chemical shifts

PROTEIN	N-ERR%	N-RMSD	H-ERR%	H-RMSD	C-ERR%	C-RMSD
IENH	3.9%	5.795	21.3%	1.848	2.1%	1.568
IHIK	3.1%	4.660	18.4%	1.578	2.4%	1.721
IIGD	4.6%	6.484	17.5%	1.651	2.9%	2.240
3OBL	3.5%	5.418	15.6%	1.558	2.6%	1.914
IQZM	3.8%	5.644	19.0%	1.678	3.1%	2.268
IUBQ	3.7%	5.197	16.6%	1.559	2.1%	1.526
MEAN	3.8%	5.533	18.2%	1.645	2.5%	1.873
STD	0.5%	0.612	02.2%	0.111	0.4%	0.325
IUBQ20	3.7%	5.261	16.8%	1.555	2.1%	1.503

ERR% = average of (abs(calculated-observed)/observed) for each residue x 100
 IUBQ20 = average ERR% and RMSD values for the 20 baseline Ubiquitin trajectories