

Intraprotein Electrostatics Derived from First Principles: Divide-and-Conquer Approaches for QM/MM Calculations

PABLO A. MOLINA, HUI LI, JAN H. JENSEN

Department of Chemistry, University of Iowa, Iowa City, IA 52242

Received 17 March 2003; Accepted 22 April 2003

Two divide-and-conquer (DAQ) approaches for building multipole-based molecular electrostatic potentials of proteins are presented and evaluated for use in QM/MM calculations. One approach is a further development of the neutralization method of Bellido and Rullmann (J Comput Chem 1989, 10, 479–487) while the other is based on removing part of the electron density before performing the multipole expansion. Both methods create systems with integer charges without using charge renormalization. To determine their performance in terms of location of cuts and distance to QM region, the new DAQ approaches are tested in calculations of the proton affinity of N^ε of Lys55 in the inhibitor turkey ovomucoid third domain. Finally, the two methods are used to build a variety of MM regions, applied to calculations of the pK_a of Lys55, and compared to other computational methodologies in which force field charges are employed.

© 2003 Wiley Periodicals, Inc. J Comput Chem 24: 1971–1979, 2003

Introduction

Electrostatic interactions are in general thought to be the principal force determining the structure and function of biomolecules. Most biomolecular force fields describe electrostatic interactions by placing charges at atomic centers, i.e., by a distributed multipole expansion of the molecular electrostatic potential (MEP) truncated after the first term. These charges are obtained either from *ab initio* electronic structure theory or by empirical parameterization, for various functional groups, and are assumed to be transferable. This relatively simple approach has been used extensively, often in conjunction with a Poisson–Boltzmann treatment of solvation or polarization, in biomolecular modeling with great success.

However, there is mounting evidence that an atom-centered charge (ACC) model is not always an adequate representation of the MEP.^{1–13} For example, the ACC model tends to underestimate the directionality of hydrogen bonds,¹⁰ while models that include additional charges¹⁰ or higher-order multipoles^{1,2} reproduce *ab initio* results significantly better. Several “multipole libraries” have been or are being developed for amino acids,^{6,14–17} and at least two force fields^{9,18} employ a multipole-based electrostatic model. However, the transferability of the multipole parameters can be complicated by the conformational dependence of the higher moments and, in general, because a higher degree of accuracy (compared to charge-based models) is typically sought.^{6,13,17,19,20}

Minikis et al.¹² investigated an alternative approach for generating multipole-based MEPs (mMEPs) of proteins for use with the effective fragment potential (EFP) method.^{21,22} The EFP method is a hybrid QM/MM method in which the chemically active part of

a molecular system is treated with *ab initio* electronic structure while the rest is treated with an mMEP (charges through octupoles and dipole polarizability tensors for each valence electron pair). The EFP is generated specifically for a given protein by a divide-and-conquer (DAQ) approach in which the protein is divided into smaller overlapping pieces for which mMEPs can be generated *ab initio* and then reassembled by excluding parameters from the region of overlap. The approach is similar in spirit to other “reassociation of fragment” approaches^{16,23–25} except that significantly larger fragments are used.

One problem that arises when the EFP pieces are assembled is that the monopoles do not add to a net integer charge. The same problem occurs in atomic-charge models and various solutions have been proposed to address this issue: Bellido and Rullmann²³ presented two “neutralization” procedures to ensure a net integer charge and Young et al.²⁵ examined in detail the errors introduced by the reassembling of fragments as a function of the location of the cut. In the EFP model, a scaling procedure^{12,26} was previously used to ensure that the reassembled fragment had a net integer charge. We tested the accuracy of this scaling procedure in our study of the 56-residue protease inhibitor turkey ovomucoid third domain (OMTKY3), where the EFP region was generated by only nine *ab initio* calculations.¹² The gas-phase proton affinity (PA) of a particular residue, Lys55, was calculated using an EFP of part of the protein constructed with two spatially different fragments (one of which was built with and without cuts). Based on this testing, it

Correspondence to: J. H. Jensen; e-mail: jan-jensen@uiowa.edu

was estimated that the PA value for an EFP region describing all residues and made of several reassembled smaller pieces was within 1 kcal/mol with respect to a hypothetical EFP region obtained from a single *ab initio* calculation.

While the scaling procedure proved to be an efficient DAQ method, it has two shortcomings: (1) It scales all the charges so that all the monopoles are slightly different from the originals; (2) the procedure is somewhat cumbersome because it required manual identification and removal of duplicate expansion and polarizable points. In this article, we present two new DAQ approaches that make it possible to build a large EFP region from smaller pieces without recourse to scaling. Both methods are computationally more convenient and yield an accuracy similar to the one offered by the scaling procedure as demonstrated by our tests of calculated pK_a s of Lys55 in OMTKY3.

The article is organized as follows: First, the EFP method and the two new DAQ methods are briefly described. Second, we test our new DAQ methods and determine the optimum combination of cut location and DAQ methodology in assembling EFPs. We do this by computing PAs and pK_a s of Lys55 with various cut locations and DAQ approaches. Third, we evaluate the use of three MM force field-described regions (AMBER, OPLSAA, and CHARMM) in the prediction of the pK_a of Lys55. Fourth, we summarize our findings and describe future directions.

Computational Methodology

QM/EFP Model of Lys55 in OMTKY3

The construction of the buffer and EFP regions and the general methodology for pK_a predictions have been discussed in detail in previous articles^{12,27} and are only summarized here.

The solution structure of OMTKY3 has been determined using NMR by Hoogstraten et al.²⁸ and was obtained from the Protein Data Bank (entry 1OMU). We use the first of the 50 conformers without further refinement of the overall structure.

1. The electronic and geometric structures of the Lys55 and Tyr20 side-chains are treated quantum mechanically at the MP2/6-31+G(2d,p)//RHF/6-31G(d) level of theory [Fig.1(a)], while the rest of the protein is treated with an EFP (described in more detail below). The use of the diffuse functions on atoms near the buffer region causes self-consistent field (SCF) convergence problems due to couplings with the induced dipoles in the EFP region, so the 6-31+G(2d,p) basis set was used only for the $C^{\delta}H_2C^{\epsilon}H_2NH_3^+ \dots HO-C^{\xi}(C^{\epsilon 1,2}H)_2$ atoms in the MP2 calculation.
2. The *ab initio* region is separated from the protein EFP by a buffer region²⁹ comprised of frozen localized molecular orbitals (LMOs) corresponding to all the bond LMOs connecting the dark atoms in Figure 1(b), as well as the core and lone-pair LMOs belonging to those atoms. The Pro22 buffer is needed to describe its short-range interactions with Tyr20.¹² The buffer LMOs are generated by an RHF/6-31G(d) calculation on a subset of the system (see ref.13), projected onto the buffer atom basis functions,³⁰ and subsequently frozen in the EFP calculations by setting select off-diagonal MO Fock matrix elements to

zero.^{31,32} The *ab initio*/buffer region interactions are calculated *ab initio* and thus include short-range interactions.

3. The EFP describing the rest of the protein is generated by separate *ab initio* calculations on overlapping pieces of the protein truncated by methyl groups.¹² Depending on the treatment of the MM region (described below), the EFP is built from two, three, or nine pieces. The protein EFP is assembled by one of the three methods described below. In all cases, the electrostatic potential of each protein piece is expanded in terms of multipoles through octupoles centered at all atomic and bond midpoint centers using Stone's distributed multipole analysis,³³ while the dipole polarizability tensor due to each LMO in the EFP region is calculated by a perturbation expression.¹²

The EFP, buffer, and *ab initio* regions are combined [Figs.1(b) and 1(c)] and the geometry of the *ab initio* region is optimized at the RHF/6-31G(d) level of theory while the geometries of the buffer and EFP regions are fixed. In a second calculation a proton is removed from the amine group and the geometry of the *ab initio* region is reoptimized. Single-point energies ($E^{MP2/RHF}$) are evaluated at the MP2/6-31+G(2d,p) level of theory by excluding excitations from the buffer LMOs (and the core MOs in the *ab initio* region).²⁹

Thermochemical Free Energy

The thermochemical free energy is the sum of the vibrational, rotational, and translational terms. The vibrational free energy (G^{vib}) of the optimized part of the *ab initio* region is calculated by the partial Hessian vibrational analysis (PHVA) method developed by Li and Jensen.³⁴ This method is based on a method by Head³⁵ in which only a subset of the atoms (in our case the atoms in the *ab initio* region) are displaced during a numerical Hessian calculation to calculate a "partial Hessian." Further studies by Li and Jensen³⁴ have shown that vibrational energy and entropy changes for proton abstraction reactions calculated using frequencies obtained in this manner are within 0.2 kcal/mol of conventional values. The translational and rotational free energies (G^{trans} and G^{rot}) are calculated using the atomic masses and positions of all atoms in the protein.

Solvation Energy

The solvation energy (ΔG_s) is calculated using the ONIOM-PCM/X approach,³⁶ which combines IEF-PCM/ICOMP = 0 protein solvation energies with D-PCM/ICOMP = 4 solvation energies of model systems:

$$\begin{aligned} \Delta G_s(\text{Protein:D-PCM/ICOMP} = 4) \\ = \Delta G_s(\text{Protein:IEF-PCM/ICOMP} = 0) \\ + \Delta G_s(\text{Model:D-PCM/ICOMP} = 4) \\ - \Delta G_s(\text{Model:IEF-PCM/ICOMP} = 0) \quad (1) \end{aligned}$$

The model is shown in Figure 1(a). The GEPOL-GB tessellation³⁷ scheme and UAHF radii³⁸ are used to construct the cavity of the entire system, and the cavitation and dispersion–repulsion energies

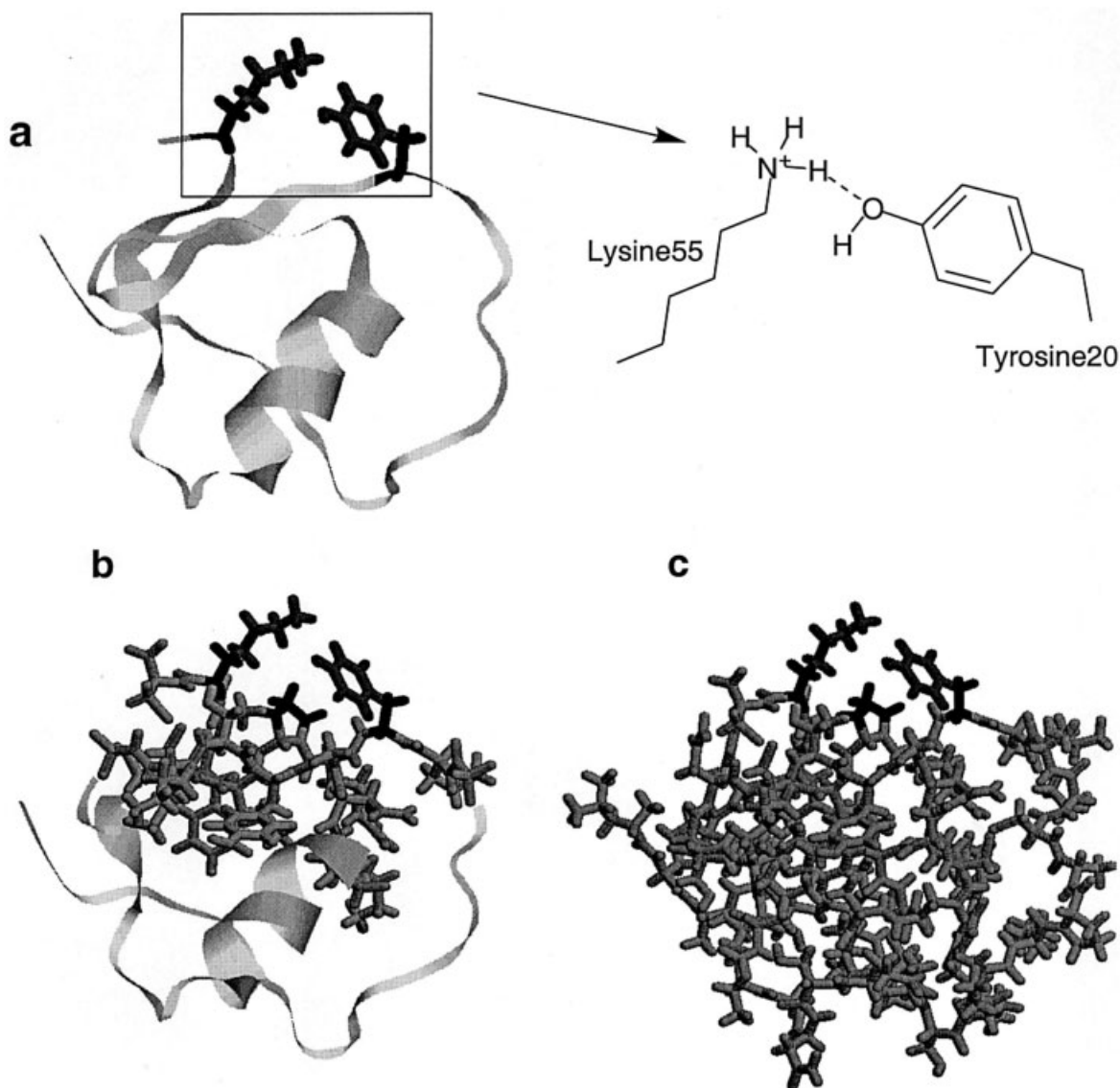


Figure 1. (a) OMTKY3 Lys55–Tyr20 *ab initio* region including buffer regions (in dark) with detail. Remainder of protein is ribbon structure. (b) Lys55–Tyr20 *ab initio* region with the surrounding 14-Å radius EFP region as measured from Lys55 NZ atom. Note proline section is now in buffer (see Fig. 5 for detail). (c) Similar to (b) except now with the entire protein as EFP.

are calculated as described previously.²⁷ The protein solvation energies are calculated using the EFP/IEF-PCM interface developed by Bandyopadhyay et al.²⁴ using the iterative approach as implemented by Li et al.³⁹

The EFP/PCM interface is similar to an all-*ab initio* PCM calculation except that the electrostatic potential (V) of the EFP region is due to its multipole representation of the electrostatic potential. The induced surface charges influence the induced dipoles and this contribution is iterated to self-consistency. Previously,²⁷ we found several cases of divergence, presumably where surface charges are close to a polarizability tensor. Thus, the polarizability tensors are removed for the single-point calculations necessary for the solvation energies.

As before,²⁷ the dispersion–repulsion contribution to the solvation energy is calculated only for the *ab initio* and buffer regions. Further, surface smoothing by the generation of additional spheres is prevented by using $RET = 100$ in the \$PCM group because the number of added spheres never converged for the protein within the memory available.

pK_a Prediction

The pK_a of Lys55 is predicted by computing the free energy (ΔG) for the following reaction:



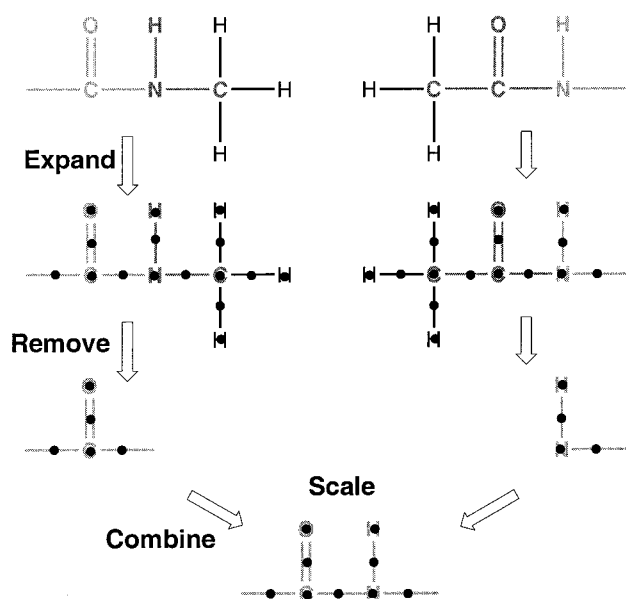


Figure 2. Schematic representation of the ERS method described in the text. The dots represent expansion points for the multipole expansion.

via calculations of the free energy of each protonation state

$$\Delta G = \{[G(\text{Lys55}) - G(\text{Lys55H}^+)] - [G(\text{CH}_3\text{NH}_2) - G(\text{CH}_3\text{NH}_3^+)]\} \quad (3)$$

Here, $G(X)$ is the total free energy of species X and is the sum of the ground-state energy ($E^{\text{MP2/RHF}}$), thermochemical free energy (G_{therm}), and solvation energy (G_s) described above. Thus,

$$\Delta G = \Delta E^{\text{MP2/RHF}} + \Delta G_{\text{therm}} + \Delta G_s \quad (4)$$

Finally, the $\text{p}K_a$ of Lys55H^+ is calculated as the $\text{p}K_a$ shift relative to the experimental $\text{p}K_a$ value of methylamine:

$$\text{p}K_a(\text{Lys55H}^+) = 10.6 + \Delta G/1.36 \quad (5)$$

Miscellaneous

The Foster–Boys localization procedure was used throughout this work to generate LMOs,^{40,41} and all calculations were done with the quantum chemistry code GAMESS⁴² except the D-PCM/ICOMP = 4 calculations, which were done using GAUSS-IAN98.⁴³ The methods described in this article are available in the January 14, 2003, version of the GAMESS program.

Expand-Remove-and-Scale Method

In the divide-and-conquer method due to Minikis et al.,¹² referred to hereafter as the expand-remove-and-scale (ERS) method, a large EFP is constructed from two smaller pieces, with a common region of overlap, in three steps (Fig. 2). First, the wave function of each

piece is computed and the MEP is expanded in terms of mono- through octupoles at all atoms and bond midpoints. In addition, dipole polarizability tensors are evaluated for each valence LMO. Second, the multipoles at duplicate expansion points are removed, as are polarizability tensors at duplicate LMOs. The remaining parameters and expansion points are then combined into a single EFP. Third, the monopoles in the final protein EFP (built from many pieces) are scaled to reflect the net integer charge of the protein.

Remove-and-Expand Method

In step 1 of the remove-and-expand (RE) method, duplicate LMOs are removed together with an assigned local nuclear charge (+2 for a doubly occupied core or lone-pair LMO or +1 for bonds) positioned at the nucleus or nuclei on which the LMOs are predominantly situated. The remaining LMOs are used to construct the density and, in step 2, the MEP is expanded as before. The expansion points in the region of overlap are chosen to be the atomic centers (and associated bond midpoints) with a nuclear charge (and associated bond midpoints) with a nuclear charge (Fig. 3). In step 3, the multipoles on duplicate expansion points are combined, and both result in a protein EFP with a net integer charge. The polarizability tensors are treated as in the ERS method.

Expand-Collect-and-Correct Method

Step 1 of the expand-collect-and-correct (ECC) method is identical to the first step of the ERS method. In step 2 dipoles, quadrupoles, and octupoles on duplicate expansion points are removed, while the corresponding monopoles are combined (giving a total charge q) and moved to the nearest remaining expansion point (Fig. 4). In step 3, the EFP pieces are combined, and at that point the method corresponds to that proposed by Bellido and Rullmann.²³

However, when two pieces (A and B) are combined the two collected charges (q_A and q_B , at \mathbf{r}_A and \mathbf{r}_B , respectively) introduce an extraneous dipole (at some origin \mathbf{r}_O)

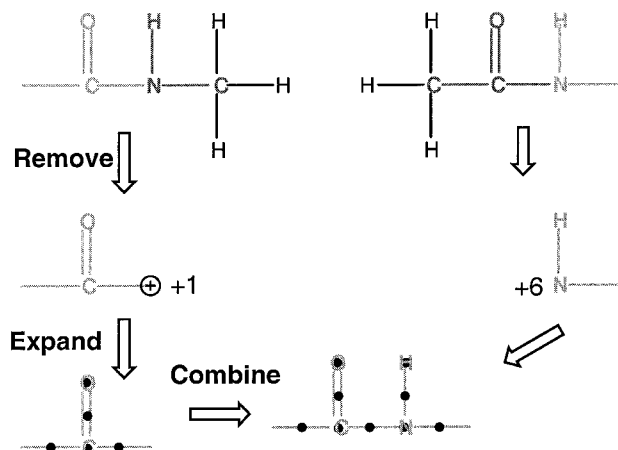


Figure 3. Schematic representation of the RE method described in the text. “+1” and “+6” indicate assigned local nuclear charges, as described in the text.

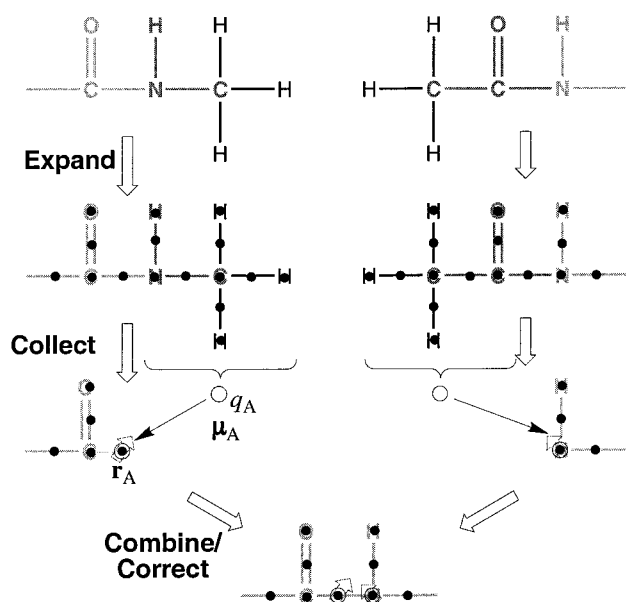


Figure 4. Schematic representation of the ECC method described in the text. The open circles and corresponding arrows represent the net charge of the overlap region and associated dipoles [eq. (2)], respectively.

$$\mu_o = q_A(\mathbf{r}_A - \mathbf{r}_o) + q_B(\mathbf{r}_B - \mathbf{r}_o) \quad (6)$$

The error in the MEP introduced by this dipole can be decreased by adding an opposing dipole ($-\mu_o$) to the MEP. Computationally, it is most convenient to compute dipole contributions independently for each piece:

$$\mu_o \approx \frac{1}{2}\mu_A + \frac{1}{2}\mu_B = \frac{1}{2}q_B(\mathbf{r}_B - \mathbf{r}_A) + \frac{1}{2}q_A(\mathbf{r}_A - \mathbf{r}_B) \quad (7)$$

Thus, when computing the multipoles of piece *A* a dipole given by $-\frac{1}{2}q_A(\mathbf{r}_A - \mathbf{r}_B)$ is placed at \mathbf{r}_A . The dipole is placed at \mathbf{r}_A , rather than \mathbf{r}_B , by adding the new dipole to the dipole already calculated for \mathbf{r}_A to avoid the introduction of an additional expansion point. Thus, a large EFP can be constructed simply by combining files, and the dipole correction does not result in additional calculations when the EFP is used.

Results and Discussion

Comparison of DAQ Methods: Proton Affinity of Lys55

The new DAQ methods described in the previous section have several practical advantages compared to the previously used ERS scaling method: (1) No scaling of charges is involved; (2) the additivity of the error in the charges and the distance dependence of this error is therefore not offset by scaling; (3) combination with force field regions representing the outer molecular environment is now possible.

The accuracies of the DAQ methods are compared for an EFP describing the protein environment within a 14-Å radius of N^ϵ of Lys55 [Fig. 5(a)]. This environment consists of two spatially distinct protein chains, composed of residues 29–34 [Fig. 5(b)] and 19–24_56–53 [Fig. 5(c)], where cysteine residues 24 and 56 are connected by a disulfide link. The 14-Å EFP generated by combining EFP from two separate RHF/6-31G(*d*) calculations on 29–34 and 19–24_56–53 results in a Lys55 PA of 231.47 kcal/mol. The PA is evaluated as follows. The EFP, buffer, and *ab initio* regions are combined and the geometry of the *ab initio* region is optimized at the RHF/6-31G(*d*) level of theory while the geometry of the buffer and EFP regions are fixed. In a second calculation a proton is removed from the amine group and the geometry of the *ab initio* region is reoptimized. The RHF/6-31G(*d*) energy differ-

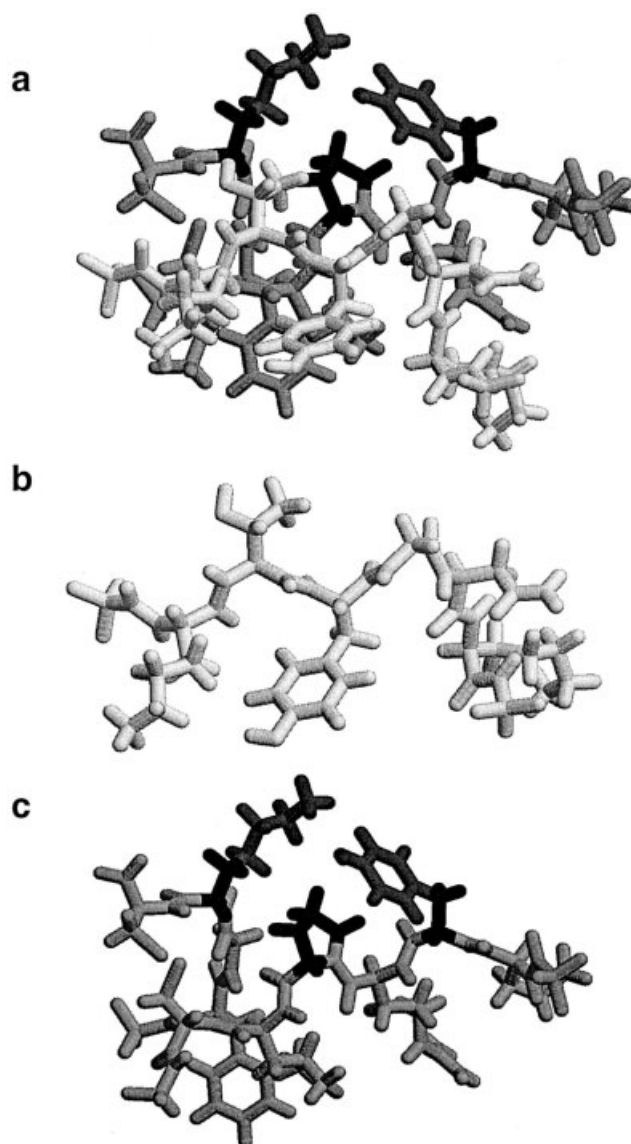


Figure 5. (a) The 14-Å EFP as a superposition of (b) chain 29–34 and (c) chain 19–24_54–56.

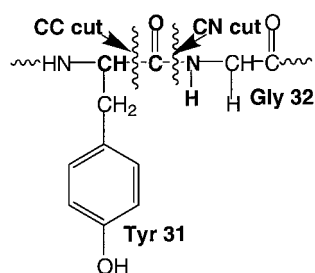


Figure 6. Illustration of the CC and CN cuts.

ence between the two geometries is taken as the PA. The PA value of 231.47 kcal/mol [PA(0)] will serve as our reference for the following overlap tests. Polarizability tensors were not included in these calculations to test the transferability of multipoles and polarizability tensors separately. Next, we analyze several aspects of the assembling procedure.

The 29–34 piece [Fig. 5(b)] of 14-Å EFP was built from two smaller pieces (29–31 and 32–34) using the ERS, RE, and ECC methods. Two different cuts between Tyr31 and Gly32 were studied: (1) the “CC” cut (between the α -carbon and the carbonyl group) and (2) the “CN” cut (at the peptide bonds) as shown in Figure 6. The distances from the location of the cuts to N^{ζ} of Lys55 are 10.2 and 9.7 Å, respectively (Fig. 7). PAs calculated using the CC and CN cut with the three DAQ methods are shown in Table 1. For the CC cut, the error is 0.09 kcal/mol for the ECC method and 0.13 kcal/mol for the RE method. For the CN cut, the error with respect to PA(0) increases to 0.2 kcal/mol for both DAQ methods, which is somewhat larger than the error of 0.09 kcal/

Table 1. Error with Respect to PA(0) as a Function of the Method, Cut Location, Distance Toward (N^{ζ}) of Lys55, and Maximum Error in the Charges.

Distance	Method	Cut	Max. error in charge	Error in PA(0)
9.7	ERS	CN		0.09
9.7	RE	CN	0.06	-0.20
9.7	ECC	CN	0.24	-0.20
9.7	EC	CN	0.24	-0.31
10.2	RE	CC	0.03	-0.13
10.2	ECC	CC	0.10	-0.09
10.2	EC	CC	0.10	0.17
19.2	RE	CN		-0.09
19.2	ECC	CN		-0.09
19.2	EC	CN		-0.11
19.2	RE	CC		-0.09
19.2	ECC	CC		-0.08
19.2	EC	CC		-0.03

mol¹² introduced by the ERS method for the same cut. The greater error observed for the CN cut compared to the CC cut for both the RE and ECC is in agreement with previous findings^{16,25} for a cut between the α -carbon and the carbonyl group.

The “remove-and-collect” (RE) method, which is the REC method without the dipole correction, yields larger errors for both cuts (0.17 for the CC cut and -0.31 for the CN cut), indicating the importance of the dipole correction in the REC approach.

The monopoles in the overlap region of the DAQ-constructed fragments can be compared to the monopoles of the 29–34 frag-

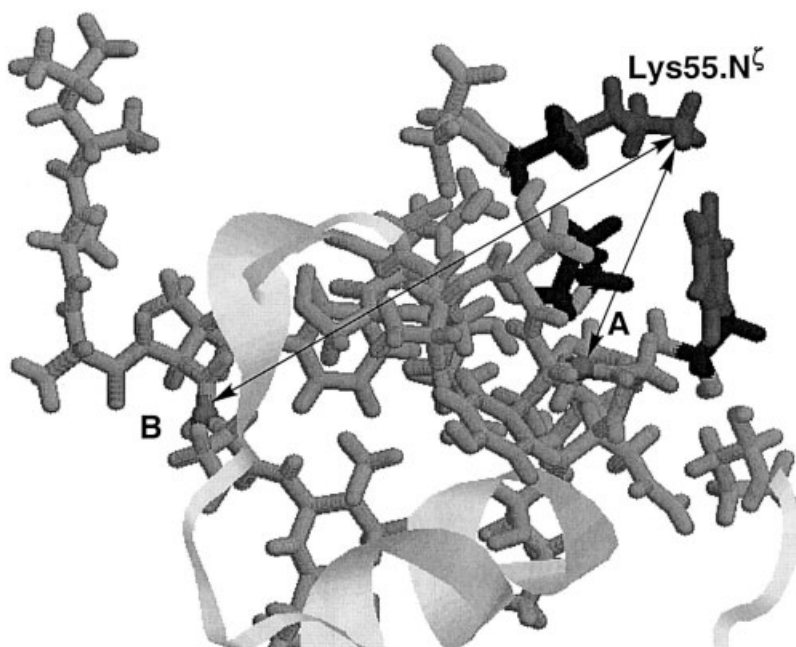


Figure 7. Location of the CC and CN cuts in OMTKY3 at two different distances from N^{ζ} . For A, CN is at 9.7 Å and CC is at 10.2 Å. For B, both CN and CC are at 19.2 Å.

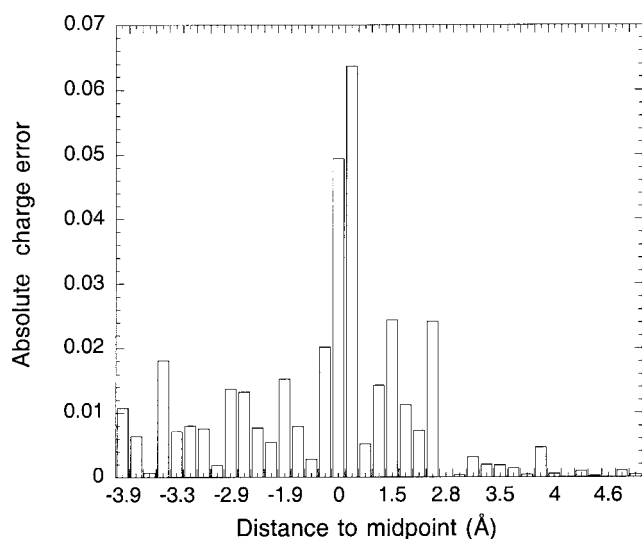


Figure 8. Plot of the error in the charges as a function of distance to the peptide midpoint for cut CN using the RE assembling method. The plot is generated by comparing the charges to identical charges in a non-overlap fragment.

ment to define a charge error. As expected, the error is largest in the charges closest to the center of the overlap region. A representative example is shown in Figure 8. The error in the PA due to the charge errors will depend on both their magnitude and spatial distribution. Accordingly, there is no clear correlation between the maximum error in the charge and PA. For example, the maximum error in the charges for cut CN is 0.06 for the RE method and 0.24 for ECC. Still, ECC results in lower errors in the PA compared to RE.

In summary, we note that all three DAQ methods considered here yield satisfactorily low errors in the PA. Similarly, both the CC and CN cuts yield satisfactory results. Thus, the optimum choice of cut location and DAQ method can be decided based on convenience.

A second region including residues 1–7 was added to the 14-Å region to investigate the distance dependence of the error in the PA (see Fig. 7). The fragment 1–7 was built from two smaller pieces, 1–4 and 5–7. The midpoint in the overlapping region between residues 4 and 5 is located about 19.2 Å from N^ε of Lys55 for both the CC and CN cuts. For both the RE and ECC methods the error appears to converge as the distance toward N^ε of Lys55 increases. All errors relative to PA(0) are less than 0.1 kcal/mol. Note also that at long distances neglect of dipole correction in the ECC method has no effect on the magnitude of the error.

pK_a of Lys55

Comparison of DAQ Methods

To provide a more realistic test of the DAQ methods, we studied their accuracy by calculating the pK_a of Lys55. The accuracy of the DAQ methods are now judged by direct comparison to experiment, and the effect of solvation on errors in the electrostatic potential are considered. The pK_a of Lys55 is computed by constructing a QM/buffer/EFP model as described above. The EFP is constructed with all three DAQ methods using the CN cut discussed in the previous section. The CN cut was chosen to facilitate the combined use of EFP and standard force fields, as described below. The construction of the EFP region comprises nine pieces: 19–24, 53–56, and 29–34—the two pieces within a 14-Å sphere [Fig. 1(b)]—and these other pieces that model the rest of the molecular environment: 1–4, 9–15, 25–28, 5–8, 38–35, 15–18, 39–44, and 45–52. The underscore indicates that residues are

Table 2. Computed pK_a s Using Different DAQ Methods and MM Regions [eqs. (4) and (5)].

Method	$\Delta E^{\text{MP2/RHF}}$	ΔG_{therm}	ΔG_s	pK_a
Experimental				11.1
Comparison of DAQ methods				
All EFPs (ERS)	26.10	1.03	−26.08	11.4
All EFPs (RE)	27.81	0.69	−27.7	11.2
All EFPs (ECC)	27.98	0.69	−27.7	11.3
Comparison of electrostatic models				
14-Å EFPs (RE) + CHARMM	24.45	0.70	−24.99	10.7
14-Å EFPs (RE) + OPLSAA	24.98	0.67	−25.33	10.8
14-Å EFPs (RE) + AMBER	24.85	0.67	−25.25	10.8
14-Å EFPs (ECC) + AMBER	25.03	0.67	−25.27	10.9
14-Å EFPs (ECC) + AM1	29.67	0.69	−30.62	10.4
9-Å EFPs (ECC) + AMBER	25.77	0.68	−25.17	11.5
All AMBER	38.10	0.47	−40.06	9.5
All CHARMM	39.88	0.63	−42.00	9.5
All OPLSAA	39.93	0.62	−41.69	9.8
All atom-centered EFP (ECC)	35.35	0.41	−30.99	14.1

Methodologies are presented in decreasing order of theoretical sophistication. In the “four-layer” approaches, the DAQ method used to treat the EFP subregion is shown in parentheses.

connected by a disulfide bridge. The average distance of the overlapping regions to N^ε of Lys55 is 18.5 Å. The EFP is then added to the QM and buffer regions and used to compute $\Delta E^{\text{MP2//RHF}}$, ΔG_{therm} , and ΔG_{sol} [cf. Eq (5)], which, when combined, yield the p*K*_a via eq. (6).

The p*K*_as calculated with EFPs constructed using the ERS, RE, and ECC methods are listed in Table 2. All three methods yield similar p*K*_as, all within 0.3 pH units of the experimental value of 11.1. Interestingly, the RE and ECC methods give energy components that are similar, while the ERS values differ by as much as 1.9 kcal/mol from their RE and ECC counterparts. Because the gas-phase and solvation energy terms are lower and higher, respectively, by roughly the same amount, the p*K*_a is not affected significantly. Thus, the PA is seen to be more sensitive to the choice of DAQ method than the p*K*_a, presumably due to solvent screening of the error. The solvation effect on differences in gas-phase properties is further illustrated for the combined use of EFP and standard force fields, as discussed next.

Comparison of Electrostatic Models

One of the main advantages of the RE and ECC methods (when used in conjunction with the CN cut) is that they allow for an easy interface with the atom-centered charges from common biomolecular force fields, such as AMBER, CHARMM, and OPLSAA. Here, we use this interface to test the use of these force field charges to model regions of the protein far removed from the ionizable residue of interest.

Table 2 lists the p*K*_a computed by a QM/buffer/EFP/MM model in which the protein environment within 14 Å is treated by an EFP while the rest of the protein is treated with either AMBER, CHARMM, or OPLSAA charges. The junction between EFP and MM is treated by the RE method. The p*K*_as predicted by this “four-layer” approach are all within 0.4 pH units of the experimental value, compared with a maximum error of 0.3 pH units for the all-EFP approaches. Thus, the EFP does not need to be calculated *ab initio* for the entire protein, leading to significant CPU time savings. The charges from all three force fields appear of equal quality.

Both DAQ methods work equally well for the EFP/MM interface. EFP/AMBER interfaces constructed with the RE and ECC method yield nearly identical p*K*_as of 10.8 and 10.9 pH units, respectively. However, using another computationally inexpensive approach such as AM1 for the > 14-Å region leads to a less satisfactory p*K*_a of 10.4.

Decreasing the EFP region to 9 Å and using the AMBER force field for the rest of the protein does not increase the error appreciably. Still, the sole use of force field charges increases the error by as much as 1.6 pH units. It is interesting to note that the “error” (relative to the all-EFP value) in the gas-phase PA is as much as 12 kcal/mol but roughly 10 kcal/mol of this error is “screened” by the PCM. The error is likely due to inherent limitations in the atom-centered charge model rather than the numerical values of the charges themselves. For example, using an all-EFP representation consisting only of atom-centered charges (but otherwise calculated as before) results in a p*K*_a error of 3 pH units. The better performance of the MM charges is presumably due to the underlying parameterization of the force fields against high-level *ab initio*

calculations. In all cases discussed so far the effect of the various representations of the protein electrostatic potential have small (≥ 0.62 kcal/mol) effects on the thermochemical energy contribution to the p*K*_a.

Conclusions

Two DAQ approaches for the construction of EFPs without the use of charge scaling are presented and compared to a previously developed DAQ method that requires charge scaling (Fig. 2).¹² In the RE (Fig. 3) method, select LMO contributions to the total density are removed prior to the multipole expansion. Alternatively, in the ECC method (Fig. 4) select charges are collected at the nearest expansion point (rather than being removed) and an additional dipole term is added to correct the electrostatic potential. Without the correction step, this method corresponds to the neutralization method of Bellido and Rullmann.²³

The methods are compared by computing the PA and p*K*_a of Lys55 in OMTKY3 for two different kinds of cuts on the protein backbone (at the CN and C_α—C bonds; Fig. 7). Both DAQ methods give rise to acceptably low errors for both the PA (< 0.2 kcal/mol, Table 1) and p*K*_a (< 0.2 pH units, Table 2) for both cut locations.

The new DAQ methods can easily be used in conjunction with standard force fields when using the CN cut. We show that the use of CHARMM, AMBER, and OPLSAA charges to describe the molecular electrostatic potential of the protein > 9 Å from Lys55 give p*K*_as within 0.4 pH units of experiment. However, using the force field charges for the entire MM region results in p*K*_a errors of up to 1.6 pH units. Use of *ab initio*-computed atomic charges give an even larger error, which indicates that the atom-centered charge model may be inadequate at short range in this case, consistent with previous findings.^{2–14}

The PCM method is shown to screen relatively large errors in the gas-phase energy introduced by the various approximate treatments of the protein electrostatic potential (different DAQ methods, force fields, etc.). Thus, tests of DAQ methods,^{1,17,24,25} or other aspects of the QM/MM methodology,^{44,45} by comparing gas-phase energetics are perhaps overly stringent.

Acknowledgments

This work was supported by a Research Innovation Award from the Research Corporation and the National Science Foundation (MCB 0209941). H.L. gratefully acknowledges a predoctoral fellowship from the Center for Biocatalysis and Bioprocessing at the University of Iowa. Calculations were performed on IBM RS/6000 workstations obtained through a CRIF grant from the NSF (CHE-9974502).

References

1. Buckingham, A. D.; Fowler, P. W. *J Chem Phys* 1983, 79, 6426–6428.
2. Buckingham, A. D.; Fowler, P. W. *Can J Chem* 1985, 63, 2018–2025.

3. Stone, A. J.; Price, S. L. *J Phys Chem* 1988, 92, 3325–3335.
4. Price, S. L.; Harrison, R. J.; Guest, M. F. *J Comput Chem* 1989, 10, 552–567.
5. Sokalski, W. A.; Maruszewski, K.; Harihan, P. C.; Kaufman, J. J. *Int J Quantum Chem Quant Biol Symp* 1989, 16, 119.
6. Faerman, C. H.; Price, S. L. *J Am Chem Soc* 1990, 112, 4915–4926.
7. Price, S. L.; Richards, N. G. J. *J Comp-Aided Mol Design* 1991, 5, 41–54.
8. Wiberg, K. B.; Rablen, P. R. *J Comput Chem* 1993, 14, 1504–1518.
9. Dudek, M. J.; Ponder, J. W. *J Comput Chem* 1995, 16, 791–816.
10. Dixon, R. W.; Kollman, P. A. *J Comput Chem* 1997, 18, 1632–1646.
11. Kosov, D. S.; Popelier, P. L. A. *J Chem Phys* 2000, 113, 3969–3974.
12. Minikis, R. M.; Kairys, V.; Jensen, J. H. *J Phys Chem A* 2001, 105, 3829–3837.
13. Tiraboschi, G.; Fournie-Zaluski, M. C.; Roques, B. P.; Gresh, N. *J Comput Chem* 2001, 22, 1038–1047.
14. Matta, C. F.; Bader, R. F. W. *Proteins* 2000, 40, 310–329.
15. Sokalski, W. *Amino Acids* 1994, 7, 19–26.
16. Dardenne, L. E.; Werneck, A. S.; Neto, M. O.; Bisch, P. M. *J Comput Chem* 2001, 22, 689–701.
17. Kedzierski, P.; Sokalski, W. A. *J Comput Chem* 2001, 22, 1082–1097.
18. Gresh, N. *J Chim Phys Phys Chim Biol* 1997, 94, 1365–1416.
19. Koch, U.; Stone, A. J. *J Chem Soc Faraday Trans* 1996, 92, 1701–1708.
20. Price, S. L.; Stone, A. J. *J Chem Soc Faraday Trans* 1992, 88, 1755–1763.
21. Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J Chem Phys* 1996, 105, 1968–1986.
22. Bandyopadhyay, P.; Gordon, M. S. *J Chem Phys* 2000, 113, 1104–1109.
23. Bellido, M. N.; Rullmann, J. A. C. *J Comput Chem* 1989, 10, 479–487.
24. Vignemaeder, F.; Claverie, P. *J Chem Phys* 1988, 88, 4934–4948.
25. Young, L.; Topol, I. A.; Rashin, A. A.; Burt, S. K. *J Comput Chem* 1997, 18, 522–532.
26. Molina, P. A.; Sikorski, R. S.; Jensen, J. H. *Theor Chem Acc* 2003, 109, 100–107.
27. Li, H.; Hains, A. W.; Everts, J. E.; Robertson, A. D.; Jensen, J. H. *J Phys Chem B* 2002, 106, 3486–3494.
28. Hoogstraten, C. G.; Choe, S.; Westler, W. M.; Markley, J. L. *Protein Sci* 1995, 4, 2289–2299.
29. Kairys, V.; Jensen, J. H. *J Phys Chem A* 2000, 104, 6656–6665.
30. King, H. F.; Stanton, R. E.; King, H.; Wyatt, R. E.; Parr, R. G. *J Chem Phys* 1967, 47, 1936.
31. Stevens, W. J.; Fink, W. H. *Chem Phys Lett* 1987, 139, 15–22.
32. Bagus, P. S.; Hermann, K.; Bauschlicher, C. W. *J Chem Phys* 1984, 80, 4378–4386.
33. Stone, A. J. *Chem Phys Lett* 1981, 83, 233–239.
34. Li, H.; Jensen, J. H. *Theor Chem Acc* 2002, 107, 211–219.
35. Head, J. D. *Int J Quantum Chem* 1997, 65, 827–838.
36. Vreven, T.; Mennucci, B.; da Silva, C. O.; Morokuma, K.; Tomasi, J. *J Chem Phys* 2001, 115, 62.
37. Cossi, M.; Mennucci, B.; Cammi, R. *J Comput Chem* 1996, 17, 57–73.
38. Barone, V.; Cossi, M.; Tomasi, J. *J Chem Phys* 1997, 107, 3210–3221.
39. Li, H.; Pomelli, C. S.; Jensen, J. H. *Theor Chem Acc* 2003, 109, 71–84.
40. Boys, S. F. In: Lowdin, P. O., ed. *Quantum Science of Atoms, Molecules and Solids*; Academic Press: New York, 1966, 283.
41. Edmiston, C.; Ruedenberg, K. *Rev Mod Phys* 1963, 35, 457.
42. Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J Comput Chem* 1993, 14, 1347–1363.
43. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98 revision A.6*; Gaussian, Inc.: Pittsburgh, PA, 1998.
44. Lennartz, C.; Schafer, A.; Terstegen, F.; Thiel, T. *J Phys Chem B* 2002, 106, 1758–1767.
45. Cui, Q.; Karplus, M. *J Phys Chem B* 2002, 106, 1768–1798.