AM1 Parameters for the Prediction of ^{1}H and ^{13}C NMR Chemical Shifts in Proteins†

Duane E. Williams, Martin B. Peters, Bing Wang, Adrian E. Roitberg, and Kenneth M. Merz, Jr.*

Department of Chemistry, Quantum Theory Project, 2328 New Physics Building, P.O. Box 118435, University of Florida, Gainesville, Florida 32611-8435

Received: March 30, 2009; Revised Manuscript Received: August 23, 2009

The semiempirical quantum mechanical description of NMR chemical shifts has been implemented at the AM1 level with NMR-specific parameters to reproduce experimental ^{1}H and ^{13}C NMR chemical shifts. The methodology adopted here is formally the same as that of the previously published finite perturbation theory GIAO-MNDO-NMR approach [Wang, B.; et al. *J. Chem. Phys.* **2004**, *120*, 24.]. The primary impetus for this parametrization was the accurate capture of chemical environments of atoms in biological systems. Protein-specific parameters were developed on a training set that comprised five globular protein systems with varied secondary structure and a range in size from 46−61 amino acid residues. A separate set of parameters was developed using a training set of small organic compounds with an emphasis on functional groups that are relevant to biological studies. Our approach can be employed using semiempirical (AM1) geometries and can be executed at a fraction of the cost of ab initio and DFT methods, thus providing an attractive option for the computational NMR studies of much larger protein systems. Analysis carried out on 3340 ^{1}H and 2233 ^{13}C chemical shifts for protein systems shows significant improvement over the standard AM1 parameters. Using ^{1}H and ^{13}C specific parameters, the rms errors are from 1.05 and 21.28 ppm to 0.62 and 4.83 ppm for hydrogen and carbon, respectively.

Introduction

The ability to predict NMR chemical shifts for protein systems (ranging in size from a few hundred atoms to many thousands) routinely, accurately, and quickly can aid in structure elucidation, give insight into the binding modes of ligands in proteins, and add valuable information on dynamics of these systems. The versatility and accuracy of modern quantum mechanical (QM) methods make them a preferred approach to predicting chemical shifts for these systems. Ab initio and density functional theory (DFT) methods perform very well in reproducing experimental chemical shifts for small molecules,[1] and significant progress has been made to improve the speed of these calculations to make them useful for larger systems.[1–9] Still semiempirical methods offer a far less computationally expensive route to attaining these quantities for very large systems. Here we present a semiempirical QM methodology that scales well enough to predict NMR chemical shifts for large protein systems, and is specifically parametrized for that purpose.

Not only are the state-of-the-art empirical methods of NMR chemical shift prediction fast enough to be used for high throughput applications, but they also have good accuracy for the prediction of chemical shifts of many carbon and hydrogen atoms found in protein systems.[10] One drawback of empirical methods is that they generally take advantage of atom typing and cannot easily be applied to the large variety of organic molecules that are of interest as ligands bound to proteins in biochemical and medicinal studies. Therefore, these approaches are generally not amenable to proteins with nonstandard amino acids or protein−ligand complexes. It is in these nonstandard systems that our approach shows its greatest promise.

The finite perturbation theory (FPT) developed in the framework of the MNDO[11] Hamiltonian using gauge including atomic orbitals (GIAO[12]) has shown promising results for the calculation of ^{13}C chemical shifts.[13] The FPT-MNDO GIAO method has recently been implemented using a divide-and-conquer strategy for the diagonalization of the complex Fock matrix to yield the perturbed density matrix with respect to the magnetic field.[14] This method has been used with NMR-specific parameters developed for ^{1}H,^{13}C, ^{15}N, and ^{17}O, and was shown to give fast and accurate results that can be applied to large protein−ligand complexes.[14]

In previous applications, the AM1[15] Hamiltonian was used for geometry optimizations, and a subsequent single-point NMR calculation was performed using the MNDO Hamiltonian. This approach was chosen because AM1 has improved on some structural features that are important to biological systems. Important for proteins, is the qualitative ability of AM1 to account for the energetics associated with hydrogen bonding interactions that are critical. An improved description of hydrogen bonding makes AM1 a more appropriate Hamiltonian than MNDO for geometry optimizations of protein systems. It would be more consistent to arrive at the geometry and perform the subsequent chemical shift calculations at the same level of theory. Therefore, AM1 was chosen as the Hamiltonian for both geometry optimization and the NMR calculation in the present work. Furthermore, NMR-specific MNDO parameters have already been developed:[16] thus, to make the distinction between these protein-specific NMR parameters and the more general NMR parameters, a different Hamiltonian was chosen for this parametrization. The new NMR-specific AM1 parameters are given in Table 1.

† Part of the "Walter Thiel Festschrift".

* Corresponding author. E-mail: merz@qtp.uf.edu.

10.1021/jp9028722 CCC: $40.75
Published on Web 10/02/2009

TABLE 1: Comparison of Standard versus NMR-Optimized AM1 Parameters

atom	parameter	AM1	AM1-NMR	AM1-NMR-H	AM1-NMR-C
H	ζ_s (au)	1.1880780	1.15036640	1.15994203	1.14478000
	β_s (eV)	−11.3964270	−14.96943835	−14.69571729	−15.15458000
C	ζ_s (au)	1.8086650	1.76509177	1.68883970	1.68679400
	ζ_p (au)	1.6851160	1.63588167	1.62877282	1.65943900
	β_s (eV)	−15.7157830	−18.89678664	−17.95998381	−17.66492000
	β_p (eV)	−7.7192830	−11.94651020	−14.64670379	−12.08854000
N	ζ_s (au)	2.3154100	2.34452630	2.16847239	2.12992700
	ζ_p (au)	2.1579400	2.05160150	2.06083458	2.11709500
	β_s (eV)	−20.2991100	−26.05351514	−27.62130377	−29.62571000
	β_p (eV)	−18.2386660	−18.66264348	−12.20412355	−14.56151000
O	ζ_s (au)	3.1080320	3.10365447	3.63641586	3.40861800
	ζ_p (au)	2.5240390	2.54342588	2.35923609	2.39947230
	β_s (eV)	−29.2727730	−29.05642758	−28.10519165	−26.94522300
	β_p (eV)	−29.2727730	−30.23181062	−29.27050269	−29.68913700
S	ζ_s (au)	2.3665150	2.36651500	2.45028965	2.51893987
	ζ_p (au)	1.6672630	1.66726300	1.60125168	1.66726300
	β_s (eV)	−3.9205660	−3.92056600	−4.26878731	−13.57459970
	β_p (eV)	−7.9052780	−7.90527800	−10.11576964	−8.83340467

Methods

Parametrization. Reparametrization of semiempirical QM methods in previous studies has been shown to significantly improve the agreement between experimental and calculated NMR chemical shifts for ^{1}H, ^{13}C, ^{15}N, ^{17}O, and ^{19}F nulcei.[13,16,17] The NMR-specific parameters for the MNDO approximation and a detailed explanation of the choice of parameters to be optimized was given in these previous works. Our method for the fast semiempirical QM NMR calculations has been outlined in a previous publication,[14] in which the optimized $\beta_{(s/p)}$ and $\zeta_{(s/p)}$ parameters are adopted.[16] The $\zeta_{(s/p)}$ parameters are the Slater orbital exponents, and they describe the character of the atomic orbitals. The $\beta_{(s/p)}$ parameters are the two-center/one-electron resonance integrals. The formalism for AM1-NMR is the same as our MNDO-NMR procedure and the choice of the type of parameters to be optimized is also the same. Although only ^{1}H and ^{13}C chemical shifts are reproduced in this study, it was necessary to optimize the $\beta_{(s/p)}$ and $\zeta_{(s/p)}$ parameters for C, H, N, O, and S to achieve the best agreement between experimental and calculated values.

The chemical shifts are calculated as a difference between the calculated shielding constant and a chosen reference value according to eq 1.

$$\delta_{calc} = \sigma_{ref} - \sigma_{calc} \tag{1}$$

In addition to generating new parameters for the calculation of the shielding constants, optimization of the reference value in eq 1 was performed to address systematic problems of over- or underestimation of the chemical shifts. During the parametrization this reference value was initially set to the value of shielding constants calculated for methane for a given set of parameters. Next, the average signed error was calculated (see eq 2) between the experimental and calculated chemical shifts. The signed error was then added to the initial reference value and this adjusted number was used as the final σ_{ref} value.

$$\text{average signed error} = \frac{\sum(\delta_{exp} - \sigma_{calc})}{N} \tag{2}$$

For a given set of parameters, this method essentially sets the average signed error to zero and minimizes the root-mean-squared (rms) error for the data set. The scoring function was then used to optimize the parameters by minimizing the rms error between experimental and calculated chemical shifts via eq 3. All chemical shifts and errors for QM calculations in this work were evaluated using the reported σ_{ref} value chosen to minimize the average signed and rms error. In all instances the calculations using the new parameters are single-point NMR calculations. The geometries were generated using standard AM1 parameters because the new parameters are NMR-specific and were not tested for their ability to provide realistic geometries, nor were they developed for that purpose.

$$G = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_{i(exp)} - \delta_{i(calc)})^2} \tag{3}$$

The new AM1-NMR parameters were optimized using a modified genetic algorithm (GA). A full description of the type of GA used in this work has previously been published.[18] A GA has been successfully used previously in several semiempirical parametrizations, and the applicability of this type of optimization routine has previously been discussed.[17–19] The GA and all handling of the data were performed using an in-house molecular tools package MTK++ (Molecular Tool Kit [written in C++]).[20] All semiempirical calculations were performed using our linear scaling semiempirical DivCon program.[21] The results obtained for protein systems using the AM1-NMR approach are compared to those given by the MNDO-NMR protocol and the empirical SHIFTX[10] program that is freely available. SHIFTX version 1.1, the version that is available as of the date of this publication, does not have the capability to predict chemical shifts for all atoms in the protein systems. Most notably, the chemical shifts for most side-chain carbon atoms could not be predicted.

Experimental Data. The large number of protein structures in the Protein Data Bank (PDB) that were solved by NMR is one of many testaments to the important role that NMR plays in the study of protein systems. It was therefore chosen as a primary goal in this parametrization to improve the ability of the semiempirical NMR protocol to predict the chemical shifts of atoms in this particular environment. The MNDO-NMR parameters were developed with the broad aim of reproducing the chemical shifts in a wide variety of small organic compounds. Various functional groups present in that training set

TABLE 2: Small Molecule ^{1}H and ^{13}C RMS Errors Using σ_{ref} Values Optimized on the Small Molecule Data Set[a]

		NMR method				
		MNDO-NMR	AM1-NMR-H	AM1-NMR-C	AM1-NMR	AM1
^{1}H	rms error	0.61	0.63	0.64	0.48	1.42
154 shifts	R^2	0.96	0.94	0.94	0.96	0.92
	av signed error	0	0	0	0	0
	av unsigned error	0.49	0.59	0.49	0.37	1.24
	σ_{ref}	50.010	50.423	49.672	49.716	47.680
^{13}C	rms error	10.54	12.59	10.00	9.86	21.54
176 shifts	R^2	0.97	0.96	0.97	0.97	0.86
	av signed error	0	0	0	0	0
	av unsigned error	7.91	9.54	7.51	7.78	15.66
	σ_{ref}	44.226	85.079	71.956	63.354	−58.802
	σ_{ref} (C−C π bonds)	44.226	85.079	51.403	63.354	−58.802

[a] Errors include the complete data set of training and test set using the average signed error as the σ_{ref} value in eq 1. The average signed errors are zero because they are absorbed in the σ_{ref} values.

TABLE 3: Small Molecule ^{1}H and ^{13}C RMS Errors Using σ_{ref} Values Optimized on the Protein Data Set[a]

		NMR method				
		MNDO-NMR	AM1-NMR-H	AM1-NMR-C	AM1-NMR	AM1
^{1}H	rms error	0.71	0.71	0.68	0.76	1.43
154 shifts	R^2	0.96	0.94	0.94	0.96	0.92
	av signed error	+0.37	+0.33	+0.22	+0.59	+0.14
	av unsigned error	0.59	0.54	0.53	0.65	1.28
	σ_{ref}	49.603	50.062	49.418	49.090	47.503
^{13}C	rms error	10.82	12.74	10.10	9.91	24.15
176 shifts	R^2	0.97	0.96	0.97	0.97	0.86
	av signed error	−2.46	+1.96	+1.30	−0.95	−10.91
	av unsigned error	7.84	9.61	7.54	7.73	19.85
	σ_{ref}	46.689	83.114	70.905	64.309	−47.8900
	σ_{ref} (C−C π bonds)	46.689	83.114	49.32	64.309	−47.8900

[a] Errors include the complete data set of training and test set.

TABLE 4: Protein Systems Used in the Parametrization

PDB ID	description	$N_{conormers/}$ resolution	model/chain	missing residues	BMRB accession
1HA8	pheromone	20	1	H, I, K, M, R, V, W	4979
1N87	pre-mRNA splicing factor PRP19	20	1	H, W	5594
1Q2N	Z domain- staphylococcal protein A	10	1	C, G, M, T	5656
1RZS[a]	P22 Cro protein	21	1	C	6185
1SZL	F-spondin protein	20	1	H	6175
2CA7[a]	Conkunitzin-S1 Kunitz-type neurotoxin	20	2	H, M, V, W	6506
1YV8[a]	crambin	20	1	H, K, M, Q, W	6455
2B7E[a]	pre-mRNA processing protein 1st FF domain	12	1	C, H	6850
2FS1	bacterial albumin binding protein	20	1	C, F, H, R, W	6945
2JN0[a]	ygdR protein from *E. coli*	20	1	C, F, W	15079
1EZG	antifreeze protein (*T. molitor*)	1.40Å	B	N, C, E, M, W	5323
1F94	bucandin	0.97Å	A	C, H, P, Q	5097
1FD3	β-defensin-2 (human)	1.35Å	D	C, H, P, Q	4642
1L3K	ribonucleoprotein A1 (HNRP-human)	1.10Å	A		4084
1UBQ	ubiquitin	1.80Å		C, W	5387

[a] Structures used in the training set. All others were included in the test set.

are not frequently seen in biological applications. A few of these include O_3, O═C═C═C═O, CH_3—N═N═N, and ON—NO_2. Therefore, we initially sought to develop NMR-specific parameters that were suited for biological compounds using a data set of small molecules with functional groups that are more frequently seen in biology. This data set consisted of 94 small molecules, of which 65 were adopted from the small reference set used in the parametrization of MNDO-NMR. The remaining compounds were added to incorporate more functional groups that may be useful in biological applications. These structures are included in the Supporting Information, and the results for this data set are presented in Tables 2 and 3. This first set of parameters (referred to as AM1-NMR) resulted in only modest improvement upon our current protocol for MNDO-NMR for this data set. Furthermore, when tested on several large protein systems, the agreement with experiment was not as close as was hoped. It was therefore decided to carry out a parametrization using protein data.

The protein structures chosen for the training and test sets are listed in Table 4. These are unbound structures that range in length from 46 to 61 amino acid residues. It was essential for the molecules to be sufficiently large and exhibit a range of secondary structures, including α helices, β sheets, and random coils. The proteins could not be too large because it would impede the speed with which the parametrization could be carried out. Figure 1 illustrates the distribution of amino acid residues in the training set. Figure 2 compares the distribution of amino acid residues in the test set with that of the complete

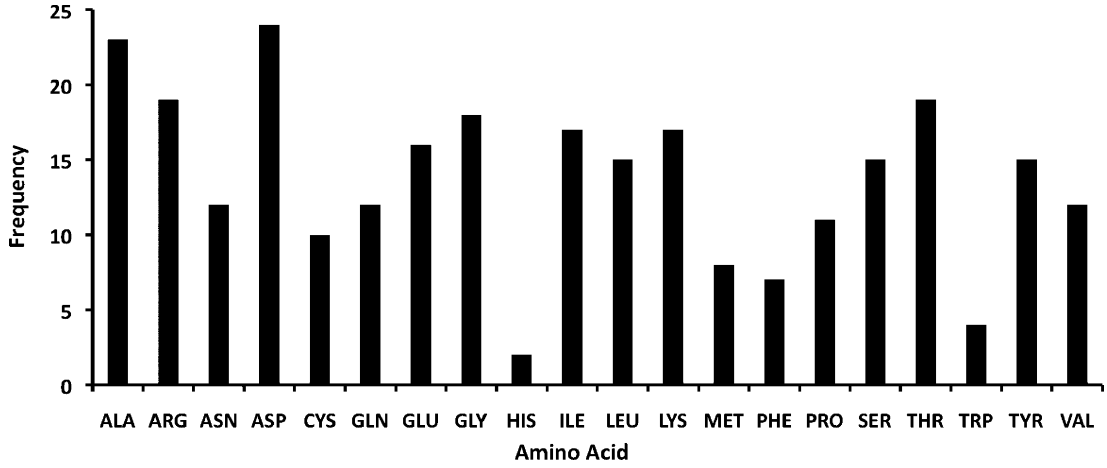

Figure 1. Distribution of amino acids in protein training set.

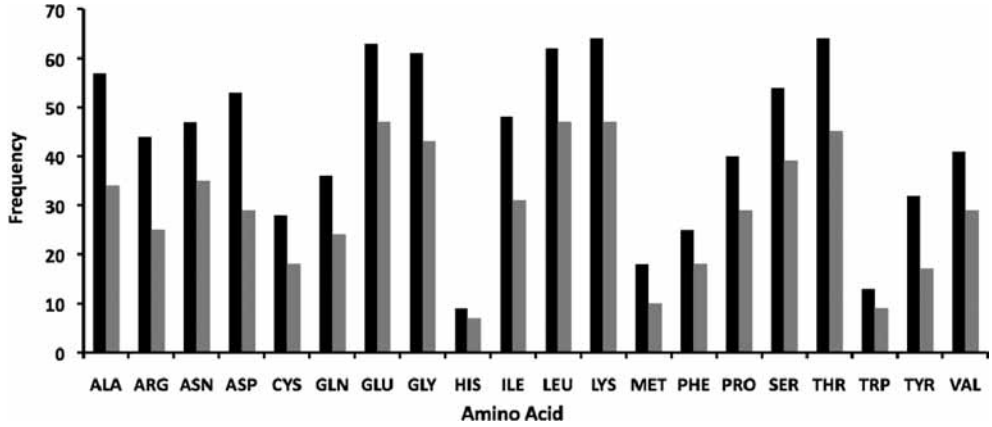

Figure 2. Distribution of amino acids in complete protein data set (black) compared to test set (gray).

protein data set including both training and test sets. The experimental chemical shifts were taken from the BMRB database.[22] There was a single instance in which the reported experimental data point was not used because it was deemed to be inaccurate (2CA7: Y24-CD1). The proteins used contained only C, H, N, O, and S atoms. No metals were present and all water molecules were removed.

NMR structures in the PDB are seldom given as a single structure. Quite frequently they are given as a series of models. The protein model used was that which the PDB file designated as the best representative conformer. The second model was used in 2CA7 because of structural flaws in the best representative conformer.

Although using high-resolution crystal structures may lower the chance of structural defects affecting the quality of the parametrization, NMR structures were used because it is more consistent to use the structures that were actually generated from the NMR experiments. Additionally, in many instances the problems associated with the use of crystal structures ultimately made the use of NMR structures a more suitable option for this type of parametrization. Among the problems encountered: First, in many instances the combination of a chosen crystal structure and matching BMRB data did not meet the general criteria outlined above because they were protein−ligand complexes, had incommensurate lengths, or lacked the secondary structure of interest. Second, they often had an insufficient number of reported chemical shifts; therefore, the number of chemical shifts that they introduced into the reference data was disproportionately small relative to the computational expense. A third important problem was that the crystal structures often had mutations that resulted in a mismatch with the BMRB data. Although several of these problems would not affect the use of these structures in training a statistical method, they were more problematic in a QM study. Because QM calculations account for the long-range interactions, even a single mutation in the corresponding structure was deemed unsuitable for training purposes, more so than small structural flaws that can be remedied by geometry optimization. Furthermore, because of the computational expense of QM models, very large systems were not suitable for training. Nonetheless, the test set was augmented with five high-resolution crystal structures.

The challenge of accurately reproducing the chemical shifts of amide protons is well-documented.[23] This is due in part to the varying rates of exchange that these protons experience with solvent during the experimental NMR procedures. For this reason, exchangeable protons were not used in this parametrization and the only hydrogen atoms included were those involved in H−C bonds.

Data Set Preparation. Because the goal of this work was to provide parameters that could be used for large biological molecules, semiempirical geometries were used; the large system sizes precludes the use of structures generated via higher level calculations. For consistency, the geometries were all optimized in vacuo using the standard AM1 parameters prior to performing the single-point NMR calculations. This was done to minimize the errors due to variations in the experimental structures. It was important to normalize the structures instead of using experimental geometries because the QM calculation of the chemical shifts is highly sensitive to bond lengths and angles, and slight variations could significantly impact the quality of

the parameters. The optimization used the steepest decent routine for 30 steps, followed by conjugate gradient, until certain convergence criteria were met. These criteria were maximum changes in the energy, gradients, and coordinates of 0.1 kcal/mol, 1.0 (kcal/mol)/Å, and 0.001 Å, respectively. The structures were then carefully inspected to ensure that no significant errors occurred during the geometry optimization. These structures were then used to generate the new parameters.

Several structures in the test set were high-resolution crystal structures. These included four structures from the cross-validation set initially used to evaluate the SHIFTX program. The last structure, ubiquitin, was chosen because it is commonly used to evaluate the predictive quality of NMR related programs. In some instances the crystal structure did not contain the coordinates for the side chains of every residue. It was therefore necessary to add all missing atoms using the LEaP program in AMBER 9.[24] After all side-chain heavy atom coordinates were built, all of the crystal structures were protonated using AMBER. In a few cases, this involved the removal of the hydrogen atoms that were present in the PDB file. This was done for consistency, and to ensure that there were no significant van der Waals clashes. All crystallographic waters were removed and a restrained minimization was then carried out using AMBER to minimize the protons with the heavy atom coordinates held constant. All of the geometry optimizations using AMBER were carried out with the ff99SB force field.[25]

Results and Discussion

General Details. Three sets of new NMR-specific parameters [ζ_s, ζ_p, β_s, β_p] for the AM1 Hamiltonian are listed in Table 1. The first set (AM1-NMR) consists of 14 parameters optimized on small molecules to reproduce both ^{1}H and ^{13}C chemical shifts. The second set (AM1-NMR-C) and third set (AM1-NMR-H) were generated using protein systems as the training set (see Table 4); they consist of 18 parameters optimized to reproduce ^{13}C and ^{1}H chemical shifts, respectively. No single parameter set was able to simultaneously achieve the highest level of accuracy possible for all three types of systems of interest. However, the results from the AM1-NMR-C set of parameters suggest that they represent the best balance of accuracy and versatility for the prediction of ^{1}H and ^{13}C chemical shifts for proteins and small molecules.

For consistency, all QM chemical shifts were calculated using a σ_{ref} value that minimizes the average signed error for this entire data set. The σ_{ref} values used in the implementation of eq 1 are listed in the various tables in which the results are summarized. It should be noted that by virtue of using the semiempirical approximations, our approach cannot explicitly account for the contributions of the core electrons to the absolute chemical shielding constants.[14] As illustrated in Table S2 of the Supporting Information, both the paramagnetic and diamagnetic contributions to the shielding constant are off when compared with DFT results at the B3LYP/6-311++G** level for ethane. However, because the core contributions to the absolute shielding are constant, they can be absorbed in the chosen σ_{ref} value enabling a good qualitative description of the chemical shifts. With the assumption that the training set was sufficiently large in this parametrization, these reported σ_{ref} values should be appropriate for future application. An alternative approach is to use the value of the shielding constants calculated for methane as the σ_{ref} values. For the protein systems examined, the signed errors using the optimized σ_{ref} values for MNDO-NMR were found to be 0.74 and 5.11 for ^{1}H and ^{13}C, respectively. Using the shielding constant for the carbon atom in methane as the σ_{ref} value increased the rms error very slightly to 5.12 for ^{13}C. A more significant difference was made for hydrogen where using methane as the reference value increased the rms error from 0.74 to 0.88 ppm. Therefore, for protein studies using our implementation of the MNDO-NMR procedure, the new σ_{ref} value is likely to improve results for ^{1}H NMR chemical shift predictions. Despite the small difference that it made for the ^{13}C results, the optimized value for carbon was still used for all calculations presented here; this allowed for consistency when the results were compared to the new parameters. In the original parametrization of MNDO-NMR a different reference value was used for hydrogen for (C−H), (O−H), and (N−H). Because polar hydrogen atoms were excluded from the present study, a single reference value was also used for all hydrogen atoms. A single reference value was also used for the AM1-NMR and AM1-NMR-H parameter sets. As discussed later, a single reference value was used for the AM1-NMR-C parameter set for hydrogen, but two reference values were used for the carbon atoms. This was the only instance in which more than one reference value was used for an atom.

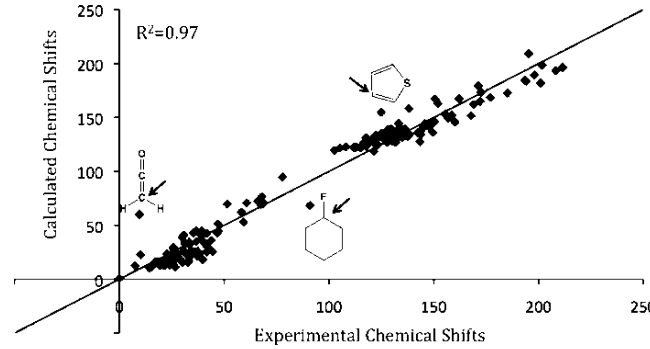

Figure 3. ^{13}C chemical shift correlation for small molecule data set using AM1-NMR parameters. Compounds with the largest deviation from experiments are drawn with the arrow pointing to the particular atom.

Both the orbital exponents [ζ] and the resonance parameters [β] were critical for obtaining good agreement with experimental NMR data, and subtle changes in either one significantly affected the results. Semiempirical methods exhibit a strong interdependence between their parameters. Changing the parameter of one atom has such a significant effect on neighboring atoms that it was necessary to change all of the ζ_s, ζ_p, β_s, and β_p parameters for C, H, N, O, and S to obtain the best agreement between experimental and calculated chemical shifts. This interdependence makes interpreting the changes that were made to the parameters less straightforward. However, as is apparent from the smaller values of the Slater orbital exponent (ζ_s and ζ_p) terms for hydrogen and carbon in Table 1, the description of both ^{1}H and ^{13}C chemical shifts benefit from the use of orbital exponents that are more diffuse than those in the standard AM1 parameter set.

The first set of parameters developed (AM1-NMR) is suited for the calculation of NMR chemical shifts for small molecules using the AM1 Hamiltonian. Encouraging results using these parameters for the small molecule data set are plotted in Figures 3 and 4. However, an improvement in the agreement between experimental and calculated chemical shifts for protein systems was reached by narrowing the scope of the parametrization "basis set" in subsequent parametrizations. These improvements are made clear by examining the results for the individual protein systems as outlined in Tables 5 and 6. As shown in Table 7, the new NMR parameters improved the average rms errors of

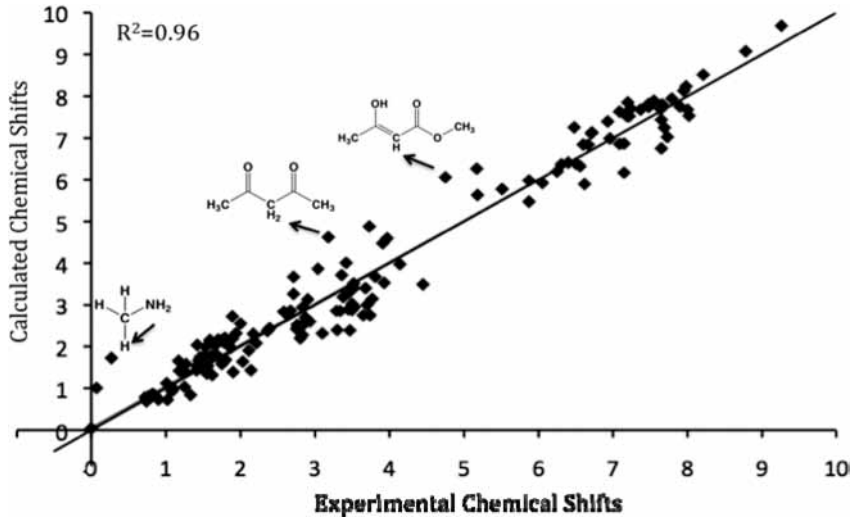

Figure 4. ^{1}H chemical shift correlation for small molecule data set using AM1-NMR parameters. Compounds with the largest deviation from experiments are drawn with the arrow pointing to the particular atom.

TABLE 5: Comparison of ^{1}H RMS Errors (ppm) by Protein

PDB ID	MNDO-NMR	AM1-NMR hydrogen	AM1-NMR carbon	AM1-NMR	AM1	no. of ^{1}H shifts
1HA8	0.86	0.67	0.76	0.85	1.28	182
1N87	0.78	0.62	0.70	0.74	1.13	208
1Q2N	0.53	0.60	0.33	0.59	1.04	18
1RZS[a]	0.70	0.57	0.60	0.67	1.06	290
1SZL	0.75	0.63	0.65	0.69	1.07	277
2CA7[a]	0.67	0.57	0.60	0.64	0.94	219
1YV8[a]	0.72	0.54	0.64	0.70	1.00	169
2B7E[a]	0.79	0.63	0.68	0.75	1.13	257
2FS1	0.71	0.53	0.62	0.69	1.01	232
2JN0[a]	0.72	0.54	0.62	0.65	1.06	191
1EZG	0.73	0.67	0.69	0.72	0.98	257
1F94	0.64	0.61	0.59	0.64	1.02	269
1FD3	0.58	0.50	0.52	0.60	0.94	202
1L3K	0.67	0.54	0.59	0.64	0.93	255
1UBQ	0.90	0.83	0.85	0.86	1.07	314

[a] Structures used in the training set. All others were included in the test set.

TABLE 6: Comparison of ^{13}C RMS Errors (ppm) by Protein

PDB ID	MNDO-NMR	AM1-NMR hydrogen	AM1-NMR carbon	AM1-NMR	AM1	no. of ^{13}C shifts
1HA8	4.49	5.40	4.59	7.34	8.21	51
1N87	6.02	7.24	5.80	8.52	12.98	141
1Q2N	4.62	17.70	4.03	7.78	26.06	191
1RZS[a]	5.09	6.50	4.41	6.43	14.62	210
1SZL	5.26	9.67	4.59	7.57	24.81	243
2CA7[a]	5.05	9.95	4.68	8.15	24.92	208
1YV8[a]	3.87	10.64	3.64	6.77	25.17	162
2B7E[a]	4.82	6.53	4.72	7.44	12.41	179
2FS1	4.32	10.01	4.45	7.10	24.16	209
2JN0[a]	4.89	9.18	4.36	7.25	22.88	183
1L3K	5.20	6.10	4.91	7.53	10.73	181
1UBQ	6.37	10.29	6.43	8.37	23.21	275

[a] Structures used in the training set. All others were included in the test set. Experimental ^{13}C chemical shifts were not available for 1EZG, 1F94, and 1FD3.

^{1}H and ^{13}C chemical shifts by 0.12 ppm and 0.28 ppm, respectively, over the more general MNDO-NMR parameters. To best reproduce the chemical shifts of ^{13}C and ^{1}H nuclei for protein systems, it was necessary to develop two separate parameter sets (AM1-NMR-C and AM1-NMR-H). Although the AM1-NMR-C parameters perform well for the prediction of ^{1}H chemical shifts, by sacrificing the accuracy of ^{13}C, a much better description of the ^{1}H chemical shifts was attainable with the AM1-NMR-H parameters set (see Tables 5 and 7). Because both the MNDO-NMR and AM1-NMR parameters already serve as more general-purpose sets, we decided to enhance the AM1-NMR-H parameters for ^{1}H NMR in the protein systems without the limitation of having to ensure good performance for any other type of chemical shifts.

The encouraging results obtained for the proteins of the test set indicate that the protocol used in this study does present an improvement upon other semiempirical QM methods to evaluate NMR chemical shifts in protein systems. This improvement is significant in comparison to standard AM1 parameters. Accounting for other factors might also lead to improvements in this method. First, the inclusion of solvent may be a first step toward achieving greater accuracy. The parameters were developed to reproduce solution phase chemical shifts but all of the calculations, including both geometry optimizations and single-point NMR calculations, were run in vacuum. Second, it is not clear what artifacts may have been brought about in this study by the use of NMR structures and the AM1 Hamiltonian to optimize the geometry of these protein systems. If there were undetected significant artifacts, addressing this issue may lead to improved results. Although a recent study has investigated the effect of geometry optimization of protein systems using PM6,[26] it is not clear how similar the artifacts of AM1 geometry optimization would be on protein systems. A detailed examination of the effect of the geometry on the prediction of NMR chemical shifts will be described in a future publication.

Higher level ab initio and DFT calculations cannot easily be performed on the protein systems used here; therefore, in Tables 8 and 9, the results for the protein systems are compared to those of the SHIFTX program. The SHIFTX calculations were performed on the experimental structures without any editing of the structure. Because SHIFTX could not perform ^{13}C NMR predictions on most of the side-chain carbon atoms, only C′, C_α, and C_β are compared. On average the SHIFTX program yielded results that were ~2.5 ppm better than those predicted using the ^{13}C parameters. The best semiempirical models currently have errors roughly twice as large as those of the

TABLE 7: Protein ^{1}H and ^{13}C RMS Errors[a]

		NMR method				
		MNDO-NMR	AM1-NMR-H	AM1-NMR-C	AM1-NMR	AM1
^{1}H	rms error	0.74	0.62	0.66	0.71	1.05
3340 shifts	R^2	0.86	0.86	0.87	0.86	0.86
	av unsigned error	0.58	0.48	0.51	0.56	0.84
	σ_{ref}	49.603	50.062	49.418	49.090	47.503
^{13}C	rms error	5.11	9.95	4.83	7.60	21.28
2233 shifts	R^2	0.99	0.97	0.99	0.98	0.92
	av unsigned error	3.98	7.70	3.67	6.43	15.51
	σ_{ref}	46.689	83.114	70.905	64.309	−47.8900
	σ_{ref} (C−C π bonds)	46.689	83.114	49.320	64.309	−47.8900

[a] Errors include the complete data set of training and test set. The σ_{ref} value was chosen to minimize the average rms error and sets the average signed error to zero.

TABLE 8: RMS Errors of ^{1}H and ^{13}C NMR Chemical Shifts for Complete Protein Set

		NMR method					
		MNDO-NMR	AM1-NMR-H	AM1-NMR-C	AM1-NMR	SHIFTX	no. of shifts
rms error	H_α	0.81	0.60	0.67	0.82	0.34	844
	C_α	4.16	7.03	4.00	7.25	1.98	668
	C_β	5.80	7.60	5.94	7.65	1.93	579
	C′	3.46	17.23	3.38	8.26	1.82	358

TABLE 9: Correlation of ^{1}H and ^{13}C NMR Chemical Shifts for Complete Protein Set

		NMR method					
		MNDO-NMR	AM1-NMR-H	AM1-NMR-C	AM1-NMR	SHIFTX	range (ppm)
R^2	H_α	0.22	0.22	0.24	0.26	0.62	1.70−6.14
	C_α	0.34	0.06	0.39	0.28	0.84	40.7−68.4
	C_β	0.87	0.68	0.86	0.86	0.98	15.4−73.4
	C′	0.02	0.17	0.01	0.01	0.38	170.2−181.7

SHIFTX. This is not surprising since SHIFTX and related programs can use extensive atom typing, whereas this study has one H or C parameter set to directly work along with indirect effects generated through modification of allied parameters for N, O, and S. SHIFTX, through atom typing, outperforms the present approach, but a QM approach is still more versatile and can be used to facilitate computational studies of chemical shift perturbation upon ligand binding, and investigations involving nonstandard amino acids.

Drawbacks. While the overall rms error is quite low when using the new sets of parameters for protein systems, this method does not appear to differentiate well among the same atom-types in dissimilar chemical environments. One example of this can be seen in the C′ chemical shifts of residues ASP31 and THR6 of the 2JN0 structure. The experimental chemical shifts are reported as 176.93 and 177.34 ppm, respectively; the calculated chemical shifts using AM1-NMR-C are179.10 and 174.99, respectively. While the order is incorrect, both calculated chemical shifts are well within the rms error of 4.36 ppm. This is illustrated in a decomposition of the R^2 values for 2JN0 using the AM1-NMR-C parameters in Figure 5. This clustering effect is present in all of the semiempirical methods evaluated here for the protein systems, and it is very evident in carbon atoms. While the overall R^2 values are quite high for all carbon atoms, the R^2 values for different clusters are quite low, as is illustrated in Table 9. Tests were run to determine whether the buffer size in the divide-and-conquer scheme limited the environmental effect on the chemical shifts. These tests showed no indication that this was the case. It appears that rms error is sufficiently large to encompass a significant portion of the range of experimental chemical shifts observed for each particular atom type. This renders the method weak at differentiating among these similar atoms in some cases. This is supported by the fact that the clusters with the smaller ranges exhibit worse R^2 values. This effect is also found in SHIFTX as illustrated in Figure 6, but to a lesser extent because their rms errors are lower (by a factor of 3 in this instance).

AM1-NMR. The first set of parameters was generated using the small molecule data. The variety of functional groups in this data set makes these parameters more general than the others; therefore, this set of parameters is referred to simply as AM1-NMR. The chemical shifts of this small molecule data set ranged from 0 to 211.5 ppm and 0 to 9.3 ppm for ^{13}C and ^{1}H, respectively. As shown in Table 2, this parameter set performed very well for the small molecule data set. The average rms errors were 9.86 and 0.48 ppm corresponding to 4.7% and 5.2% of the chemical shift range, respectively. Therefore, using these parameters, the goal of obtaining 5% of the chemical shift range was met for carbon and a very close result was achieved for hydrogen. However, these results were generated using a σ_{ref} value that minimized the error for this data set. It is desirable to have a single σ_{ref} value that can be used for a variety of systems. To determine how extensible this parameter set was, tests were run for the small molecule data set using the σ_{ref} value that minimized the rms error for the protein data set.

As shown in Table 3, using the σ_{ref} value optimized for the protein systems resulted in a significant difference in error being observed for ^{1}H NMR calculations in the small molecule data set, changing the rms error from 0.48 to 0.76 ppm. This difference suggests that the effect of the environment in protein systems was more important than was accounted for in the small molecule parametrization. The difference in error for ^{13}C NMR calculations was much smaller, only changing the rms error from 9.86 to 9.91 ppm. Furthermore, even when the different σ_{ref} value was used, the ^{13}C NMR results were best for the small molecule data set using these AM1-NMR parameters. The larger errors were observed for both ^{13}C and ^{1}H NMR calculations for small crowded systems and those neighboring heteroatoms—sulfur and nitrogen in particular. The molecules with the largest errors are given in the scatter plot in Figures 3 and 4 for carbon and hydrogen respectively. Because the data set was more oriented toward biological compounds, there was an overabundance of aromatic systems, and consequently the parameters performed better in these cases. A detailed comparison of the experimental chemical shifts with those calculated using all parameters sets is given in Table S1 of the Supporting Information.

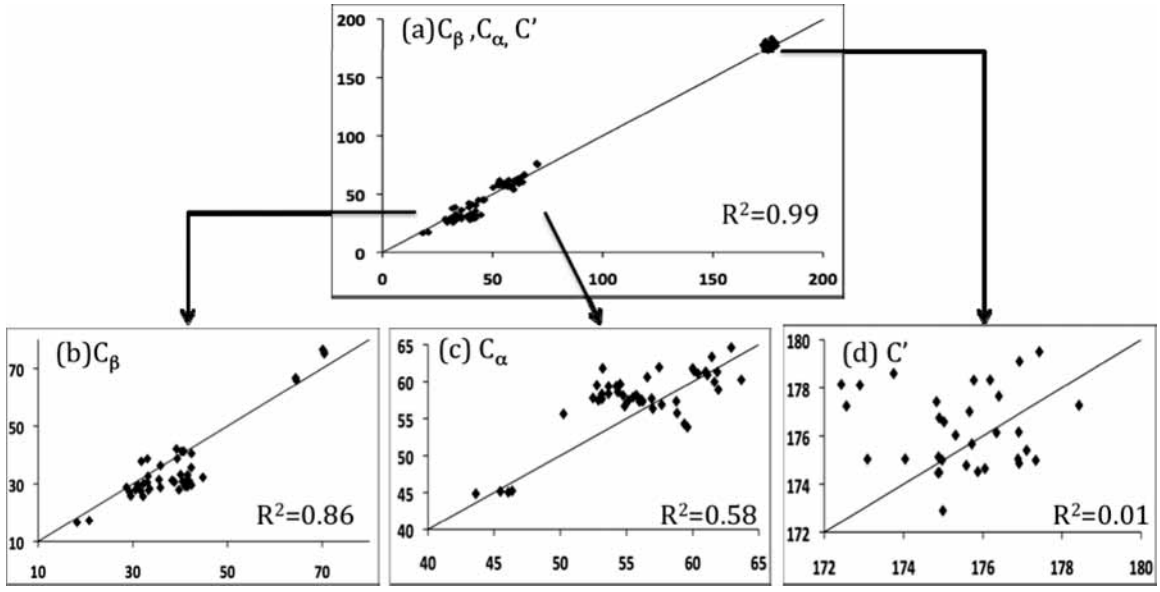

Figure 5. Decomposition of ^{13}C NMR R^2 value for PDB ID 2JN0 using AM1-NMR-C parameters. For direct comparison with SHIFTX results in Figure 6, only C_β, C_α, and C′ atom types are included. Figures b, c, and d are the C_β, C_α, and C′ subsets of figure a, respectively. The *x*-axis is the experimental and the *y*-axis is calculated chemical shifts. The average ^{13}C rms error for this protein is 4.36 ppm. (a) The overall R^2 for all C_β, C_α, and C′ atoms is 0.99. (b) R^2 for C_β is relatively high at 0.86 because large variety in the bonding situations for C_β atoms results in a large chemical shift range of >50 ppm. (c) C_α $R^2 = 0.58$. With less variety in bonding situations the chemical shift range decreases and the rms error has a greater impact as exhibited in the decreased R^2 value. (d) For very small chemical shift ranges such as C′, the rms error is too large to distinguish among different atoms of this atom type. It is important to note that despite the poor correlation, the magnitude of the errors is still quite small.

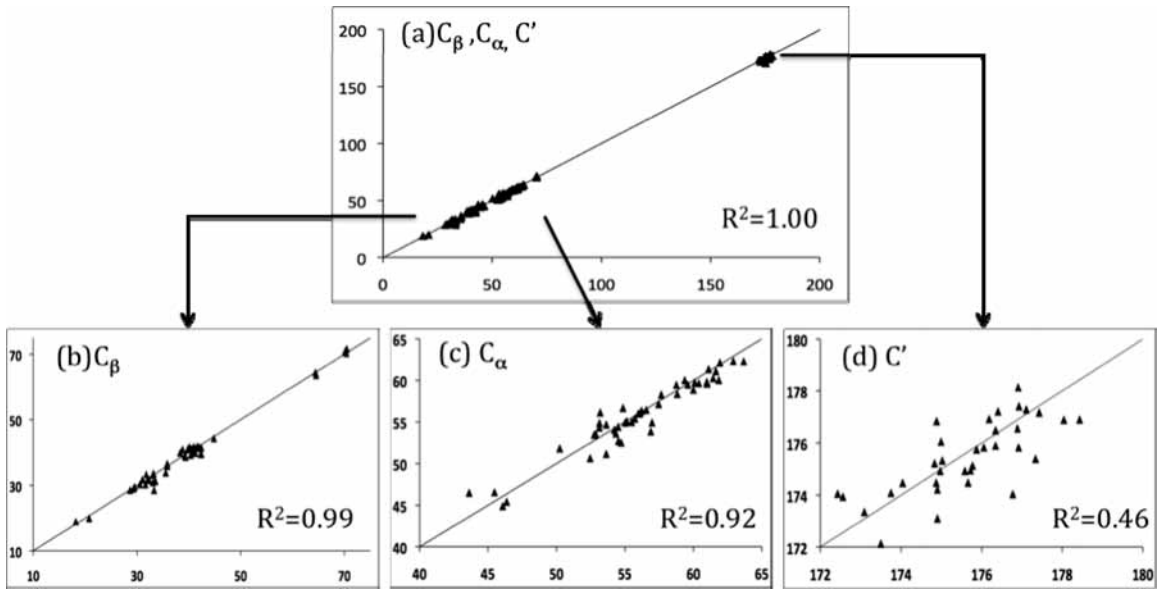

Figure 6. Decomposition of the ^{13}C NMR R^2 value for PDB ID 2JN0 with SHIFTX. Figures b, c, and d are the C_β, C_α, and C′ subsets of figure a, respectively. The *x*-axis is the experimental and the *y*-axis is calculated chemical shifts. The average ^{13}C rms error for this protein is 1.33 ppm. (a) The overall R^2 for all C_β, C_α, and C′ atoms is 1.00. (b) R^2 for C_β is still high, at 0.99, because large variety in the bonding situations for C_β atoms results in a large chemical shift range of >50 ppm. (c) C_α $R^2 = 0.92$. With less variety in bonding situations the chemical shift range decreases and the rms error has a greater impact as exhibited in the decreased R^2 value. (d) For C′, which covers a much smaller range of chemical shifts $R^2 = 0.46$.

This parameter set showed significant improvement compared to the standard AM1 parameters for both the small molecule and protein data sets. It also exhibited an improvement relative to MNDO-NMR for the overall prediction of ^1H NMR chemical shifts for the protein systems. However, the performance for ^{13}C NMR prediction for protein systems was worse than that of the MNDO-NMR procedure. Because no improvement was made in this area using the small molecule training set, more parametrization was done using globular proteins in the data set, which represent an important target application area for us.

AM1-NMR-C (Carbon). The second set of parameters was optimized to reproduce ^{13}C and ^1H NMR data for a set of protein systems. To differentiate between these and the third set of parameters (which were generated to reproduce ^1H NMR only), these parameters are referred to as AM1-NMR-C, even though they do perform very well for ^1H NMR calculations. The ^{13}C chemical shifts for the complete protein data set ranged from 8.49 to 181.70 ppm. When the AM1-NMR-C parameter set was used, the average rms errors for the complete protein data set was 4.83 ppm, corresponding to just below 2.5% error. When

compared with the original AM1 parameters, the final parameters show a significant improvement in agreement with experimental results for both the small molecule and protein data sets. This parameter set also performed very well for both ^{1}H and ^{13}C NMR predictions in the small molecule data set. As shown in Table 3, the σ_{ref} value generated to optimize performance for the protein system was still very well suited for the small molecule data set. In fact, this parameter set was least affected by this change, indicating that it is likely to be the most extensible of the methods developed and tested here. The quality of this set of parameters is also demonstrated in Table 7, where they show good improvement upon both the standard AM1 and MNDO-NMR results for the protein data set.

Large signed errors (−20 ppm) were noticed for carbon atoms that participate in one or more C−C π bonds when using the AM1-NMR-C parameters. In protein systems, this only affects the residues with aromatic side chains (HIS, PHE, TRP, TYR). However, the problem was observed for all of the small molecules in which these bonding situations were present. For this reason a different σ_{ref} value was used for these carbon atoms. It is important to note that these errors only involved the C−C bonds, and errors associated with carbon having multiple bonds to other atoms were not affected. The use of more than one reference value significantly improves the performance of this parameter set and may be a large part of the reason that these parameters outperform the other methods tested here. For no other semiempirical method tested did a particular functional group contain a large signed error that could be easily addressed by the use of a different reference value. Although significant improvements would be made by addressing other instances of systematic errors in this fashion, this procedure would lead to atom typing, which limits the versatility of these methods and was therefore avoided.

As with the AM1-NMR parameters, larger errors were found in crowded areas of the molecule. For example, the rms error for carbon atoms in buried residues (less than 25% solvent accessible) was 3.43 ppm for 94 C′ atoms and 4.45 for 195 C-α atoms. In contrast, the rms error in exposed residues (greater than 65% solvent accessible) was 3.04 ppm for 109 C′ atoms and 3.67 for 180 C-α atoms.

AM1-NMR-H (Hydrogen). The third set of parameters was developed to specifically reproduce only ^{1}H chemical shifts in protein systems. Attaining high accuracy for proton chemical shifts in protein systems is a significant challenge for a QM model. However, with the focus limited to improving the accuracy of ^{1}H chemical shift calculations for protein systems only, a reasonable goal was to achieve accuracy of 5% of the chemical shift range. The ^{1}H chemical shifts for the complete set of protein systems ranged from −0.53 to +7.74 ppm. The average rms error for the complete data set, including training and test sets was 0.62 ppm, corresponding to 7.5% of the chemical shift range for this data set. This is an improvement upon the MNDO-NMR procedure, which yielded an average rms error of 0.74 ppm, corresponding to 8.9% of the chemical shifts range. While 7.5% error is at the high end of our target for the prediction of ^{1}H chemical shifts, the R^2 of 0.86 is still encouraging. The high R^2 value suggests that the general trends are still well preserved. Furthermore, the results suggest that this parameter set is the best available semiempirical method for the prediction of ^{1}H chemical shifts in protein systems.

Conclusions

The addition of three new NMR-specific parameter sets for the AM1 Hamiltonian has now extended the semiempirical QM methodology for a near quantitative description of NMR chemical shifts. This approach to calculating the chemical shifts is more consistent than the previously implemented semiempirical approach for protein systems, because the geometries are also generated with the AM1 Hamiltonian. When compared to previously available semiempirical protocols, the reduction in error in protein systems is significant. Furthermore, the methods can be executed at a fraction of the cost of ab initio and DFT methods. The rationale for the development of three parameter sets is outlined and the possible limitations of the method are given in detail in the Results and Discussion. The results from the AM1-NMR-C parameter set suggest that they represent the best balance of accuracy and versatility for the prediction of ^{1}H and ^{13}C chemical shifts for proteins and small molecules.

Clearly the problem associated with using large systems for a training set in semiempirical QM methods is the computational cost. However, the benefit is seen in the significant reduction of the errors associated with this type of procedure being applied to relevant systems. The large number of chemical shifts examined in this study, combined with the promising results seen in the test set makes it likely that the current parameters are extendible to other protein systems.

While semiempirical methods are quantum mechanical and therefore do not require atom typing, they are limited in their flexibility due to the use of a minimal basis set and other approximations. The MNDO-NMR parameters were not developed with such a specific goal as to capture atoms in a protein environment. Therefore, the close agreement seen between the experimental chemical shifts and those calculated using the MNDO-NMR methodology for protein systems is a testament to the quality of the parametrization carried out by Patchkovskii and Thiel. However, by having a more focused target of achieving good agreement for the ^{1}H and ^{13}C chemical shifts found in protein environments, higher accuracy was achieved in these particular systems of interest.

Acknowledgment. We thank NSF-SEAGEP and NSF (MCB-0211639) for financial support.

Supporting Information Available: Experimental and calculated chemical shifts for small molecule data set. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293.
(2) Haser, M.; Ahlrichs, R.; Baron, H. P.; Weis, P.; Horn, H. *Theor. Chim. Acta* **1992**, *83*, 455.
(3) Kussmann, J.; Ochsenfeld, C. *J. Chem. Phys.* **2007**, *127*, 054103/1.
(4) Ochsenfeld, C.; Kussmann, J.; Koziol, F. *Angew. Chem., Int. Ed.* **2004**, *43*, 4485.
(5) Kussmann, J.; Ochsenfeld, C. *J. Chem. Phys.* **2007**, 127.
(6) Gao, Q.; Yokojima, S.; Kohno, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Fujihira, M.; Nakamura, S. *Chem. Phys. Lett.* **2007**, *445*, 331.
(7) Cui, Q.; Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 3721.
(8) Dedios, A. C.; Oldfield, E. *Chem. Phys. Lett.* **1993**, *205*, 108.
(9) Pickard, C. J.; Mauri, F. *Phys. Rev. B* **2001**, 6324.
(10) Neal, S.; Nip, A. M.; Zhang, H. Y.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215.
(11) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
(12) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251.
(13) Wu, W. X.; You, X. Z.; Dai, A. B. *Sci. Sinica Ser. B-Chem. Biol. Agric. Med. Earth Sci.* **1988**, *31*, 1048.
(14) Wang, B.; Brothers, E. N.; van der Vaart, A.; Merz, K. M. *J. Chem. Phys.* **2004**, *120*, 11392.
(15) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(16) Patchkovskii, S.; Thiel, W. *J. Comput. Chem.* **1999**, *20*, 1220.
(17) Williams, D. E.; Peters, M. B.; Wang, B.; Merz, K. M. *J. Phys. Chem. A* **2008**, *112*, 8829.
(18) Brothers, E. N.; Merz, K. M. *J. Phys. Chem. B* **2002**, *106*, 2779.
(19) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *233*, 231.
(20) Peters, M. B.; Williams, D. E.; Merz Jr., K. M. Unpublished results.
(21) Wang, B.; Raha, K.; Liao, N.; Peters, M. B.; Kim, H.; Westerhoff, L. M.; Wollacott, A. M.; van der Vaart, A.; Gogonea, V.; Suarez, D.; Dixon, S. L.; Vincent, J. J.; Brothers, E. N.; Merz, K. M., Jr. DivCon.
(22) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Wenger, R. K.; Yao, H. Y.; Markley, J. L. *Nucleic Acids Res.* **2008**, *36*, D402.
(23) Moon, S.; Case, D. A. *J. Biomol. NMR* **2007**, *38*, 139.
(24) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.
(25) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712.
(26) Stewart, J. J. P. *J. Mol. Model.* **2008**, *14*, 499.

JP9028722