JCTC Journal of Chemical Theory and Computation

Development of a Parametrized Force Field To Reproduce Semiempirical Geometries

Andrew M. Wollacott[†] and Kenneth M. Merz, Jr.*[,‡]

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received January 11, 2006

Abstract: Here we describe the development of a classical force field parameter set to reproduce the geometry of proteins minimized at the semiempirical quantum mechanical level. The overall goal of the development of this new force field is to provide an inexpensive, yet reliable, method to arrive at geometries that are more consistent with a semiempirical treatment of protein structures. Since the minimization of a large number of protein structures at the semiempirical level can become cost-prohibitive, a "preminimization" with an appropriately parametrized classical treatment could potentially lead to more computationally efficient methods for studying protein structures through semiempirical means. Here we demonstrate that this force field allows for more rapid and stable geometry optimizations at the semiempirical level and can aid in the adoption of quantum mechanical calculations for large biological systems.

Introduction

Although ab initio and Density Functional (DFT) methods have proven reliable in the modeling of chemical systems,[1,2] they cannot be routinely applied to larger biological systems as they scale poorly with the size of the system. Through various approximations in the quantum mechanical (QM) formulation more computationally feasible methods have been developed. One such approach is the semiempirical QM treatment of chemical systems.[3−5] These methods were developed in the late 1970s and 1980s to model smaller chemical systems at the quantum mechanical level. To account for the various approximations in the semiempirical QM treatment, these methods have been highly parametrized to reproduce experimental data. In fact, these methods have proven to model chemical systems reliably at accuracies rivaling higher-level treatments. Additionally, with the recent development of linear-scaling semiempirical methods it is now feasible to apply these methods to the study of protein systems.[6−14]

Semiempirical QM methods have proven valuable in the study of protein structures, and their usefulness has been demonstrated for related applications such as protein−ligand interactions.[15] While more costly than molecular mechanics-based (MM) methods, semiempirical calculations have been shown to outperform their classical counterparts in the discrimination of native structures from misfolded models (Wollacott and Merz, unpublished results). Semiempirical methods are also more sensitive to changes in protein geometry (Wollacott et al., unpublished results). The choice of starting structure for any modeling exercise is extremely important, and this requirement is exaggerated in the case of QM methods because they are more dependent on the internal geometry. To improve molecular structures before analysis, it is typical to first optimize the structure at the level of theory for which the model is being investigated. To speed up convergence, optimizations can be carried out in multiple steps, starting at lower levels of theory followed by an optimization at the desired level of theory so as to bring the model closer to the local energy minimum.

* Corresponding author e-mail: merz@qtp.ufl.edu.

[†] Current address: Department of Biochemistry, University of Washington, Seattle, WA.

[‡] Current address: Department of Chemistry, Quantum Theory Project, University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, Florida 32611-8435.

10.1021/ct0600161 CCC: $33.50 © 2006 American Chemical Society
Published on Web 06/16/2006

This is not usually necessary for classical methods such as employed in AMBER[16] but is frequently used with QM methods.

While a divide-and-conquer approach to semiempirical treatments can be used to minimize the energies of small proteins very rapidly relative to other QM methods,[6−9,14] it is currently infeasible to apply semiempirical minimizations to large-scale biological problems, such as ab initio protein folding simulations. For these systems, it would be desirable to first minimize structures with a fast classical potential before scoring or optimizing with semiempirical QM methods, thereby potentially reducing the computational expense. The assumption here is that the MM potential results in a structure more consistent with a semiempirical treatment. This procedure has been applied to the identification of native structures from sets of misfolded structures (Wollacott and Merz, unpublished results), with very promising results.

The feasibility of reparametrizing a classical force field, in this case AMBER, to better reproduce structures minimized at the semiempirical level has been investigated. The advantage of such a new parameter set would be 2-fold: (1) it would potentially speed up minimizations by utilizing an MM minimization to arrive at structures that are lower in energy with respect to a semiempirical treatment, and (2) it would reduce the overall strain on the system and potentially remove large instabilities during the minimization process that can lead to bond cleavage. For general protein minimization it is undesirable for amino acid groups to undergo bond rearrangement; in such applications the bonding configuration of residues should remain intact. Several aspects of the AMBER force field have been chosen for reparametrization, starting with the parm94 parameter set,[17] to reproduce the geometries of proteins minimized at the PM3[5] and AM1[4] levels. These new parameter sets have been named parmPM3 and parmAM1 for the respective Hamiltonian that they were parametrized against.

Bond lengths, bond angles, atomic charges, and Lennard-Jones parameters from the AMBER parm94 force field were reparametrized. An analysis of the proper and improper torsions in proteins minimized at the semiempirical level indicates that these values may not be optimal for proteins (Wollacott et al., unpublished results). Out of plane bending (controlled by improper torsions) and unfavorable rotameric states of side chains were noted as unfavorable artifacts of semiempirical QM minimizations. Thus, the torsion parameters from AMBER were retained as semiempirical QM methods poorly model the dihedrals, whereas classical methods are better suited to treat these terms.

It should be stressed that the purpose of the parmAM1 and parmPM3 parameter sets is not to yield geometries that are in better agreement with experimentally determined structures. Rather, the parmAM1 and parmPM3 sets have been developed to reproduce protein structures minimized using semiempirical QM methods, regardless of whether these geometries are more or less nativelike. The resulting structures can then be more reliably used in large-scale semiempirical calculations on biological systems.

Table 1. Protein Systems Comprising the Training Set for the Parameterization of ParmAM1 and ParmPM3

PDB ID	description	resolution (Å)	N_{res}
1A80	HIV capsid C-terminal domain	1.70	70
1AIL	N-Ter fragment of Ns1 protein	1.90	70
1B0X	Epha4 receptor tyrosine kinase	2.00	72
1BCG	scorpion toxin Bixtr-It	2.10	74
1BMG	bovine beta-2 microglobulin	2.50	98
1CEI	colicin E7 immunity protein	1.80	85
1CQY	starch-binding domain of *Bacillus beta-amylase*	1.95	99
1CSP	major cold shock protein	2.50	67
1DSL	beta crystallin (C-ter)	1.55	88
1EM7	helix variant of B1 domain from strep protein G	2.00	56
1ENH	engrailed homeodomain	2.10	54
1F0M	ephrin type-B receptor	2.20	71
1FAS	fasciculin 1 (toxin)	1.80	61
1FNA	fibronectin cell-adhesion module	1.80	91
1H75	glutaredoxin-like protein Nrdh	1.70	76
1HPT	human pancreatic secretory trypsin inhibitor	2.30	56
1HYP	hydrophobic protein from soybean	1.80	75
1KW4	polyhomeotic sam domain structure	1.75	70
1LPL	hypothetical 25.4 Kda protein	1.77	95
1MJC	major cold shock protein	2.00	69
1MWP	amyloid A4 protein	1.80	96
1OPS	type III antifreeze protein	2.00	64
1ORC	Cro repressor insertion mutant	1.54	64
1PWT	alpha spectrin SH3	1.77	61
1WHO	allergen Phl P 2	1.90	94
1R69	phage 434 repressor (N-ter)	2.00	63
1SN1	neurotoxin Bmk M1	1.70	64
1UBI	ubiquitin	1.80	76
2CRO	434 Cro protein	2.35	65
2OVO	ovomucoid third domain	1.50	56

Methods

To arrive at a set of parametrized values for the bond lengths, angles, atomic charges, and van der Waals parameters, a set of small proteins from the protein databank was collected, ranging in size from 600 to 1500 atoms (Table 1). Due to the computational expense associated with optimization at the semiempirical level, the training set was limited to 30 small proteins. These proteins were selected to obtain a fairly representative set of topological features, including structures with secondary structural content composed of all α-helices, all β-sheets, a mix of helices and sheets, and random coils. All structures were solved using X-ray crystallography, ranging in resolution from 1.54 Å to 2.5 Å, and contained no cofactors or metal ions. The protein systems used for the training set are listed in Table 1. Hydrogen atoms were added to all proteins using the LEaP module of AMBER (AMBER 8.0). Hydrogen atoms were minimized using the Sander package from AMBER with the parm94 force field for 300 steepest descent steps, followed by 700 conjugate gradient steps. These protein systems were then minimized with either the PM3 or AM1 Hamiltonians using conjugate gradient as the minimization protocol. The resulting optimized structures were chosen as targets for the parametrization. In some cases hydrogen atom transfer reactions occurred between charged

Table 2. Amino Acid Frequencies in the Training Set of Proteins

amino acid	frequency	percentage (%)
ALA	130	6.18
CYS	56	2.66
ASP	122	5.80
GLU	137	6.51
PHE	73	3.47
GLY	160	7.60
HIS	31	1.47
ILE	110	5.23
LYS	152	7.22
LEU	154	7.32
MET	53	2.52
ASN	98	4.66
PRO	96	4.56
GLN	96	4.56
ARG	107	5.09
SER	125	5.94
THR	139	6.61
VAL	155	7.37
TRP	31	1.47
TYR	79	3.75

groups during optimization in vacuo. These were fixed by removing and then rebuilding the transferred hydrogen atoms using AMBER,[16] followed by a short optimization of only the rebuilt hydrogen atom coordinates.

In order for the training set of protein systems to be used as targets for parametrization, there should be an adequate representation of each type of amino acid to obtain reliable statistics. Although only 30 proteins were used, the majority of amino acids were well represented, as shown by the frequency of residue types in the training set in Table 2. Cysteine residues were found as either single residues or as part of disulfide bridges, although the two forms were not distinguished between when parametrizing the bond length and angle values. In general, the frequencies of amino acids in the training set are similar to those found across the protein database.[18]

In developing the parmPM3 and parmAM1 parameter sets for AMBER, the atomtype designations used in parm94 were retained. Since parmPM3 and parmAM1 were only parametrized against proteins, only those atomtypes found in protein systems were included in the parametrization. While current versions of these parameter sets are applicable only for proteins, all other parameters were kept unaltered from their parm94 values, retaining the ability to model other biologically relevant molecules.

Reparametrizing Bond Lengths and Angles. The parameters for bond lengths and bond angles were taken as the average found in the training set of minimized structures for bonds between atoms of designated AMBER atomtypes. The force constant for each bond length and angle (K_{eq}) was taken from the AMBER parm94 parameter set, with only the equilibrium value of the lengths and angles (R_{eq}) modified to match the average from the target set. In general the difference in internal geometries between AM1 and PM3 minimized structures is small. The frequency of different bond types found in the protein systems varied considerably, with a large number of peptide and aliphatic bonds being represented in the database of bond lengths, but only few bond types specific to underrepresented amino acid side chains such as tryptophan and cystine. However, a comparison of the underrepresented bond types to those found in a large set of pentapeptides (10 000 of sequence GGXGG where X is any amino acid) minimized at the semiempirical level revealed very small differences between the two geometries. Furthermore, the deviation in bond lengths across systems was limited. Thus, undersampling of bond lengths and angles does not seem to be a major problem for this data set. For the case of disulfide bonds, there were very few S−S bonds in the training set, so this bond type was not included in the parametrization. The values for bond lengths and angles for the parmPM3 and parmAM1 parameter sets are listed in the Supporting Information in Tables S1 and S2.

Parametrizing Atomic Charges. To parametrize the charges for each amino acid, the average CM2[19] charge on each atom in the protein training set was determined from the semiempirical calculation. These charges were taken from vacuum calculations, although the differences in average atomic charges from in vacuo or solvation calculations is small (Wollacott et al., unpublished results). In comparison, the atomic charges used in the parm94 set were derived by fitting with a restrained ESP-fit (RESP) model.[20]

Since a semiempirical treatment allows for charge transfer to occur, the average charge on each residue did not sum to an integral value. The charge for each atom in a residue was, therefore, normalized via the scaling of charges such that the residue possessed an integral charge. In addition, charges were averaged for chemically equivalent groups. For example, the three hydrogens in the methyl group of alanine (H^{β}'s) possessed slightly differing charges, so the charge assigned to each atom was taken as the average of the three. Charges for the cysteine residue were not reparametrized because the residue could be found in disulfide bridges, which affected the charge distribution. While the parm94 force field differentiates between cysteine (CYS) and cystine (CYX), the limited number of cysteines in the training set prevented their inclusion in the reparametrization effort. In this case, charges from the parm94 set were used for both forms of cysteine. The charges derived for the parmPM3 and parmAM1 parameter sets are listed in the Supporting Information in Table S3.

Reparametrization of Lennard-Jones Parameters. The shape of the Lennard-Jones potential can be specified by two variables, the R^* value and the ϵ value.[17] The R^* parameters define the closest approach of two atoms, while the ϵ parameters describe the well depth. These parameters were derived for the parm94 parameter set by fitting to Monte Carlo liquid simulations to reproduce the densities and enthalpies of vaporization for hydrocarbons.[21]

For the development of the parmAM1 and parmPM3 parameters, the van der Waals parameters were adjusted such that minimization with AMBER using these parameters would best reproduce protein conformations generated by minimization at the AM1 or PM3 level. To accomplish this would require a parametrization scheme whereby van der Waals parameters are initially modified from their parm94 values, followed by an AMBER minimization of the crystal structures with the new force field and an evaluation of the RMSD of these minimized models to the target. Parameters that minimized the RMSD of the AMBER minimized structures compared to the semiempirical-minimized structures would then be accepted. On current computer hardware (2.4 GHz AMD Opteron), an AMBER minimization of a small protein can take up to 30 minutes. With 30 proteins in the training set, and 20 van der Waals parameters that must be optimized, this parametrization scheme quickly becomes difficult even on modern hardware.

To parametrize van der Waals parameters for the parmAM1 and parmPM3 parameter sets, the *R** Lennard-Jones parameters were modified such that the AMBER energy and gradients were reduced for the semiempirical-minimized target proteins. This has the advantage that a full AMBER minimization does not have to be performed for each step of the parametrization. However, adjusting the *R** values so as to reduce the AMBER energy of these semiempirical-minimized proteins does not guarantee that a minimization with these parameters will reach the target geometries. With this limitation in mind, the *R** values for the 20 protein-relevant atomtypes were parametrized using a genetic algorithm[22] to yield lower energies for structures that have been minimized at the semiempirical level. The ϵ parameters were not, however, modified from their parm94 values. The van der Waals parameters derived for the parmAM1 and parmPM3 parameter sets are listed in the Supporting Information in Table S4.

Results and Discussion

Minimizing with parmAM1 and parmPM3. Optimizing protein geometries with either parmAM1 or parmPM3 results in structures that are similar to those minimized with parm94. The average all-atom RMSD compared to the crystal structure after optimization with parm94 was 0.73 Å, with parmPM3 was 0.83 Å, and with PM3 was 0.95 Å. The all-atom RMSD between parm94 minimized structures and parmPM3 structures was smaller, on average 0.47 Å. Minimizing with parmAM1 lead to structures that had an average RMSD of 0.81 Å compared to the crystal structure. Structures minimized with parmAM1 were very similar in overall topology to those structures minimized with parmPM3 (RMSD 0.3 Å). In general, the parmAM1 and parmPM3 parameter sets are similar, so it is not surprising that when optimized under the same potential energy function they result in comparable structures.

Minimizing with parmAM1 or parmPM3 maintained many of the favorable traits of AMBER models, such as reducing large side-chain motions that form salt bridges. This effect can be most likely attributed to the inclusion of implicit solvation with AMBER minimizations, that were missing with semiempirical optimizations. When minimizing with MM force fields, planar groups retained their planarity, which was problematic when minimizing with semiempirical methods for nitrogen-containing groups such as the guanidyl group of arginine. The MM treatment of these groups explicitly enforces improper torsions to maintain planarity.

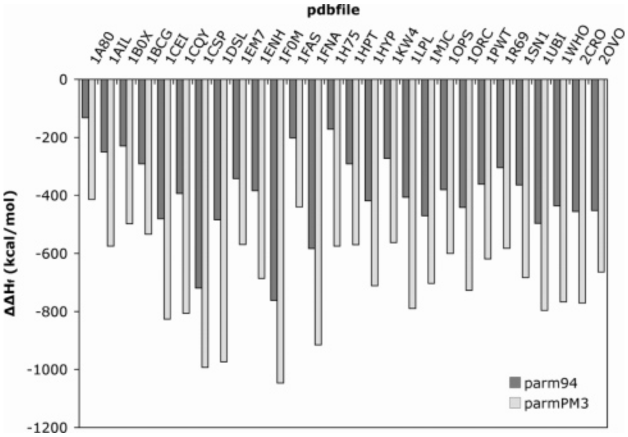

Figure 1. Improvements in the calculated heat of formation for structures minimized with parm94 and parmPM3 relative to their crystal structures for the training set.

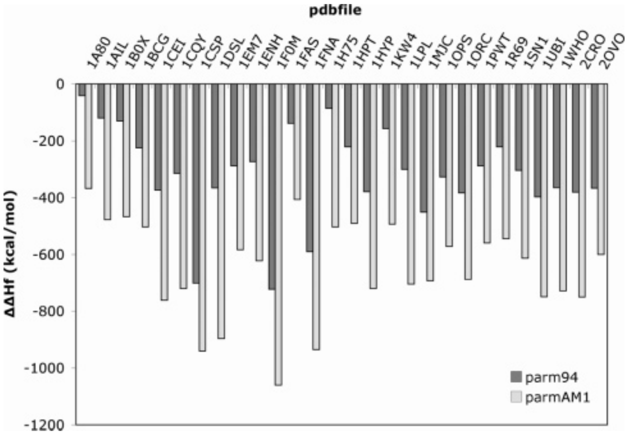

Figure 2. Improvements in the calculated heat of formation for structures minimized with parm94 and parmAM1 relative to their crystal structures for the training set.

The improvement of preminimizing with MM force fields can be seen by the reduction in the heats of formation for proteins in the training set, as illustrated in Figures 1 and 2. Minimizing a structure with AMBER using the parm94 parameter set results in heats of formation that are lower by an average 397 kcal/mol compared to the crystal structure. This represents a significant improvement over the starting structure, resulting from improved internal geometries and the reduction of atomic clashes. Minimizing using parmPM3 improves the heat of formation by over 692 kcal/mol compared to the starting structure, while parmAM1 improves the heat of formation by approximately 649 kcal/mol. Clearly, a rapid minimization using the parmAM1 or parmPM3 parameter sets improves the structure with respect to semiempirical QM treatments of the protein.

The average force on each atom in a structure is another telling feature of its quality with respect to the potential used. The gradient on each atom is calculated as the first derivative of the energy potential with respect to the Cartesian coordinates. The gradients as evaluated by DivCon are performed numerically, and the average gradient (GAVRG)

Table 3. Summary of Improvements in the Average Heat of Formation ($\Delta\bar{H}_f$) and the Mean of the Average Gradient ($\mathrm{GAVRG}_{\mathrm{avrg}}$) for Structures Minimized with parm94, parmAM1, and parmPM3 Compared to the Crystal Structures

	crystal	parm94	parmAM1	parmPM3
$\Delta\bar{H}_f$ (kcal/mol)	0.0	−391.9	−648.6	−692.6
$\mathrm{GAVRG}_{\mathrm{avrg}}$ (kcal/mol Å)	14.1	13.0	6.5	5.1

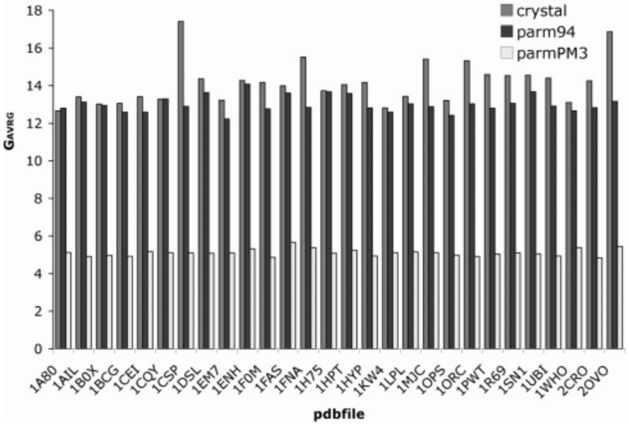

Figure 3. Average atomic gradients for crystal structures, structures minimized with parm94, and structures minimized with parmPM3 for the training set.

is calculated as shown in eq 1

$$\mathrm{GAVRG} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial E}{\partial r_i} \qquad (1)$$

where N is the total number of atoms, and $\partial E/\partial r_i$ is the derivative of the total energy with respect to the Cartesian coordinates.

There is a marked improvement in the average GAVRG for protein structures after minimization with the parmAM1 and parmPM3 parameters (Table 3). The mean of the initial average gradient for the crystal structure is 14.1 kcal/molÅ, with improvements after optimizing with the MM force fields parm94 (13.0 kcal/molÅ), parmAM1 (6.5 kcal/molÅ), and parmPM3 (5.1 kcal/molÅ). Thus, structures minimized with parmAM1 and parmPM3 are less strained according to the semiempirical calculations, as shown in Figure 3.

The observed improvements in the heats of formation and atomic forces are due primarily to the reparametrization of the bond lengths and angles. Semiempirical treatments were found to be very sensitive to the optimal bond geometry. Despite the improvements, since these minimizations were carried out with a classical force field, the method is still restricted by the limitations in MM modeling such as the use of a fixed point-charge model.

These results illustrate the utility of the parmPM3 parameters to allow proteins to be rapidly minimized with an MM potential before being evaluated at the semiempirical level. By significantly reducing the heat of formation of the protein systems and the atomic forces, preminimizing with parmAM1 or parmPM3 significantly improves the stability of systems before use in QM calculations.

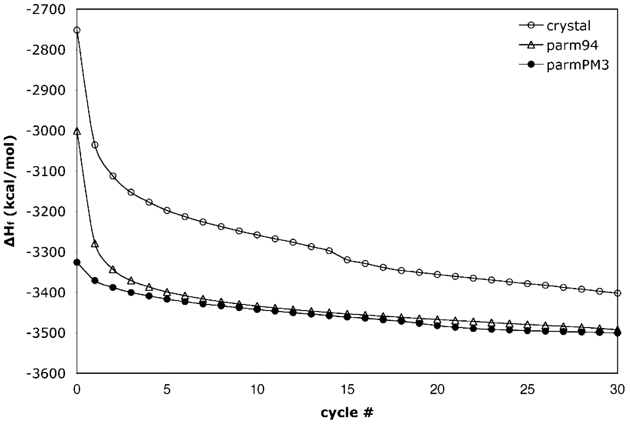

Figure 4. Minimization of 1AIL with PM3 using steepest descent for the crystal structure and the structures minimized with parm94 and parmPM3.

In general, the improvements seen by minimizing with parmAM1 or parmPM3 before scoring with their respective Hamiltonian are comparable. Optimization with parmAM1 leads to structures that are more consistent with the AM1 Hamiltionian, and the same is true for minimizing with parmPM3 in relation to the PM3 Hamiltonian. Since the observed trends and results for both parameter sets mirror each other so closely, we have focused here primarily on the improvements obtained with parmPM3, although the general results and their interpretations also hold true for parmAM1.

Improvements during Minimization. Figure 4 shows a minimization profile for a select protein system in the training set (1AIL). PM3 minimizations were performed for 30 steps using steepest descent starting with either the crystal structure or the structures minimized with parm94 or parmPM3. As shown, there is an absence of a steep drop-off in energy when minimizing the structure that had been preminimized with parmPM3. The energies of the structures preminimized with parm94 and parmPM3 converge to similar values, with structures that are close in overall conformation. A similar trend is seen across the systems in the training set, although in several cases, while the initial heat of formation for structures minimized with parmPM3 is lower than those minimized with parm94, the order of final heats of formation is reversed. This trend is seen for the 1ENH protein system and shown in Figure 5. In general, the parm94 and parmPM3 preminimized structures converge to similar heats of formation upon limited minimization with DivCon.

To investigate the stability of starting structures relative to a semiempirical treatment, structures were minimized using the LBFGS routine instead of steepest descent. The LBFGS minimization scheme reaches a local minimum structure faster than using steepest descent or conjugate gradient techniques, but the starting structure should be close to the local minimum before invoking LBFGS. Figure 6 shows the minimization profile for 1ENH taking starting structures as the crystal structure and structures preminimized with parm94 and parmPM3. As illustrated, in the early stages of minimization of the crystal structure and parm94 preminimized structures, the energy increases before the steep drop-off in energy. The parmPM3 preminimized structure,

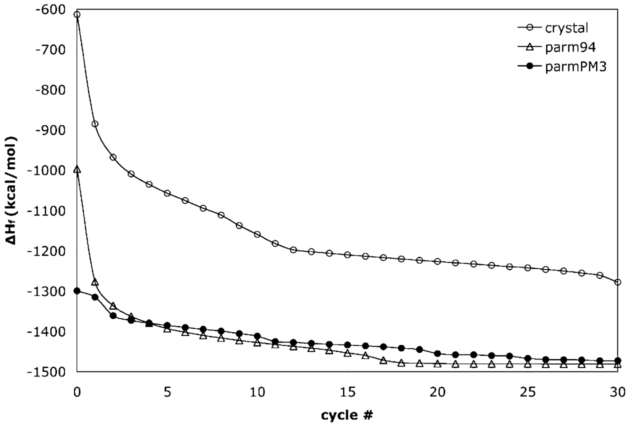

Figure 5. Minimization of 1ENH with PM3 using steepest descent for the crystal structure and the structures minimized with parm94 and parmPM3.

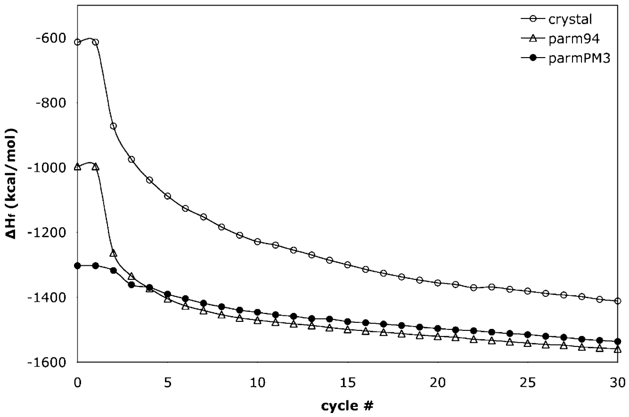

Figure 6. Minimization of 1ENH with PM3 using LBFGS for the crystal structure and the structures minimized with parm94 and parmPM3.

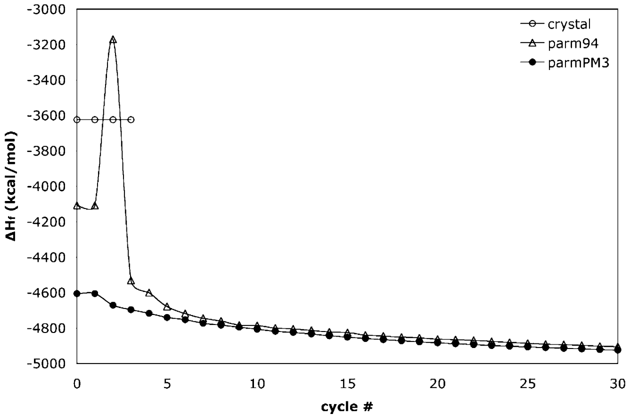

Figure 7. Minimization of 1DSL with PM3 using LBFGS for the crystal structure and the structures minimized with parm94 and parmPM3.

however, exhibits a more stable minimization profile and again demonstrates only a relatively small total decrease in heat of formation upon minimization.

The LBFGS minimization of the 1DSL structure exhibits exaggerated instabilities as seen in Figure 7. In this case, the LBFGS optimization becomes unstable for the crystal structure and is terminated prematurely as the energy continues increasing. The minimization of the parm94 preminimized structure exhibits a large increase in energy early in the minimization, while the parmPM3 preminimized structure again shows a stable minimization profile. The energy spike for the parm94 preminimized structure is caused by the large initial forces on the atoms, created by an unfavorable starting geometry. This is illustrated in Figure 8 for the Arg 13 residue of 1DSL. The large forces cause the Cartesian coordinates of the atoms to move by too great a distance, leading to elongated bond lengths and angles in one minimization step, destabilizing the system, and increasing the heat of formation by over 900 kcal/mol. Since the geometry of the parmPM3 preminimized structure is more consistent with a semiempirical approach, the atomic forces for the structure are lower in overall magnitude and so the atomic positions do not vary as much and are more stable during minimizations.

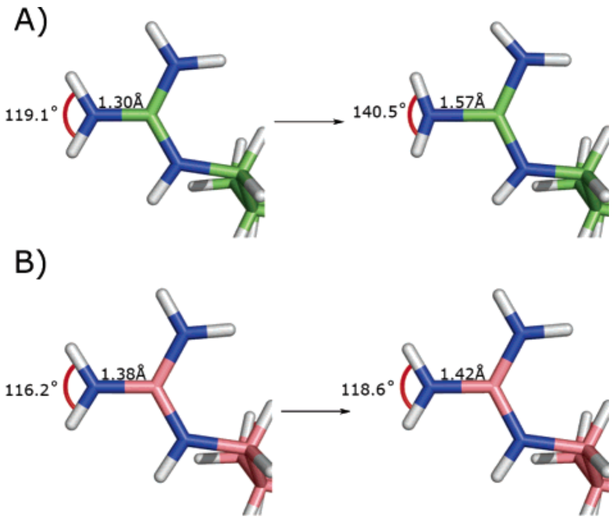

Figure 8. Structure of LYS-13 of 1DSL during LBFGS minimization with parm94 and parmPM3 preminimized structures. (A) Preminimizing with parm94 results in structures with elongated bond lengths and large angles for arginine. (B) Preminimizing with parmPM3 results in more stable bond lengths and angles during minimization.

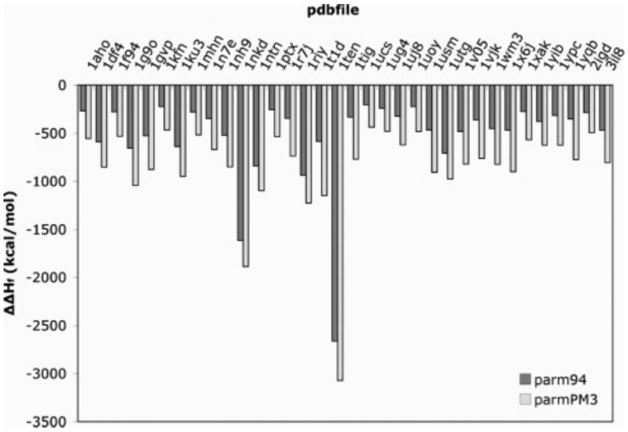

Figure 9. Improvements in the heat of formation for structures minimized with parm94 and parmAM1 relative to their crystal structures for the test set.

These results highlight the ability of the parmAM1 and parmPM3 force fields to clean up structures for subsequent semiempirical studies. ParmAM1 and parmPM3 premini-

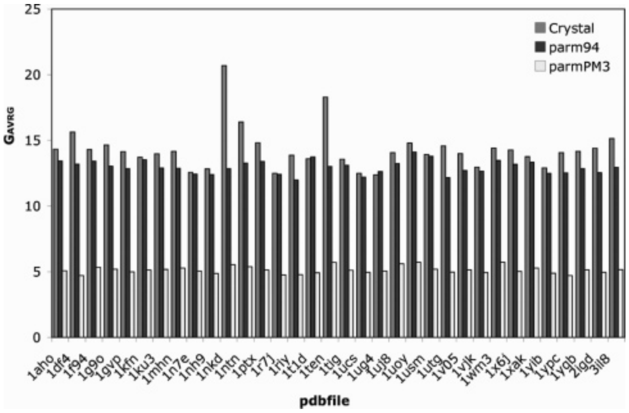

Figure 10. Average atomic gradients for crystal structures, structures minimized with parm94, and structures minimized with parmPM3 for the test set.

mized structures not only score much lower with respect to heats of formation but also result in less strained structures that are better behaved during the minimization process.

Evaluating the Extensibility of parmAM1 and parmPM3 Parameters. To assess the extensibility of parmAM1 and parmPM3 beyond the training set of proteins, a test set of 34 small proteins structures that had been solved using X-ray crystallography was chosen. As with the training set, the test set listed in Table 4 covers a range of topological features. The reduction in heats of formation and the average gradient for this data set are shown in Figures 9 and 10. As with the test set, the parmPM3 minimized structures have heats of formation over 300 kcal/mol lower than parm94 minimized structures on average, and the average gradient is cut by over 50% for parmPM3 minimized structures. Again, structures preminimized with parmPM3 were more stable during semiempirical minimizations. Overall this illustrates the general applicability of the parmPM3 force field to studying small proteins using semiempirical QM approaches.

Table 4. Protein Systems Comprising the Test Set for the Evaluation of parmAM1 and parmPM3

PDB ID	description	resolution (Å)	N_{res}
1AHO	scorpion toxin	0.96	64
1DF4	HIV-1 envelope glycoprotein	1.45	68
1F94	bucandin toxin	0.97	63
1G9O	Pdz domain	1.50	91
1GVP	gene V protein	1.60	87
1KFN	major outer membrane	1.65	56
1KU3	RNA polymerase σ subunit	1.80	73
1MHN	SMN tudor domain	1.80	59
1N7E	sixth PDZ domain of Grip1	1.50	97
1NH9	DNA binding protein Mja10B	2.00	87
1NKD	Cole1 repressor of primer	1.07	65
1NTN	neurotoxin-I	1.90	72
1PTX	scorpion toxin II	1.30	64
1R7J	DNA binding protein Sso10A	1.47	95
1RIY	histone-like DNA binding protein	1.80	90
1TLD	shaker potassium channel	1.51	100
1TEN	fibronectin type III domain	1.80	90
1TIG	translation initiation factor	2.00	94
1UCS	type III antifreeze protein Rd1	0.62	64
1UG4	carditoxin VI	1.60	60
1UJ8	hypothetical protein	1.75	77
1UOY	bubble protein	1.50	64
1USM	protein-binding transcriptional coactivator	1.20	80
1UTG	oxidized uteroglobin	1.34	70
1V05	human filamin	1.43	96
1VJK	molybdopterin converting factor	1.51	98
1WM3	human sumo-2 protein	1.20	72
1X6J	hypothetical protein Yfgy	2.00	91
1XAK	sar-coronavirus Orf7A accessory protein	1.80	83
1YIB	microtubule-associated protein Rp/Eb	1.80	76
1YPC	chymotrypsin inhibitor 2	1.70	64
1YQB	ubiquilin 3	2.00	100
2IGD	protein G IgG-binding domain III	1.10	61
3IL8	interleukin 8	2.00	72

Conclusion

Minimizations of proteins at the semiempirical level are time-consuming, even when utilizing linear scaling approaches. In addition to their computational expense, semiempirical calculations are more sensitive to the initial conformation of the starting structure. Since the internal geometries of atoms in X-ray and NMR structures are different from those found in structures minimized at the semiempirical level, many atoms experience large initial forces during minimization. Without the constraint of explicit bonds in the quantum mechanical treatment of the structure, this can lead to undesirable features during the minimization process such as bond cleavage. From these results, it appears that both the parmAM1 and parmPM3 parameter sets are a valuable addition to the semiempirical treatment of proteins. By creating structures that are more geometrically consistent with proteins minimized at the semiempirical level, these structures score and behave better in semiempirical QM calculations. This approach is suitable for use in QM/MM calculations where the QM region is treated at the semiempirical level as well as in QM studies of, for example, protein−ligand interactions.

The parmAM1 and parmPM3 parameter sets represent a fast and effective preminimization step for semiempirical quantum mechanical calculations. By providing a consistent approach to removing strain in the protein, these new parameter sets allow for subsequently more reliable calculations using semiempirical QM methods.

Abbreviations Used. Quantum mechanics, QM; molecular mechanics, MM.

Acknowledgment. We thank Kenneth Ayers for his assistance with managing the computational resources and for valuable discussions. We would also like to thank Duane Williams for his input and proofreading of this manuscript. We thank the NSF (MCB-0211639) and the NIH (GM 44974) for support.

Supporting Information Available: Parameters for parmAM1 and parmPM3; bond length, bond angle, Lennard-Jones, and charge parameters; the parameter sets in formats consistent with the AMBER molecular modeling package (parmAM1.dat and parmPM3.dat) as well as files containing the charge information (all_aminoAM1.in and all_aminoPM3.in). This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kohn, W.; Becke, A. D.; Parr, R. G. Density functional theory of electronic structure. *J. Phys. Chem.* **1996**, *100* (31), 12974−12980.

(2) Hehre, W. J.; Radom, L.; Pople, J. A.; Schleyer, P. V. R. *Ab Initio Molecular Oribital Theory*; Wiley-Interscience: 1986.

(3) Dewar, M. J. S.; Thiel, W. Ground States of Molecules. 38. The MNDO method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99* (15), 4899−4907.

(4) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(5) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 1. Method. *J. Comput. Chem.* **1989**, *10* (2), 209−220.

(6) Dixon, S. L.; Merz, K. M., Jr. Semiempirical Molecular Orbital Calculations with Linear System Size Scaling. *J. Chem. Phys.* **1996**, *104*, 6643−6649.

(7) Dixon, S. L.; Merz, K. M., Jr. Fast, Accurate Semiempirical Molecular Orbital Calculations for Macromolecules. *J. Chem. Phys.* **1997**, *107*, 879−893.

(8) van der Vaart, A.; Gogonea, V.; Dixon, S. L.; Merz, K. M., Jr. Linear Scaling Molecular Orbital Calculations of Biological Systems Using the Semiempirical Divide and Conquer Method. *J. Comput. Chem.* **2000**, *21*, 1494−1504.

(9) van der Vaart, A.; Suarez, D.; Merz, K. M. Critical assessment of the performance of the semiempirical divide and conquer method for single point calculations and geometry optimizations of large chemical systems. *J. Chem. Phys.* **2000**, *113* (23), 10512−10523.

(10) Raha, K.; Merz, K. M., Jr. A Quantum Mechanics Based Scoring Function: Study of Zinc-ion Mediated Ligand Binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020−1021.

(11) Yang, W.; Lee, T.-S. A Density-matrix Divide-and-conquer Approach for Electronic Structure Calculations of Large Molecules. *J. Chem. Phys.* **1995**, *103* (13), 5674−5678.

(12) Lee, T. S.; York, D. M.; Yang, W. Linear-scaling semiempirical quantum calculations for macromolecules. *J. Chem. Phys.* **1996**, *105*, 2744−2750.

(13) Stewart, J. J. P. *MOPAC2000*; Fujitsu Ltd.: Tokyo, 1999.

(14) Daniels, A. D.; Milliam, J. M.; Scuseria, G. E. Semiempirical Methods with Conjugate Gradient Density Matrix Search to replace Diagonalization for Molecular Systems Containing Thousands of Atoms. *J. Chem. Phys.* **1997**, *107*, 425−431.

(15) Raha, K.; Merz, K. M., Jr. Calculating Binding Free Energy in Protein−ligand Interaction. *Ann. Reports Comput. Chem.* **2005**, *1*, 113.

(16) Case, D. A.; Darden, T. A.; Cheatham, I. T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8.0*; 2004.

(17) Cornell, W. D.; Cieplak, P.; Baylay, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(18) Gilis, D.; Massar, S.; Cerf, N. J.; Rooman, M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol.* **2001**, *2* (11), 1−12.

(19) Li, J. B.; Zhu, T. H.; Cramer, C. J.; Truhlar, D. G. New Class IV Charge Model for Extracting Accurate Partial Charges from Wave Functions. *J. Phys. Chem.* **1998**, *102*, 2 (10), 1820−1831.

(20) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges − the Resp Model. *J. Phys. Chem.* **1993**, *97*, 7 (40), 10269−10280.

(21) Jorgensen, W. L.; Pranata, J. Importance of Secondary Interactions in Triply Hydrogen-Bonded Complexes − Guanine-Cytosine Vs Uracil-2,6-Diaminopyridine. *J. Am. Chem. Soc.* **1990**, *112* (5), 2008−2010.

(22) Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: San Mateo, CA, 1989.

CT0600161