Force Field Validation Using Protein Side Chain Prediction

Matthew P. Jacobson,*,† George A. Kaminski, and Richard A. Friesner

Department of Chemistry, Columbia University, New York, New York 10027

Chaya S. Rapp

Department of Chemistry, Stern College, Yeshiva University, New York, New York 10016

Received: July 8, 2002; In Final Form: August 8, 2002

The prediction of protein side chain conformations is used to evaluate the accuracy of force field parameters. Specifically, new torsional parameters have recently been reported for the OPLS-AA force field, which achieved substantially better accuracy with respect to high level gas-phase quantum chemical calculations [*J. Phys. Chem. B* **2001**, *105*, 6474]. Here we demonstrate that these new parameters also lead to qualitatively improved side chain prediction accuracy. The primary emphasis is on the prediction of *single* side chain conformations, with the rest of the protein held fixed at the native configuration. Errors due to incomplete sampling can thus be essentially eliminated, using a combination of rotamer search and energy minimization. In addition, the protein environment is modeled realistically using implicit solvation and an explicit representation of crystal packing effects. Aided by the development of new algorithms, these calculations have been performed with modest computational requirements (a cluster of PCs) on a database of 36 proteins (∼5000 total residues). The side chain prediction tests that we employ are quite general and can be used to evaluate nonbonded or solvation parameters as well. As such, they provide a useful complement to decoy studies for force field validation.

I. Introduction

The development of an accurate molecular mechanics force field for protein modeling is a critical challenge for computational molecular biology. In principle, an accurate atomic resolution protein force field and model of aqueous solvation, in conjunction with robust sampling methods, should be capable of selecting the experimentally observed protein structure as the one that has the lowest free energy in a protein simulation. However, it is not yet possible in practice to realize this objective on a routine basis. There are several major difficulties:

1. Inaccuracies in the protein force field,
2. Errors in the treatment of aqueous solvation, and
3. Inadequate sampling of conformational space.

In practice, these problems are coupled. Evaluation of the quality of the force field and solvation model (as well as improvement of the models based on comparisons with experimental structures) has been hindered by difficulties associated with exploring adequately the large number of local minima that characterize atomic-level protein potential energy surfaces.

A common method of force field evaluation is the use of decoy sets to test the ability of energy functions to distinguish native from non-native folds (e.g., refs 1−4). Here we employ a different strategy for force field evaluation, based on the prediction of side chain conformations, which we believe will be particularly valuable for evaluating the suitability of energy functions for high resolution structural predictions. The principal focus of this paper is the prediction of single side chain conformations (i.e., keeping the remainder of the protein fixed at the native), as this is the least computationally intensive task which provides a nontrivial evaluation of force field quality. The primary precedent for interrogating the energy function in this way is work by the Karplus group.[7,8] Our new contributions include more extensive conformational sampling (all side chain torsional angles, and in fact all atoms, are free to move, not just the first two torsional angles), a much larger test set of proteins, and, especially, the systematic evaluation of alternate energy functions. We also perform a somewhat less restrictive side chain prediction test, which involves optimizing the conformations of all side chains on a loop and then energy minimizing the entire loop. Adequate sampling can be performed for this test with modest computational effort, and it provides additional information about the accuracy of side chain and backbone force field parameters.

Our specific objective in the present paper is the evaluation of several variants of one protein force field, the OPLS-AA force field of Jorgensen and co-workers.[5] Recently, new torsional parameters were presented for the OPLS-AA force field in which systematic refitting was carried out for all of the amino acids. Specifically, qualitatively better agreement was obtained with high level gas-phase quantum chemical calculations for several amino acid dipeptides.[6] However, at that time, it was not possible to verify that the new parameters actually provide better results in biologically relevant calculations (i.e., proteins in the condensed phase). The data in this paper unambiguously demonstrate that the new parameters indeed reduce errors in protein structural prediction in a realistic condensed phase environment. This result not only highlights the powerful role that gas phase quantum chemistry calculations can play in force field development but also suggests that further refinement of protein force fields and solvation models may be achievable, through multiple cycles of parameter refinement and validation against large experimental data sets of protein structures. We

* To whom correspondence should be addressed.
† Current address: UCSF Department of Pharmaceutical Chemistry, Box 0446, San Francisco, CA 94143-0446. E-mail: matt@cgl.ucsf.edu.

10.1021/jp021564n CCC: $22.00
Published on Web 10/15/2002

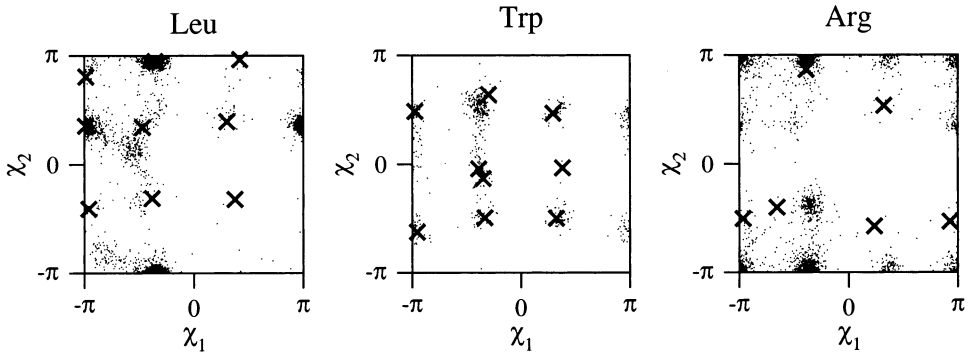

Figure 1. Local minima on the dipeptide quantum chemical potential energy surfaces (crosses) which correlate well with observed side chain conformations (dots). The experimental results are taken from 300 high quality X-ray crystal structures, chosen from a "culled PDB" list compiled by Dunbrack.[19] There are 4336 Ile conformations, 708 Trp conformations, and 2262 Arg conformations represented. In the case of Arg and other charged residues, the local minima on the dipeptide potential energy surface were located using a continuum solvent environment (Self-Consistent Reaction Field; SCRF); all other quantum chemical calculations were performed for isolated gas-phase molecules. The periodicity of both the χ_1 and χ_2 axes should be considered when interpreting these plots.

believe that this approach will play a critical role in overcoming many of the challenges that have confronted the development of force fields/solvation models that can be used reliably for high-resolution structural refinement.

The achievable side chain prediction accuracy is dependent upon several factors other than the protein force field. First, we have found that it is essential to model explicitly the crystalline environment to obtain quantitatively meaningful predictions with X-ray crystal structures for many of the solvent-exposed side chains on the protein exterior. Second, to achieve computational tractability, we employ a continuum solvation model, using the surface generalized Born (SGB) approach.[9] Details of the solvation model (including treatment of bound waters) have a substantial effect on the results. In this paper, we briefly discuss modeling of both the crystalline and aqueous environments, leaving however a more extensive exploration of these topics to other publications.

The paper is organized as follows. In section II, we provide a summary of the OPLS-AA protein force field variants, as well as brief discussions of our new sampling, solvation, and crystalline modeling approaches. Section III presents results for side chain prediction. Finally, in section IV, the conclusion, we discuss future directions.

II. Computational Models and Methods

A. OPLS-AA Force Field. The OPLS-AA force field is a fixed charge molecular mechanics force field developed over a period of many years in the research group of Jorgensen and co-workers. The parameters in the original version of OPLS-AA[5] were developed as follows:

1. Nonbonded parameters (charges, van der Waals radii) were obtained by fitting parameters for small molecules to reproduce liquid state thermodynamic data (heats of vaporization, densities) and then transferring these parameters, via an atom typing scheme, to the appropriate protein atoms.

2. Stretching and bending parameters were taken from the AMBER force field.[10]

3. Torsional parameters were fit to quantum chemical data on small model systems, typically at the HF/6-31G* level of theory, which provides a reasonably good description of conformational energetics.

Recently, our laboratory has collaborated with the Jorgensen group in developing a new version of OPLS-AA for peptides.[6] In this version, the stretching, bending, and nonbonded parameters were generally retained from the previous version (with a few exceptions, indicated below).

The primary modification to the force field involved refitting key backbone and side chain torsional parameters to a large data set of quantum chemical energies obtained for various conformational states of dipeptides. Geometry optimization of the dipeptides was carried out at the HF/6-31G** level followed by single-point energy calculations at the LMP2/cc-pVTZ (-f) level; this approach has been shown to produce substantial improvements in relative conformation energetics of small organic molecules as compared to HF/6-31G* calculations.[11] For charged amino acids, geometry optimization was carried out using continuum solvent rather than in the gas phase, thus focusing fitting effort on the relevant portion of conformational space. In addition to enumerating the minima of the dipeptide, energy scans along key torsional coordinates were carried out (within $\pm 40°$ of the local minima), yielding a data set of $\sim$2000 points on the various dipeptide energy surfaces.

Although the quantum chemical calculations were extensive, it is not obvious that the resultant sampling of the dipeptide potential energy surfaces, with points centered around the quantum chemical local energy minima as described above, should necessarily correlate well with those portions of the side chain conformational space that are observed most frequently in protein structures. In Figure 1, we plot the local minima obtained from the quantum chemical calculations (crosses) along with a representative sampling of observed side chain conformations (dots) for three amino acids: Leu (4336 conformations), Trp (708 conformations), and Arg (2262 conformations). The experimental results are obtained from 300 high quality ($<$2.0 Å resolution; R value $<$0.2) X-ray crystal structures, chosen from a "culled PDB" list (maximum sequence identity 30%) compiled by Dunbrack.[19] The local minima of the quantum chemical potential energy can be seen to correlate rather well with the observed clustering ("rotamers") of the observed side chain conformations (the periodicity of the χ_1 and χ_2 axes should be kept in mind when interpreting these plots). There are a few, high energy quantum chemical local minima which lack a significant number of experimental counterparts (e.g., the cross in the lower right quadrant of the Leu plot). The correlation between the quantum chemical and observed conformations is somewhat weaker for the charged Arg residue, but the quantum chemical points still provide a reasonably good sampling of the χ_1/χ_2 conformational space (13 single-point energy calculations were performed in the vicinity, $\pm 40°$, of each local minimum).

For five of the amino acids, multiple new sets of parameters were reported. For Leu and Val, one parameter set was constructed to fit both of the dipeptide data sets simultaneously, and two others were constructed to fit each dipeptide individually. For Ser and Thr, parameter sets were generated with only χ_1 parameters refitted and with both χ_1 and χ_2 refitted (i.e., refitting parameters involving the hydroxyl hydrogen as well). Finally, two parameter sets were developed for Asp, because of concerns about overfitting the data. That is, in the original fit, one of the torsional parameters became very large, raising suspicions that it might be unphysical, and a second fit restricted this parameter to a lower value. The present results provide an opportunity to test each of these alternative parameter sets and understand better the tradeoffs involved in the different fitting strategies.

Finally, the nonbonded parameters for sulfur (in Cys and Met) were also modified in the new version of OPLS-AA. Although the liquid state properties of small molecule analogues (dimethyl sulfide, methane thiol) yielded good agreement with experimental data with the earlier parameters, the gas phase hydrogen bond energy using the original parameters was substantially overbound. By increasing the dispersion terms for sulfur and decreasing its partial atomic charge, a new parameter set was developed which also fit the liquid-state properties quite well but provided a more accurate description of the hydrogen bond interactions.

B. Solvation Model. We employ the SGB continuum solvation model,[9] which has previously been shown to give good agreement with a Poisson−Boltzmann (PB) treatment of peptide conformational energetics. However, there has been only limited testing of the transferability of the parametrization to proteins. In the course of the present project, we have modified the SGB parametrization described in ref 9, including the development of specialized correction terms which are necessary to predict quantitatively the formation of solvent exposed salt bridges or hydrogen bonds. This work will be described in a subsequent publication. Here, we note that, although the improved parametrization is critical for *absolute* prediction accuracy of surface side chains, it does not strongly impact the *relative* accuracy achieved by the different versions of the force field that are tested here.

X-ray crystal structures of proteins often specify the positions of certain water molecules which can be imaged in the electron density (these waters are presumably sterically or electrostatically restrained). The ability of implicit solvent models to represent the environment created by these crystallographic waters is unclear, and for this reason, we perform simulations both with and without these waters explicitly represented. When the crystal waters are included, the positions of the oxygen atoms are held fixed, and the positions of the hydrogens are determined by energy optimization, prior to any simulations. The SPC parameters are employed for the explicit waters.[12]

C. Optimization Algorithms. We use a hierarchical approach to single side chain prediction. Initially, side chain conformations are sampled using a highly detailed (10° resolution) rotamer library developed by Xiang and Honig.[13] This library contains, for example, 2086 rotamers for Lys. The use of such a detailed library ensures adequate sampling. The associated computational expense is reduced by prescreening the rotamers using only hard sphere overlap as a criterion (this can be made very rapid with the use of a cell list); thus, many rotamers can be excluded before performing energy evaluations. After choosing the lowest energy rotamer, the side chain is completely energy minimized (<0.001 kcal/mol/Å final root-mean-square gradient; all side chain atoms unconstrained in Cartesian space) using a novel minimization algorithm that we developed.[14] This algorithm combines the powerful Truncated Newton method[15] with other techniques based on multiple length scales, leading to a qualitative reduction in computational effort (more than an order of magnitude in preliminary tests) as compared to alternative optimization algorithms. This approach is readily applied to optimization in both the gas phase and with the generalized Born solvation models.[16,9,17] The overall timings for the complete single side chain optimizations are ∼30 s per residue, including all computational overhead, on a 600 MHz Pentium processor. This level of efficiency allows us to study large data sets for single side chain prediction using a small PC cluster.

It should be noted that the effects of side chain entropy are neglected in these calculations. That is, ideally one would wish to calculate *free energy* differences among alternative side chain conformations (local minima on the potential surface, generally separated from each other by large potential barriers). In addition, it would in principle be possible to compare experimentally observed atomic fluctuations (i.e., using the so-called "temperature factors" reported in protein crystal structure files) with predicted values. However, the necessary sampling for an accurate estimate of entropic effects, especially those associated with correlated motions of multiple side chains, would require, at the present time, excessive computational effort. Thus, comparison of alternative conformations is accomplished with the sum of the internal protein *energy* and the solvation *free energy* (as parametrized to reproduce experimental data at standard conditions), as estimated by the SGB model. Highly significant improvements in prediction accuracy are observed using this level of approximation (which would be typical in most protein structure prediction work).

D. Crystal Packing. One final, critical detail of the calculations is that crystal packing effects are included in order to make the most meaningful comparison with the crystal data. To our knowledge, the only prior treatment of crystal packing in the context of side chain prediction is very early work by Gelin and Karplus,[7] on a single protein.

Crystal unit cells are explicitly reconstructed using the dimensions and space group reported in the Protein Data Bank files. For most proteins, the crystal unit cell contains too many atoms for explicit lattice summation techniques (e.g., Ewald summation) to be computationally feasible. Instead, the simulation system consists of one asymmetric unit (which may contain more than one protein chain) and all atoms from other surrounding asymmetric units that are within 20 Å. Every copy of the asymmetric unit is identical at every stage of the calculation; that is, if the conformation of a side chain is modified, all copies of the side chain in the simulation system are updated simultaneously. Inclusion of the crystal environment in this way ensures that discrepancies between the observed and predicted side chain conformations are due to errors in the energy function and not due to deficiencies in the representation of the physical system. Indeed, as we discuss in detail elsewhere, crystal packing effects play a strong role in stabilizing observed conformations of surface side chains.[18]

III. Results

A. Single Side Chain Prediction. A diverse set of high-resolution protein structures, solved by X-ray crystallography, were chosen for use in this study. Specifically, 36 proteins were selected from a "Culled PDB" list compiled by Dunbrack[19,20] which consists of 909 protein structures solved to 2.0 Å resolution or better (R value <0.2) with maximum pairwise

TABLE 1: Single Side Chain Prediction Results with Crystal Waters Excluded[a]

	QM fit			side RMSD		% correct χ_1		% correct χ_{1+2}	
residue	old	new	*N*	old	new	old	new	old	new
Val	0.39	0.08/0.16	368	0.58	0.55/0.56	94.8	95.7/95.4	N/A	N/A
Ile	0.88	0.38	314	0.46	0.44	98.7	99.0	95.5	95.8
Leu	0.37	0.34/0.38	470	0.61	0.61/0.63	98.3	98.3/98.1	93.6	93.6/93.8
Met	1.00	0.59	79	1.33	1.47	91.1	88.6	76.9	75.6
Cys	1.91	0.35	32	0.92	0.54	90.6	96.9	N/A	N/A
Ser	0.47	0.44/0.34	311	1.18	1.17/1.17	73.0	73.0/73.6	N/A	N/A
Thr	0.77	0.87/0.87	354	0.68	0.70/0.70	91.5	91.0/91.0	N/A	N/A
Asn	1.30	0.16	289	1.50	1.40	82.4	87.2	72.7	72.0
Trp	0.56	0.50	88	0.42	0.43	100.0	100.0	98.9	98.9
Gln	0.98	0.96	227	1.68	1.63	86.3	87.7	77.9	78.8
His	0.79/2.05	0.85/0.97	117	1.49	1.54	94.0	94.0	87.3	87.3
Asp	4.15	0.16/1.95	313	1.42	1.29/1.32	78.3	86.9/84.0	70.3	75.1/73.8
Glu	2.24	1.53	295	1.80	1.76	81.4	80.3	74.5	67.0
Lys	1.09	0.88	334	1.67	1.67	88.3	87.4	83.6	84.2
Arg	1.50	1.15	253	2.04	2.05	92.1	91.7	87.4	86.6

[a] The columns labeled "QM fit" are the (RMS) energy residual (kcal/mol) of the force field fit to the quantum chemical calculations, as reported in ref 6 (the two values for His are unprotonated/protonated). *N* is the total number of residues in our data set. Note that Cys residues involved in disulfide bonds are excluded. The side chain RMSD and % correct χ_1 and χ_{1+2} are defined in the text. "old" refers to the OPLS-AA force field as defined in ref 5; "new" is the refitted force field of ref 6. The multiple values shown for the "new" parameters are, as reported in ref 6, Version 1/2 for Leu, Version 2/3 for Val, Version 1/2 for Ser and Thr, and Version 1/2 for Asp. See text for explanations.

TABLE 2: Single Side Chain Prediction Results with Crystal Waters Included (See Caption for Table 1)

	QM fit			side RMSD		% correct χ_1		% correct χ_{1+2}	
residue	old	new	*N*	old	new	old	new	old	new
Val	0.39	0.08/0.16	368	0.56	0.53/0.53	95.9	95.9/95.9	N/A	N/A
Ile	0.88	0.38	314	0.36	0.31	99.7	100.0	95.8	96.2
Leu	0.37	0.34/0.38	470	0.60	0.60/0.60	98.7	98.9/98.7	93.6	93.4/93.6
Met	1.00	0.59	79	1.16	1.08	94.9	94.9	78.2	83.3
Cys	1.91	0.35	32	0.57	0.24	96.9	100.0	N/A	N/A
Ser	0.47	0.44/0.34	311	1.13	1.13/1.13	75.2	74.6/74.9	N/A	N/A
Thr	0.77	0.87/0.87	354	0.58	0.56/0.60	93.8	94.4/93.5	N/A	N/A
Asn	1.30	0.16	289	1.37	1.26	88.6	93.8	74.0	74.0
Trp	0.56	0.50	88	0.22	0.22	100.0	100.0	100.0	100.0
Gln	0.98	0.96	227	1.31	1.34	90.7	89.4	86.7	87.2
His	0.79/2.05	0.85/0.97	117	1.11	1.10	98.3	99.1	86.4	86.4
Asp	4.15	0.16/1.95	313	1.31	1.20/1.26	84.3	90.1/87.9	70.6	74.8/74.1
Glu	2.24	1.53	295	1.70	1.71	85.4	83.7	73.8	72.1
Lys	1.09	0.88	334	1.74	1.80	89.5	88.3	80.9	80.5
Arg	1.50	1.15	253	1.84	1.85	92.9	92.9	87.4	87.4

sequence identity of 30%. Proteins with nonpeptide ligands or nonstandard (chemically modified) residues were excluded from study, as were proteins with large disordered regions. The largest protein contained 285 residues; the total number of residues represented is 4808. The Protein Data Bank[21] codes for the proteins included are 1ew4, 1u9a, 5icb, 2pth, 1bk7, 1dvo, 3vub, 1et1, 1aie, 1ej8, 2fcb, 1nps, 1whi, 1aho, 1bv1, 1c44, 1edm, 2igd, 1d4t, 1dhn, 1qto, 1ay7, 5hpg, 1f94, 3ezm, 1pbv, 1qtw, 1bue, 2btc, 1sur, 1b2p, 1a8l, 1byi, 1ako, 1tvd, 2plc, and 1qts.

We report accuracy of single side chain prediction using a variety of standard measures. First, side chain dihedral angles are considered to be "correct" if they are within ±40° of the experimental value. This criterion is chosen because it has been employed in numerous previous studies to quantify side chain prediction accuracy and because thermal fluctuations and experimental uncertainties in the dihedral angles are generally well within this range (but the range is still small enough to distinguish qualitatively different "rotamer" states). Thus, for example, "% correct χ_{1+2}" indicates the fraction of side chains with two or more heavy-atom dihedral angles in which the first two are both within 40° of the native. Second, we calculate the side chain root-mean-square deviation (RMSD) for all heavy atoms, excluding the C^β (which is largely fixed by the backbone position).

The results are summarized in Tables 1 (no crystal waters) and 2 (crystal waters included). As is to be expected, the inclusion of crystallographic waters leads to higher accuracy, in part because of the resulting spatial constraints and in part because of the improved description of the solute−solvent electrostatics. However, the general trends with respect to changes in the protein potential function are similar for both models. In examining the results, it should be kept in mind that virtually all side chains buried in the interior of the protein (and a nontrivial fraction that are partially solvent exposed) are severely conformationally constrained by geometrical packing considerations. Predictions will be successful for such cases independent of the potential energy function. Thus, in practical terms, a 5−10% increase in χ_1 prediction accuracy, for example, represents a significant improvement.

The results are also summarized graphically in Figure 2. Here, the improvement in the χ_1 prediction accuracy is plotted against the improvement in the fit to the calculated dipeptide quantum mechanical energies, as reported in ref 6. [Note that the improvements in the fit to the quantum chemical data contain contributions from improvements to the backbone parameters. The single side chain prediction results discussed here are not sensitive to these parameters.] The correlation between these two quantities is remarkable. The amino acids for which little or no improvement was made in the fit (<1 kcal/mol improvement, i.e., smaller than or similar to RT at room temperature) show little or no improvement in the prediction accuracy (improvement in % correct χ_1 of <2%). However, the three

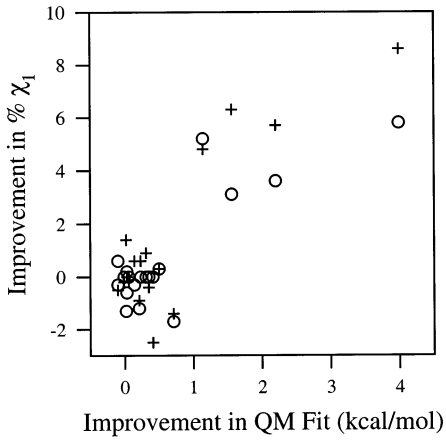

Figure 2. Improvement of the accuracy of χ_1 prediction (measured by % correct χ_1) which correlates strongly with improvement in the fit to the calculated quantum chemical energies (RMS deviation from ab initio results, in kcal/mol). Circles/crosses are results with/without crystal waters. In the cases of amino acids with multiple fits reported in ref 6, each fit is represented on the plot.

amino acids with the most significant improvements in the parameter fitting, Cys, Asn, and Asp (which is represented by two points on the plot, because two fits were reported), show substantial improvements in side chain prediction accuracy. More detailed comments on each of these three cases, as well as the cases where little improvement was observed, follow.

The results for Asp are certainly among the most interesting, and are represented graphically in Figure 3. In the original OPLS-AA force field,[5] the torsion parameters for this side chain were taken from generic OPLS-AA parameters for small molecules (the χ_1 potential is depicted in the left panel, bottom row of Figure 3). In the work of Kaminski et al., however, specialized parameter sets were developed specifically for Asp, which brought the force field results for dipeptide rotamers into much better correspondence with the quantum chemical data than the generic parameters. The development of the specialized parameters can be justified by the close proximity of the amide group of the side chain to the backbone; the torsional parametrization probably counteracts errors in the nonbonded terms and takes into account differences in local chemistry. Prior to the present paper, however, it had yet to be demonstrated that such a reparametrization would improve results in the condensed phase.

An additional complication is that two fits were reported for Asp in ref 6, one of which was unconstrained and resulted in rather large torsional parameters (right bottom panel of Figure 3), and the second was constrained such that the torsional parameters did not exceed 4.5 kcal/mol (middle bottom panel). Both of these refitted parameter sets result in substantial improvements in single side chain prediction accuracy, but the results unambiguously demonstrate that the unconstrained fit provides significantly better accuracy than the constrained fit. That is, as the size of the torsional barrier is increased, the single side chain prediction accuracy increases substantially by all measures. In the optimal parameter set, the prediction accuracy for aspartic acid is comparable to that of other polar and charged amino acids, as opposed to the results for the original parameter set which are anomalously poor. Thus, it is highly unlikely that the very large torsional parameters in the unconstrained (highest accuracy) fit are due to overfitting of the quantum chemical data or any other artifacts. We suspect that the magnitude of the torsional parameters will be reduced substantially in a

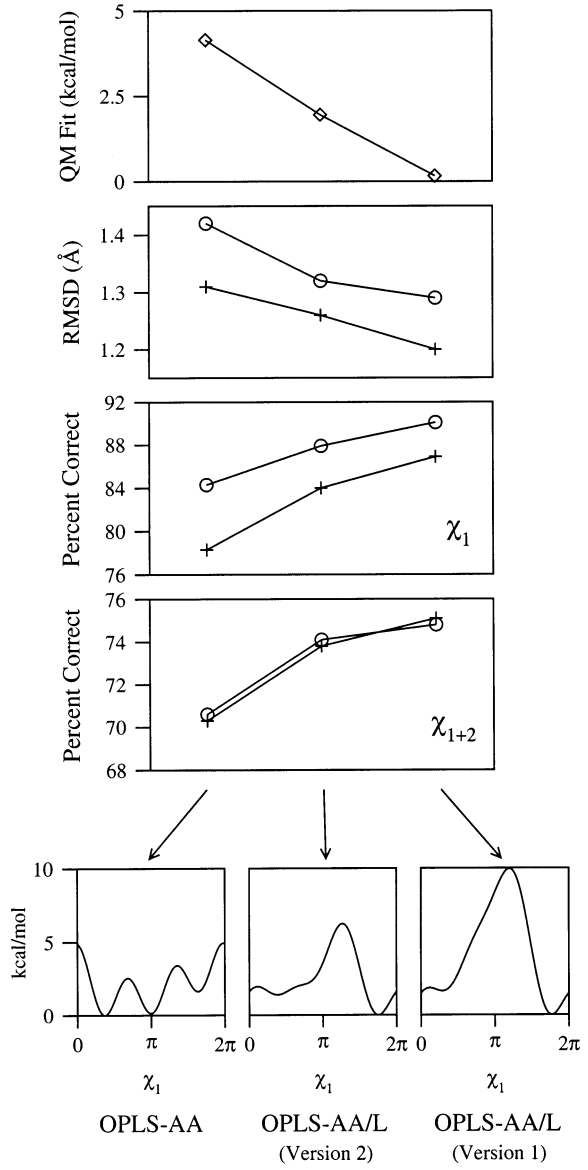

Figure 3. Graphical representation of the results for Asp. Top row: Root-mean-square energy residual of the fit to the peptide quantum chemical calculations. Middle three rows: Circles/crosses are single side chain prediction results with/without crystal waters. Bottom row: Potential functions along χ_1 for the three different parameter sets.

polarizable force field (i.e., that the effects of polarization may be particularly apparent in Asp, which has the shortest charged side chain).

The form of the unconstrained refitted χ_1 torsional potential for Asp, as depicted at the bottom of Figure 3, may appear to be surprising at first glance. In particular, the large barrier near $\chi_1 = 180°$ might appear to preclude the existence of side chain rotamers with near-"trans" conformation. That is, 30% of experimentally observed Asp side chains are found with $\chi_1 = 180° \pm 60°$[22] (in the data set used here, 32% have this conformation). In fact, 38% of the *predicted* Asp side chain conformations fall into this conformational class; that is, the "trans" conformation is slightly *overpredicted*, despite the large barrier in the torsional potential (The g^+ conformations [$\chi_1 = +60° \pm 60°$] account for 18% of both the observed and predicted conformations, whereas for the g^- conformations [$\chi_1 = -60° \pm 60°$], the values are 44% predicted and 50% observed.). The resolution of this discrepancy is simply that the other components of the potential energy function, particu-

larly electrostatics for charged side chains, also contribute strongly to the observed distribution of side chain conformations. Favorable "1,4" Lennard-Jones and electrostatic interactions, as well as favorable solvation free energy and interresidue interactions, stabilize the trans conformations. Because these interactions vary as a function of the backbone conformation and local environment of the particular side chains, they are not included in the plots in Figure 3.

The results for Asn are similar to those for Asp, although the improvement in accuracy is seen only in χ_1 and the RMSD, not χ_{1+2}.

Although the number of data points for Cys is smaller (32; the majority of Cys residues in our data set were involved in disulfide bonds and thus were excluded), there is an unambiguous improvement in all measures of side chain prediction accuracy. Cys benefited from refitting of not only the torsional parameters but also the nonbonded parameters (primarily a reduction of the partial charge on sulfur, with compensating changes in the dispersion). Note, however, that Met underwent a similar refitting of both torsional and nonbonded parameters, and these new parameters do not result in a clear improvement in prediction accuracy. In the absence of crystal waters, the Met results actually show a slight decrease in prediction accuracy with the new parameters, although with the inclusion of crystal waters there is a small increase in accuracy, particularly at χ_2. Met of course has a much larger side chain than Cys, and the refitting did not reduce the energy error nearly as much as for Cys. It should also be noted that only the parameters associated with the χ_1 angle of Met were refitted, despite the fact that the sulfur is located near the end of the side chain.

The remaining amino acids demonstrate little or no improvement in prediction accuracy, consistent with their relatively small (if any) improvement in the fit to the quantum chemical data. The majority of these amino acids, including Val, Ile, Leu, Thr, Trp, and His, demonstrate excellent prediction accuracy with all of the parameter sets tested. In the case of Trp, the excellent results may be biased by the large size of the side chain, which strongly constrains the possible conformations in single side chain prediction because of simple steric effects. It should also be noted that we have made only a rudimentary effort to assign protonation states for His, based simply upon the pH reported in the PDB files. Because the protonation states for many of the His residues have not been determined experimentally, all predicted results are averaged together.

The long polar and charged side chains (except for Glu, which is discussed below) demonstrate no clear improvement in prediction accuracy. All except Gln showed some improvement in the quality of the fit to the quantum chemical data, although the new parameter sets still have relatively large discrepancies. Further improvement in the quality of the fits may be possible, for two reasons. First, in the cases of Lys and Arg, only the χ_1 torsional parameters were refitted. Second, for all of the long side chains, it is much more difficult to ensure adequate coverage of the conformational space with a reasonable number of ab initio data points. Ultimately, however, the prediction accuracy for the long polar/charged side chains may be affected much more strongly by the description of the nonbonded interactions (electrostatics, van der Waals) and solvation than the torsional parameters. As discussed in Section II, some progress has already been made toward improved description of solvated hydrogen bonds and salt bridges, which contributes to the absolute accuracies reported here.

The refitted parameters reported for Glu in ref 6 also demonstrated little improvement in accuracy over the original OPLS-AA parameters. However, in this work, we have improved the prediction accuracy for Glu by obtaining new torsional parameters through a minor modification to the prior refitting procedure. Specifically, seven local minima were located on the glutamic acid dipeptide surface, and as described in detail previously, 13 single-point calculations were performed in the vicinity of each minimum in a "cross-like" pattern. These data points were used for the torsional fit, and the results are presented in Table 3. Note that local minimum #7 is a gross outlier in this fit, with an energy residual of 3.6 kcal/mol. In the absence of an independent method of validation, it is difficult to judge whether this outlier should be excised, on the basis that it creates a much poorer fit to the other data points, or retained, on the basis that it points to a critical deficiency of the model. With the validation procedure described here, it is a simple matter to evaluate the prediction accuracy of the Glu parameter obtained with and without point #7. As is clear from Table 3, the prediction accuracy increases substantially (%χ_1 improves by ~5%) when point #7 is omitted from the torsional fit.

TABLE 3: Results of Fits to Quantum Chemical Data for Glu[a]

conformer	ab initio	old	new	this work
1	0.00	−2.19	−1.28	−0.46
2	7.89	8.48	7.89	8.65
3	3.68	6.04	3.19	3.38
4	14.09	12.38	13.62	12.13
5	7.20	4.95	6.05	7.44
6	12.79	12.04	12.60	14.51
7	10.95	14.91	14.55	N/A
QM fit		2.24	1.53	1.14
% correct χ_1		81.4	80.3	85.8
side RMSD		1.80	1.76	1.68
NCCC V_1		0.845	4.952	3.589
NCCC V_2		−0.962	−0.257	2.120
NCCC V_3		0.713	−0.235	−1.060
CCCC V_1		−1.697	−1.618	−1.527
CCCC V_2		−0.456	−0.571	3.406
CCCC V_3		0.585	0.0	0.0

[a] The "conformer" number is an arbitrary label for the quantum chemical local energy minima, as designated in ref 6. The column "ab initio" lists relative quantum chemical energies of the local minima. "old" and "new" are the fitted results for OPLS-AA and OPLS-AA/L, respectively. "this work" reports a new fit in which conformer 7 is omitted, resulting in substantially lower RMS error with respect to the quantum chemical data ("QM fit") as well as improved side chain prediction results ("% correct χ_1" and "side RMSD"). The bottom portion of the table lists force field parameters obtained from each fit. "NCCC" and "CCCC" refer to the two heavy atom torsions which together parametrize the χ_1 torsional potential. V_n refers to the coefficient of the $\cos(n\phi)$ term.

Finally, the prediction accuracy for Ser, as measured by % correct χ_1, is by far the worst of all of the amino acids. The reported torsional fits have low RMS energy discrepancy from the quantum data, and because of the small size of the side chain, there is little possibility of inadequate sampling. In addition, note that Thr, which has a poorer fit to the quantum chemical data, has much higher prediction accuracy. The picture that emerges is that the small polar side chain of Ser, which can generally rotate without any steric hindrance, is exquisitely sensitive to the local electrostatic environment (which may itself be incorrect because of errors in assigning protonation states, for example). Work is in progress to investigate whether reduction of the partial charges (i.e., in a manner similar to Cys) or adjustment of relevant parameters in the solvation model is capable of improving the single side chain prediction results.

B. Loop Side Chain Prediction. Both backbone and side chain torsional parameters were refitted in ref 6, but single side chain prediction is sensitive only to the side chain parameters. The backbone parameters can in principle be validated by performing backbone sampling, for example, on the flexible loop regions. Note, however, that in a realistic test, the side chains would need to be sampled concomitantly with the backbone. That is, because of coupling between the backbone and side chains, there is no simple way of validating the backbone parameters in isolation from the side chain parameters. In addition, a serious technical challenge associated with validating the backbone parameters in this way is the requirement of adequate sampling. That is, the results would only be meaningful if it were possible to locate the global minimum or at least a local minimum with an energy less than that of the minimized native, in a reasonable amount of computational time. There exist numerous techniques for loop sampling, but at the present time, we have not found a satisfactory solution for adequate sampling of a large statistical sample of loops with reasonable computational expense. Algorithmic evaluation and development are ongoing, however.

We have instead devised a simpler test which provides some information about the improvement in accuracy achieved by the new parameters for *both* side chain and backbone conformations. The test involves predicting the conformations for all side chains in a loop and then energy minimizing the entire loop, including both the backbone and side chain atoms. The remainder of the protein is held fixed at the native. Thus, the accuracy of the backbone achieved in this test will reflect both the side chain and backbone parameters; no rigorous decomposition of the results is possible. However, as will be seen below, the loop side chain prediction results complement the single side chain prediction results discussed above by providing additional evidence for the superiority of the new parameters, particularly as the loop length increases. The procedure is fast enough that it can be performed on all 379 loops (as defined by DSSP[23]) in the 36 protein data set used for single side chain prediction. These loops range from single-residue turns to 24-residue regions. No crystal waters are used, but the crystal environment is included in the same manner as in the single side chain prediction results.

Side chain prediction for the loops was accomplished using a very simple sampling method described by Xiang and Honig.[13] In brief, all side chains are placed in random rotamer states, and then single side chain optimization is performed sequentially for each residue in the loop. Convergence is achieved when the rotamer state for each side chain ceases to change. Xiang and Honig demonstrated that this simple, fast procedure is remarkably effective for side chain optimization on entire proteins.[13] We have tested this sampling procedure for loop side chain prediction by comparing the energies of the final structures obtained (after minimization) with the minimized native loop. A reasonable, although somewhat arbitrary, criterion for adequate sampling is that the final energy obtained is lower than, or <1 kcal/mol greater than, that of the minimized native loop. A total of 88% of the loops in the test set satisfied this criterion after side chain prediction and minimization. In the analysis of the results, we have excluded any cases that do not satisfy the adequate sampling criterion. The results are not strongly dependent on the precise cutoff used.

The loop side chain prediction results are depicted in Figure 4. The average RMSDs for the backbone and for all heavy atoms are plotted as a function of the loop length (which is binned according to 1−2, 3−4, 5−6, 7−8, 9−10, >10 residues). For the longest loops (>8 residues), the results demonstrate a very clear improvement in prediction accuracy with the new parameters (solid line) relative to the older ones (dashed line), for both the backbone and the side chains. A smaller, but fairly consistent, improvement can also be observed for intermediate loop lengths (3−8 residues). These shorter loops of course are constrained more severely by the (fixed) protein surroundings and thus demonstrate less conformational flexibility and can be expected to be less sensitive to the parameter variation. Oddly, the very shortest "loops" (better described as turns), with only one or two residues, demonstrate slightly less accuracy with the new parameters. However, this small effect is strongly outweighed by the improvement observed for the larger loops.

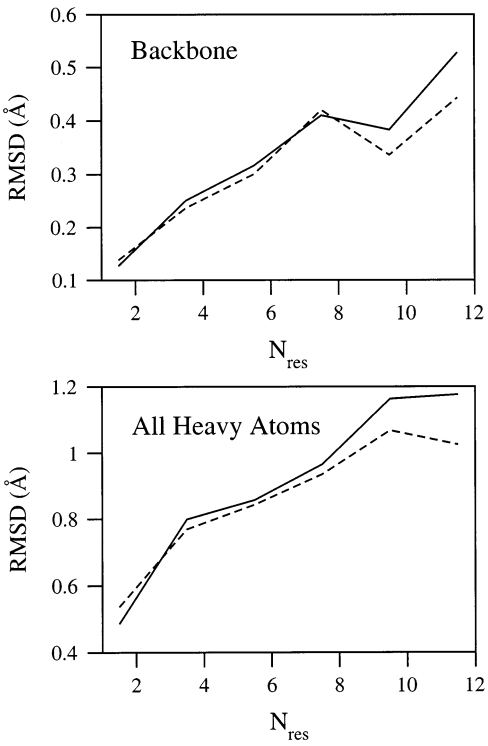

Figure 4. Loop side chain optimization with the old (dashed line) and new (solid line) OPLS torsional parameters. The loop side chains are optimized as described in the text, and then the full loop (backbone plus side chains) is energy minimized. The average RMSD between the generated and native structures, for all heavy atoms (bottom) and only the backbone atoms (top), is plotted as a function of loop length, binned in the following intervals: 1−2, 3−4, 5−6, 7−8, 9−10, >10 residues. Cases with significant sampling error (energy of optimized structure >1 kcal/mol above that of the minimized native) have been excluded (∼12% of the total).

IV. Conclusion

We have shown in this paper that the accuracy of the protein force field, as measured by agreement with *gas phase* quantum chemical data for dipeptide conformations, is an important factor in determining the accuracy of protein side chain prediction in the *condensed phase*. Modifications of the OPLS-AA force field have led to substantial improvement in prediction accuracy for single side chains. Residual errors in the force field, which were not eliminated in ref 6, have a number of sources, including neglect of polarization effects, failure to reoptimize stretching and bending terms or to incorporate a more complex functional form for these terms, and inaccuracies in the nonbonded interactions. Further improvements in the force field to enforce a better fit, including explicit incorporation of polarizability, will be investigated in subsequent papers.

We note however that not all errors in side chain prediction accuracy can be attributed to problems with the force field; errors may also arise from the solvation model, description of nonbonded interactions, specification of the protonation state of ionizable side chain groups, neglect of solute entropy (i.e., we calculate energies, not free energies), incomplete sampling, and uncertainty in the experimental data itself. In the long run, progress must be made in all aspects of the model in order to obtain robust and highly precise prediction accuracy. Future papers will discuss our efforts in the other areas mentioned above, some of which must be approached more heuristically than the present one. It is, however, highly encouraging that systematic improvement is possible, and this augurs well not only for protein force field development but for treatment of a wider chemical space (e.g., protein−ligand interactions) as well.

The simple side chain prediction tests that we have performed here provide upper bounds for the prediction accuracy that can be achieved in more realistic modeling situations. For example, in the context of predicting side chains for a homology-based protein model, more extensive sampling will be required (optimization of all side chains), and the backbone will not be perfectly accurate (especially loop regions, which may require backbone sampling). Overall, the combination of the refined all-atom OPLS force field and the SGB solvent model makes possible very high accuracy predictions. Some problems remain, e.g., Ser side chain conformations, but the methodology that we have introduced here makes possible a systematic, iterative process of model improvement. That is, the side chain prediction tests described here make it possible both to isolate errors in existing energy functions and to validate refined parameters, with a large and diverse test set of proteins in a realistic condensed phase environment.

Finally, we note that, although we have focused on the OPLS force field, it is of course possible to apply the same methods of refinement and validation described here to other force fields intended for use on proteins. On the other hand, obtaining fair comparisons of accuracy among different force fields is highly nontrivial (and thus we have not attempted to perform such comparisons here) because differences among the solvent models employed in conjunction with the force fields can lead to much larger differences in prediction accuracy than the differences in the force fields themselves. Although generalized Born models have been employed in conjunction with several different force fields (including CHARMM[24] and AMBER[25,26]), there are substantial differences in parametrization among these models. Moreover, our SGB model cannot be applied to other force fields (without complete reparametrization) because the parameters (atomic radii and correction terms) depend implicitly on the force field partial charges. Comparisons could be made with distance dependent dielectic or in a vacuum, but the results are unlikely to be meaningful because, as will be discussed in more detail in a future publication, the errors resulting from crude or nonexistent representation of solvent are much larger than the differences in accuracy observed here for different torsional energy parameters.

Acknowledgment. We thank Drs. Z. Xiang and B. Honig (Columbia U. Department of Biochemistry) for providing us with their rotamer libraries and for many helpful discussions. M.P.J. wishes to acknowledge support from an NSF Postdoctoral Fellowship in Biological Informatics. This work was supported in part by a grant to RAF from the NIH (GM-52018).

References and Notes

(1) Novotný, J.; Bruccoleri, R.; Karplus, M. *J. Mol. Biol.* **1984**, *177*, 787.
(2) Holm, L.; Sander, C. *J. Mol. Biol.* **1992**, *225*, 93.
(3) Park, B. H.; Huang, E. S.; Levitt, M. *J. Mol. Biol.* **1997**, *266*, 831.
(4) Bonneau, R.; Strauss, C. E. M.; Baker, D. *Proteins* **2001**, *43*, 1.
(5) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
(6) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. J. *J. Phys. Chem. B* **2001**, *105*, 6474.
(7) Gelin, B. R.; Karplus, M. *Biochem.* **1979**, *18*, 1256.
(8) Petrella, R. J.; Lazardis, T.; Karplus, M. *Fold. Des.* **1998**, *3*, 353.
(9) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983.
(10) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. J. *J. Am. Chem. Soc.* **1984**, *106*, 765.
(11) Murphy, R. B.; Pollard, W. T.; Friesner, R. A. *J. Chem. Phys.* **1997**, *106*, 5073.
(12) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331−342.
(13) Xiang, Z.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421.
(14) Jacobson, M. P.; Friesner, R. A. In preparation.
(15) Xie, D.; Schlick, T. *SIAM J. Opt.* **1999**, *10*, 132.
(16) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrikson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
(17) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.
(18) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597.
(19) Dunbrack, R. L., Jr. http://www.fccc.edu/research/labs/dunbrack/culledpdb.html
(20) Hobohm, U.; Scharf, M.; Schneider, R. *Protein Sci.* **1993**, *1*, 409.
(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.
(22) Dunbrack, R. L., Jr.; Karplus, M. *J. Mol. Biol.* **1993**, *230*, 543.
(23) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.
(24) Dominy, B. N.; Brooks, C. L., III. *J. Phys. Chem. B* **1999**, *103*, 3765.
(25) Onufriev, O.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712.
(26) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005