Accurate Intraprotein Electrostatics Derived from First Principles: An Effective Fragment Potential Method Study of the Proton Affinities of Lysine 55 and Tyrosine 20 in Turkey Ovomucoid Third Domain

Ryan M. Minikis, Visvaldas Kairys,† and Jan H. Jensen*

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

Received: September 13, 2000; In Final Form: December 21, 2000

A divide-and-conquer method by which an accurate static and induced multipole representation of the electrostatic potential of a protein can be generated using ab initio electronic structure theory is presented. The method is applied to the generation of an effective fragment potential (*J. Chem. Phys.* **1996**, *105*, 1968) for the protein turkey ovomucoid third domain. Dipoles and induced dipoles are necessary for accurate intraprotein electrostatics, as measured by their effects on the gas-phase proton affinities (PAs) of the amino acid residues lysine 55 (Lys55) and tyrosine 20 (Tyr20). Deprotonation of Tyr20 is predicted to result in *spontaneous* proton transfer from Lys55 to Tyr20, which thus have identical PAs. It is suggested that the experimentally measured (identical) pK_as of Tyr20 and Lys55 might be identical for the same reason.

I. Introduction

Electrostatics is generally believed to be the principal force determining the structure and function of proteins.[1] Thus many biomolecular force fields treat all interactions of atoms separated by more than two bonds by long-range charge−charge interactions plus short-range van der Waals terms (e.g. a 6-12 potential).[2] The successes of this approach in modeling biomolecular systems, using for example the AMBER,[3] CHARMM,[4] and GROMOS[5] force fields, are impressive.

However, comparisons to ab initio calculations on model systems reveal that the atom-centered charge model is not always an adequate representation of the molecular electrostatic potential (MEP).[6−9] For example, interactions that involve orbitals of peptide bonds or aromatic side chains may require either additional charges[10] or higher order multipoles.[11] Similarly, models based on atom-centered charges tend to underestimate the directionality of hydrogen bonds,[12] while models that include additional charges[9] or higher order multipoles[13] reproduce ab initio results well.

Intermolecular interaction potentials that make use of higher order multipoles have been used extensively to model solute−solvent interactions, crystal structures, and hydrogen bonding between relatively small biological molecules.[14] Higher order multipoles are also used in the effective fragment potential (EFP) method,[15] a hybrid QM/MM method in which only the active part of a molecular system is treated with ab initio quantum mechanics while the rest is replaced by one or more EFPs. An EFP represents the static electrostatic potential by a distributed multipole expansion[16] (charges through octupoles at all atomic centers and bond midpoints), while the electronic polarizability is represented by dipole polarizability tensors for each valence (localized) molecular orbital.[17]

Multipole expansions for interaction potentials are usually derived from electron densities calculated with ab initio electronic structure methods, using, for example, Stone's distributed multipole analysis[16] (DMA) or Bader's atoms-in-molecules[18] method. These ab initio-derived multipolar representations of a MEP (mMEP) can be systematically improved by using better electronic structure methods. However, for a protein one must address the methodological issue of how to obtain the mMEP for a system that is too large to be treated by a single ab initio calculation.

One approach to obtaining a mMEP for a protein is to generate a library of mMEPs for amino acid residues by calculations on smaller representative systems and investigate the transferability to larger systems. Work by Stone,[19] Price,[20] Bader,[21] and their co-workers have identified two factors that limit the transferability. One is the conformational dependence of the multipoles, and the other is the perturbation of the multipoles by intraprotein hydrogen bonding. The very recent work by Matta and Bader[21] is encouraging, since the multipoles calculated within the atoms-in-molecules approach appear less sensitive to conformational effects than those from the DMA approach, though intramolecular hydrogen bonds must still be identified and appropriately dealt with.

Another, more immediate, approach is to generate a mMEP specifically for a given protein by a divide-and-conquer approach. In this approach 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.

In this paper we investigate the use of the divide-and-conquer approach to generate an EFP representation[22] of the 56-residue protease inhibitor turkey ovomucoid third domain (OMTKY3). Two key issues for this approach are addressed: (1) the size of the region of overlap and (2) the efficient computation of the EFP parameters so that the protein can be divided into as few large pieces as possible.

The paper is organized as follows:

First, the general EFP methodology is outlined.

Second, a new and more efficient method for calculating the localized molecular orbital polarizabilities is introduced. With

* Corresponding author. E-mail: jan-jensen@uiowa.edu.
† Present address: Center for Advanced Research in Biotechnology, 9600 Gudelsky Dr., Rockville, MD 20850.

10.1021/jp0032659 CCC: $20.00 © 2001 American Chemical Society
Published on Web 03/21/2001

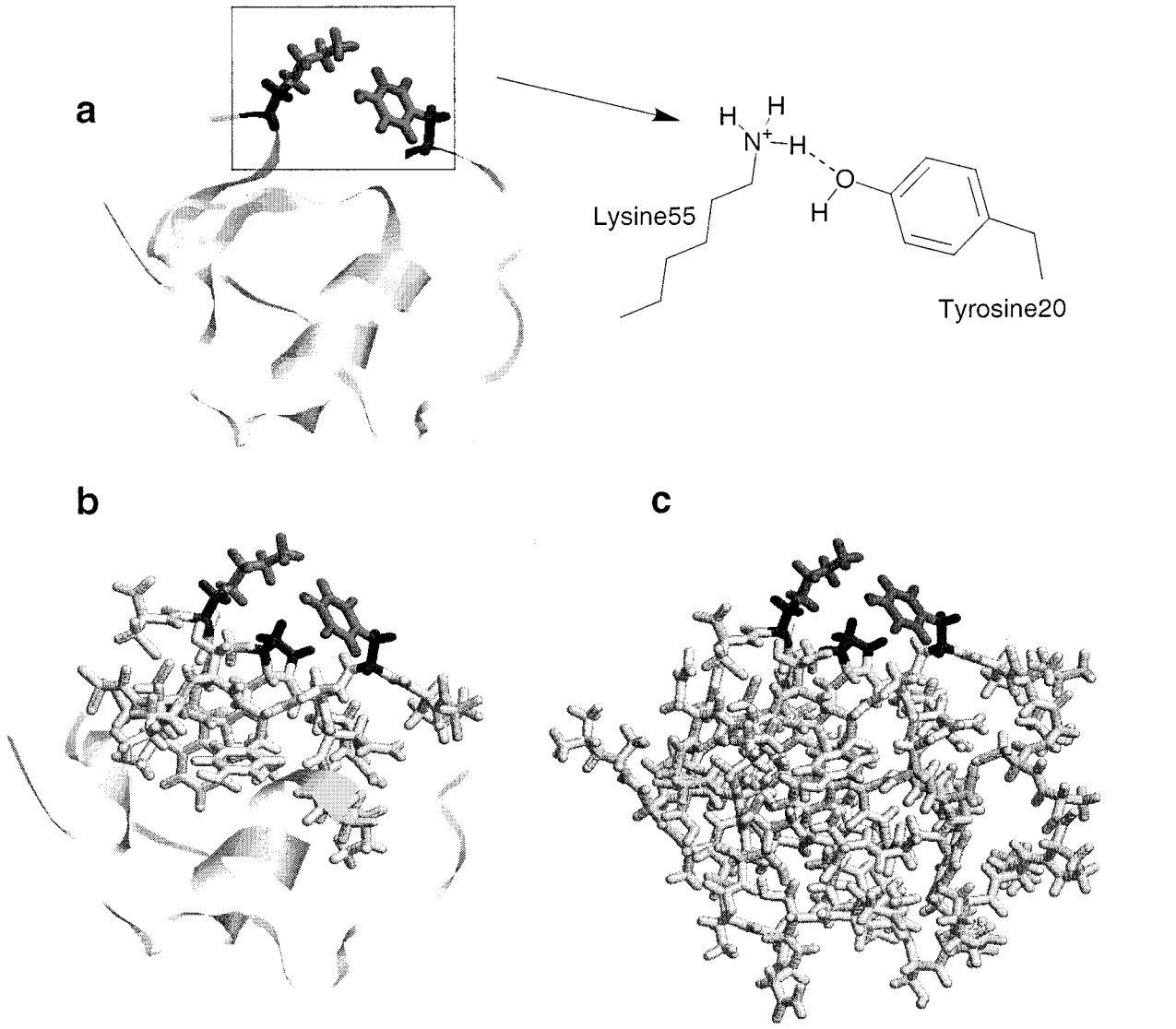


Figure 1. (a) OMTKY3 Lys55-Tyr20 ab initio region including buffer regions (in dark) with detail. The remainder of the protein is ribbon structure. (b) Lys55-Tyr20 ab initio region with the surrounding 14 Å radius EFP region as measured from the Lys55 Nζ atom. Note that the proline section is now in buffer (see Figure 2 for detail). (c) Similar to part b, except now with the entire protein as EFP.

this addition the CPU time required for the generation of the EFP parameters is essentially negligible compared with that of the SCF itself.

Third, the EFP corresponding to the protein environment within a 14 Å radius of lysine 55 (Lys55) in OMTKY3 is computed using various choices of overlap between protein pieces. The proton affinity (PA) of Lys55 is then calculated using these EFPs and used to select the optimum strategy for EFP generation. PAs have been used previously by us[23] and others[24] as a sensitive measure of the accuracy with which the molecular environment is modeled.

Fourth, this strategy is then used to create the EFP parameters for the remaining protein to yield an accurate nonempirical treatment of the internal electrostatics of the protein.

Fifth, the relative effects of intramolecular hydrogen bonding and longer range interactions within the protein on the PA of Lys55 are discussed.

Sixth, the relative PAs of Lys55 and Tyr20, as well as a possible reinterpretation of the experimentally measured pK_as of these residues, are discussed.

Finally, we summarize our findings and discuss future directions.

II. Computational Methodology

The solution structure of OMTKY3 has been determined using NMR by Hoogstraten et al.[25] 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. The electronic and geometric structures of the Lys55 and Tyr20 side chains are treated quantum mechanically at the RHF/6-31G(d)[26] level of theory, while the rest of the protein is treated with an EFP (described in more detail below). Both residues are included since they are connected by a short, strong hydrogen bond (see Figure 1a), which influences the proton affinities (PAs). The ab initio region is separated from the protein EFP by a buffer region[23] comprised of frozen localized molecular orbitals (LMOs) corresponding to the $C_\alpha-C_\beta$ bonds of Lys55 and Tyr20 and the associated CH and core LMOs, as well as part of the Pro22 ring. Our previous work[23] has shown that placing the buffer region at the $C_\alpha-C_\beta$ bond yields proton affinities within 0.5 kcal/mol of the all ab initio reference value for the tripeptide glycyl-lysyl-glycine (Gly-Lys-Gly). The Pro22 buffer is needed to describe its short-range interactions with Tyr20. The buffer LMOs are generated by an RHF/6-31G(d) calculation on a subset of the system (shown in Figure 2),

projected onto the buffer atom basis functions[27] and subsequently frozen in the EFP calculations by setting select off-diagonal MO Fock matrix elements to zero.[28,29] The ab initio/buffer region interactions are calculated ab initio and thus include short-range interactions. Other buffer regions are used for analysis purposes, as described in section III.D.

The EFP describing the rest of the protein is generated by nine separate ab initio calculations on overlapping pieces of the protein truncated by methyl groups. Two different regions of overlap are used, depending on whether it occurs on the protein backbone or on a disulfide bridge, as described in section III.B.

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.[16] The monopoles of the entire EFP are scaled to ensure a net integer charge, as described in section III.B. The dipole polarizability tensor due to each LMO in the EFP region is calculated by a perturbation expression described below.

For the protein piece containing the ab initio/buffer region, the density of the molecular region that will be described by the EFP is optimized in the presence of the frozen buffer region but in the absence of the ab initio region. The electrostatic potential of the optimized density, but not the buffer density, is expanded in terms of multipoles. Calculated in this way, these multipoles do not account for polarization of the EFP region due to the ab initio region, so that this effect is not double counted when dipole polarizabilities are added.

The EFP, buffer, and ab initio regions are combined, and the geometry of the ab initio region is reoptimized. In a second calculation the Nζ proton of lysine is removed and the geometry of the ab initio region is reoptimized. The energy difference between these two systems is taken to be the proton affinity.

The interaction energy between the EFP and buffer region is not calculated. Since the geometry of the EFP and buffer region remains unchanged in both calculations, only the induced-dipole/buffer interaction is changed during deprotonation, and this term is neglected in our calculations.

The Foster−Boys localization procedure was used throughout this work to generate localized orbitals,[30] and all calculations were done with the quantum chemistry code GAMESS.[31]

III. Results and Discussion

A. Computation of the Polarizability Tensors. The change in the electronic structure of the EFP region is modeled by polarizability tensors for each localized molecular orbital, as formulated by Garmer and Stevens.[15,17] These tensors are usually calculated by numerical differentiation of the electronic dipole of each LMO with respect to a weak (0.0001 aus) uniform field

$$\alpha_{fg}{}^{l} = \lim_{F_g \to 0} \frac{\mu_{ll}{}^{f}(F_g) - \mu_{ll}{}^{f}}{F_g} \approx \frac{-2(\langle \psi_l' | \hat{\mu}_f | \psi_l' \rangle - \langle \psi_l{}^{0} | \hat{\mu}_f | \psi_l{}^{0} \rangle)}{F_g} \quad (1)$$

where f and g refer to x, y, or z components, ψ_l' and $\psi_l{}^{0}$ are perturbed and unperturbed LMOs, respectively.

The application of this numerical approach to the computation of protein EFP parameters suffers from some practical shortcomings. The computation of the polarizability tensors requires

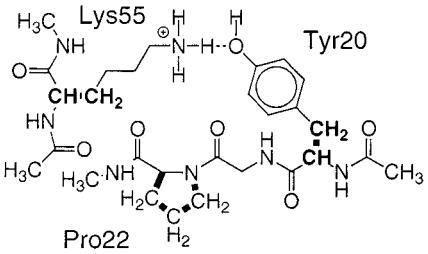


Figure 2. Subsystem of OMTKY3 used to obtain the buffer region (bold) used in this study.

three additional SCF calculations, compared to computing the static multipole expansions, which is a nontrivial consideration for the size of systems considered here. Additionally, the relatively weak perturbing field necessitates the use of stricter convergence criteria and more accurate integrals than for a normal SCF calculation. Thus, the actual CPU demands can actually increase 10-fold if polarizability tensors are to be included. Furthermore, even with strict convergence criteria, we have observed unphysical LMO polarizability tensors for conjugated systems such as phenylalanine side chains.

Webb and Gordon[32] have developed a method by which the LMO polarizability tensors can be calculated analytically from the corresponding canonical MO expression[33]

$$\alpha_{fg}{}^{i} = -4 \sum_{m}^{virt} U_{mi}^{g} \mu_{mi}^{f} \quad (2)$$

Here, the response functions are defined as[22]

$$\frac{\partial C_{\mu i}}{\partial F_g} = \sum_{m}^{\text{virt}} U_{mi}{}^{g} C_{\mu m} \quad (3)$$

and obtained by iteratively solving the coupled perturbed Hartree−Fock equation

$$U_{mi}{}^{g} = \frac{1}{\epsilon_i - \epsilon_m} \Big(\sum_{n}^{\text{virt}} \sum_{j}^{\text{do}} (4\langle mi|nj\rangle - \langle mn|ij\rangle - \langle mj|ni\rangle) U_{nj}{}^{g} + \mu_{mi}{}^{g} \Big) \quad (4)$$

The LMO equivalent of eq 2 is obtained by separately transforming the response functions and dipole integrals into the localized basis

$$\alpha_{fg}{}^{l} = -4 \sum_{m}^{\text{virt}} \Big(\sum_{i}^{\text{do}} T_{li} U_{mi}{}^{g} \Big) \Big(\sum_{j}^{\text{do}} T_{lj} \mu_{mj}{}^{f} \Big) \quad (5)$$

by using the transformation matrix that related the localized and canonical MOs

$$\psi_l = \sum_{i}^{\text{do}} T_{li} \phi_i \quad (6)$$

Calculated in this way, LMO polarizability tensors can be obtained using standard convergence criteria and integral packages, even for conjugated systems. However, solving the CPHF equations requires a partial two-electron integral transformation, which is very disk and memory intensive and thus also severely limits the size of the system that can be treated this way.

Here we propose a third approach to obtaining the LMO polarizability tensors, which retains the best features of both

methods, by approximating the response functions by a perturbation theory expression

$$U_{mi}^{g} \approx \frac{\mu_{mi}^{g}}{\epsilon_i - \epsilon_m} \tag{7}$$

Though this represents only the first iteration of the CPHF equation solution, intermolecular perturbation theory has been shown to give accurate total induction energies.[34] Here we demonstrate the utility of eqs 5 and 7 by calculating the PA of the tripeptide Gly-Lys-Gly, using polarizability tensors obtained by all three methods discussed above. This system has been used previously to demonstrate that the EFP results are relatively insensitive whether the polarizability tensors are calculated numerically or analytically (for nonconjugated systems, see Table 4 in ref 23). The relevant PAs are 232.1 and 232.2 kcal/mol, respectively, while the new perturbative method yields a PA of 232.0. All three values are thus within 0.2 kcal/mol of one another, and all are within 0.4 kcal/mol of the all ab initio value, 231.8 kcal/mol.

Since the computational cost of the perturbation method is the smallest of the three approaches (essentially negligible compared to a regular SCF calculation) and yields reasonable tensors for conjugated systems (data not shown), we use this new method for calculating polarizability tensors for the remaining calculations described in this paper.

B. Choice of Overlap. To determine what region of overlap is sufficient when building a protein EFP from smaller, computationally affordable protein pieces, we focus on the protein environment within a 14 Å radius of Lys55 (see Figures 1b and 3a). This environment consists of two spatially distinct protein chains (Figure 3), composed of residues 29−34 (Figure 3b) and 19−24—56−53 (Figure 3c, 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. This value 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 (see subsection 3 below).

(1) Overlap Along the Backbone. Residues 29−34 are used to test the overlap along the peptide backbone. A total of six sets of calculations with increasing amounts of overlap are carried out in order to determine the required amount of overlap to obtain the convergence of error in the proton affinity of Lys55 when calculated with the resulting EFPs (see Figure 4).Case 0 (Figure 4) involves no overlap and, when combined with chain 19−24—56−53, leads to the reference PA of 231.47 kcal/mol. The overlap in case 1 is a single peptide bond between Tyr31 and Gly32. The EFP parameters on the N-terminal side of this bond as well as those of the overlapping peptide bond midpoint are taken from the calculation on residues 29−31, while the parameters on the C-terminal side are taken from the calculation on residues 32−34. The final EFP describing residues 29−34 is then constructed by combining the EFP parameters from these two calculations and used together with chain 19−24—56−53 to recalculate the PA of Lys55. The new PA of 233.32 kcal/mol (Table 1) is 1.85 kcal/mol higher than the reference value, and this difference is taken to be the error due to differences in the EFP parameters resulting from end effects.

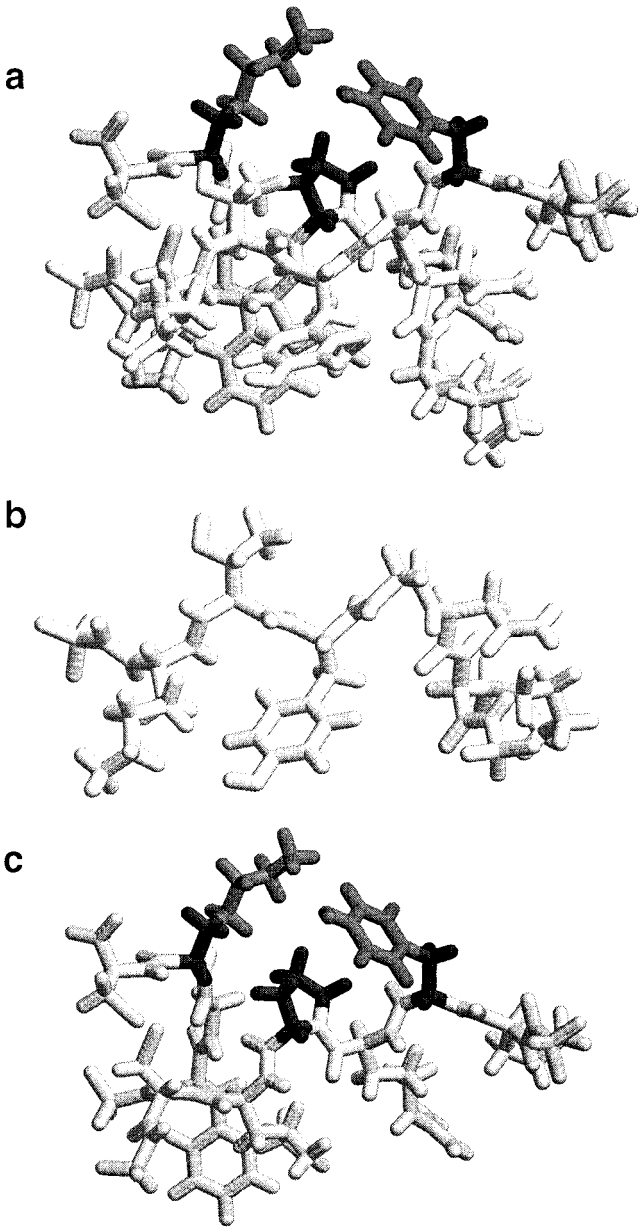


Figure 3. (a) The 14 Å EFP as a superposition of (b) chain 29−34 and (c) chain 19−24—54−56.

This process was repeated for increasing regions of overlap (cases 2−5 in Figure 4), and the results are listed in the second column of Table 1. In cases 2 and 4, where the midpoint of overlap is Tyr31, the parameters of the Tyr side chain were taken from the calculation N-terminus side, since the side chain is closest to other residues in that direction. Though there are some oscillations, the error converges to <1 kcal/mol^{-1} relatively quickly. However, even for case 5, where the overlap is considerable, the PA is still in error by 0.23 kcal/mol.

One marked difference in the EFP parameters obtained by calculations on overlapping pieces is that the monopoles no longer add up to a net integer charge. This is a well-known problem for empirical force fields and has been dealt with by scaling the charges. Here we scale the monopoles obtained in cases 1−5 to reproduce the overall integer charge of the system, *i*, by determining a scaling constant, *k*, for which

$$pk^{-1} + nk = i \tag{8}$$

where *p* and *n* are the sum of all positive and negative

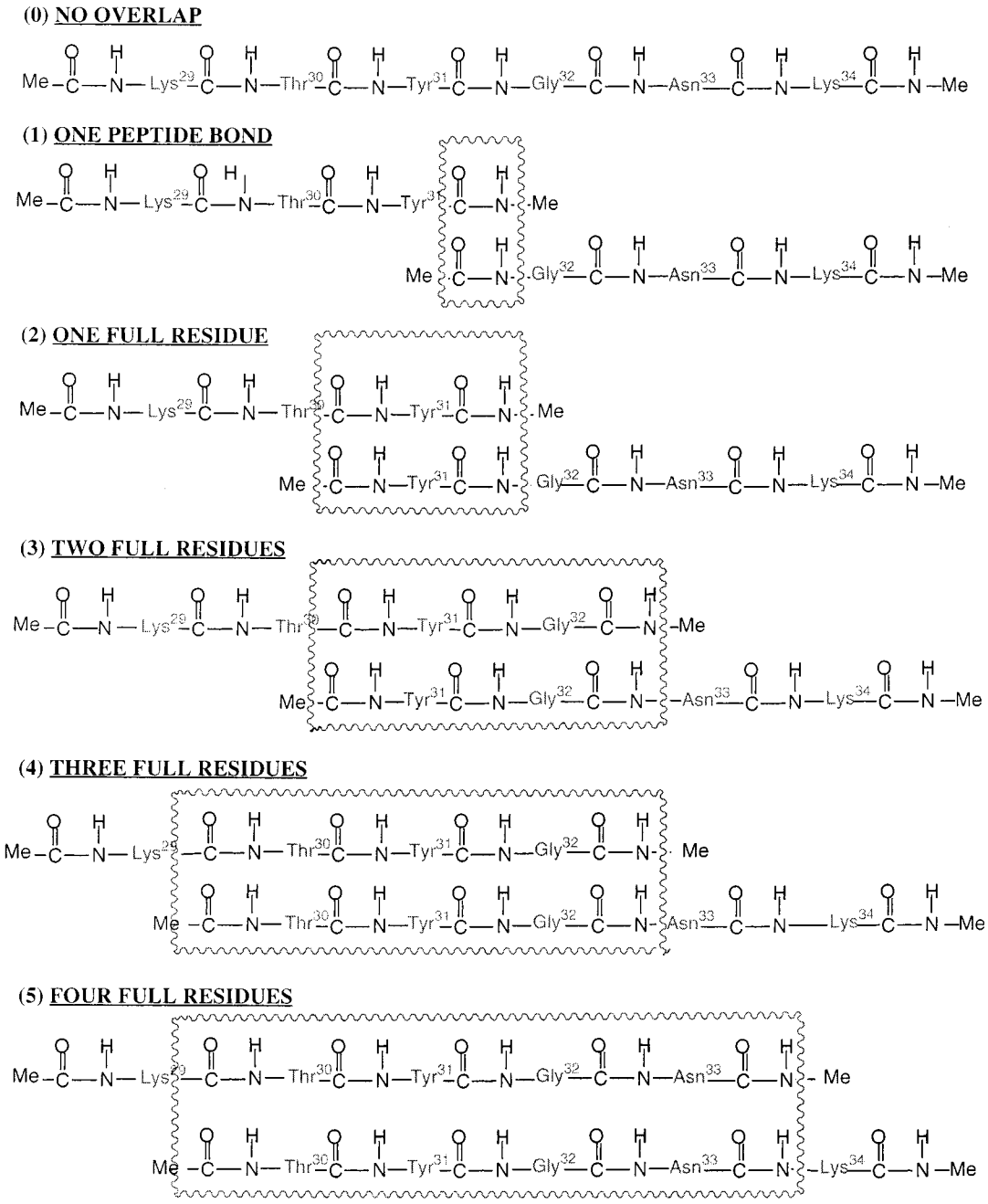


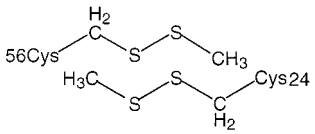
Figure 4. EFP overlap testing cases along the backbone using chain 29−34. Case 0 corresponds to Figure 3b.

TABLE 1: Proton Affinities of Lys55 in Various EFP Overlap Tests (cf. Figure 4; in kcal/mol)[a]

overlap case	without scaling	with scaling	overlap case	without scaling	with scaling
0	231.47	N/A	4	232.18	231.46
	0.0			0.71	−0.01
1	233.32	231.56	5	231.70	231.39
	1.85	0.09		0.23	−0.08
2	232.04	231.40	S−S	229.16	231.68
	0.57	−0.07		−2.31	0.21
3	231.53	231.35	one peptide	N/A	231.78
	0.05	−0.12	bond + S−S		0.31

[a] The upper number is the absolute proton affinity; the lower one is the error relative to the reference calculation, case 0.

monopoles, respectively. The resulting PAs are listed in the third column of Table 1, from which it is evident that the error quickly converges to within 0.1 kcal/mol for case 1. These results suggest that only one peptide bond of overlap is necessary to construct EFPs if scaling is used. This approach will be used for constructing the entire OMTKY3 EFPs.



(2) Disulfide Bridges. Disulfide bridges present another covalently bonded linkage between chains that must be dealt with through overlapping EFP calculations. Here we test a single overlap region, by recalculating the EFP for chain 19−24−56−53 using this $\beta-\beta$ overlap. This EFP chain is then combined with the 29−34 chain (case 0 above) and used to recalculate the Lys55 PA. Without monopole-scaling the resulting error is 2.31 kcal/mol (Table 1, case S−S), but as before, the error is significantly reduced (to 0.21 kcal/mol) by the scaling. Thus,

TABLE 2: Proton Affinities of Lys55 in Polarizability Overlap Tests[a]

overlap case	proton affinity
0′	233.78
1′	233.78
S−S′	233.79

[a] All values use scaled multipoles and are in kcal/mol.

the $\beta-\beta$ overlap with scaling will be used when constructing the rest of the EFP.

(3) Additivity of Error. As more than one region of overlap is utilized to build an EFP, the error induced on the PA will not necessarily be additive. This issue is tested by combining the one-peptide-bond (case 1) and $\beta-\beta$ disulfide overlap regions discussed previously and calculating the Lys55 PA. The resulting PA of 231.78 kcal/mol (see Table 1) is in error by 0.31 kcal/mol, which is approximately the sum of the 0.09 and 0.21 kcal/mol errors found for the two separate cases. In general, we expect the signs of the errors will be random, so that the overall error may be reduced by cancellation.

(4) Polarizability Tensors. Having determined adequate regions of overlap for the construction of the static multipolar part of the EFP, we now test whether they yield equally small errors for the polarizability tensors. The reference PA value (233.78 kcal/mol, Table 2) is calculated as before by constructing the 14 Å EFP from separate calculations on chains 29−34 and 19−24−56−53, and including the resulting polarizability tensors (Case 0′). Calculating the polarizability tensors, but not the multipoles, of chain 29−34 using the one-peptide-bond overlap (Case 1′), changes the Lys55 PA by <0.01 kcal/mol. Similarly, calculating the polarizability tensors of chain 19−24−56−53 using the $\beta-\beta$ overlap (Case S−S′), results in a PA error of only 0.01 kcal/mol. The overlap regions tested are thus adequate for calculating the polarizability tensors, which appear more transferable than the static multipoles. The good transferability is likley due to the use of LMOs, and we will investigate the use of LMOs in constructing the multipole expansion in future studies.

C. The PA of Lys55 in OMTKY3. The EFP parameters of the remaining part of the protein are calculated by seven additional calculations, using the overlap regions described above and scaling the monopoles to reflect the net −1 charge of the EFP region. The resulting ab initio/buffer/EFP calculation yields a Lys55 PA of 254.02 kcal/mol. Given the results of our previous calculations on tripeptides and the errors due to scaling demonstrated here (which likely decrease as the overlap regions occur further from the ab initio region), we estimate that this value is within 1.0 kcal/mol of the full RHF/6-31G(d) result.

Neglecting polarization introduces an error of 2.39 kcal/mol, and this term is thus crucial for an accurate PA. However, we note that this error is essentially identical to the 2.31 kcal/mol error due to neglecting polarization in the 14 Å EFP calculation, which suggests that this effect is relatively short-range in this case. Further neglect of the octupoles for the entire EFP introduces an error of only 0.09 kcal/mol, which demonstrates that *the multipolar representation of the static electrostatic potential of the protein is converged*. The quadrupoles and dipoles contribute 0.03 and 5.91 kcal/mol, respectively, indicating that the latter term is necessary for determining an accurate PA.

The use of higher order multipole terms does not result in a prohibitive computational cost. The average CPU time for an energy plus gradient of the system in Figure 1c is only 30 min on a four-node IBM 44P 270 RS/6000 workstation compared to 15 min without EFPs. Thus, the inclusion of monopoles through octupoles at 1553 EFP points leads to only a doubling of the CPU time. Furthermore, the CPU requirement is linear with respect to the number of EFP points, so that the addition of the 14 Å EFP (502 points) leads to a 36% increase in CPU time. In comparison, increasing the size of the ab initio/buffer region from the Lys55 side chain to include Tyr20 and part of Pro22 increases the CPU cost by 2500%.

D. Interpretation of the Proton Affinity. The shift in the acid/base properties of a residue induced by the protein is a measure of the intraprotein forces in that region and thus reflects the intricate relationship between protein structure and energetics. The EFP method can be used to extract this relationship by relating the PA of Lys55 to that of the isolated lysine side chain through the thermodynamic cycle displayed in Figure 5.

(1) The PA Shift Due To the Lys55···Tyr20 H-bond. The upper cycle of Figure 5 considers the effect of the Lys55···Tyr20 H-bond on the PA of the isolated lysine side chain residue. The lysine side chain residue is isolated by removing the EFP and all other MOs and nuclei (but not the basis functions), and the energy of the protonated and unprotonated structures are recalculated. The resulting intrinsic PA of 236.78 kcal/mol agrees reasonably well with the RHF/6-31G(d) PA of 233.86 kcal/mol for pentanamine.

Similarly, the buffer and ab initio region of Tyr20 can be added back on and used to compute a Lys55 PA of 247.35 kcal/mol. The presence of the H-bond therefore increases the PA by 11.52 kcal/mol, due to the fact that the protonated form of Lys55 can form a stronger H-bond with Tyr20 than the neutral form.

The latter assertion can be proved by calculating the strengths of the intramolecular H-bonds, by computing the respective energies of the H-bonded systems relative to the energy of the isolated lysine chain discussed above and a similarly computed energy of an isolated tyrosine chain. The energies of the respective buffer regions have been subtracted from the total energies to yield the energies of the ab initio region. Therefore, the H-bond strengths resulting from these corrected energies correspond to the interaction within the ab initio region and show that the 8.55 kcal/mol PA shift is a result of decreasing the Lys55···Tyr20 H-bond strength from 16.79 to 8.24 kcal/mol upon deprotonation (Figure 5). These values agree well with the respective RHF/6-31G(d) H-bond strengths of pentanamine···*p*-methylphenol in the protonated and neutral form of the amine, 15.69 and 8.77 kcal/mol.

(2) The PA Shift Due To the Protein Environment. The lower cycle of Figure 5 separates the effect of the Lys55···Tyr20 H-bond from the effect of the rest of the protein on the Lys55 PA, where the rest of the protein is considered to be EFP plus the Pro22-frozen LMOs. Again, for each system the energy of the buffer region has been subtracted from the total energy, to focus on the ab initio/protein interaction. As can be seen from Figure 5, this interaction is repulsive and further increases the PA by 8.70 kcal/mol, since the repulsive force increases from 8.58 to 17.28 kcal/mol upon deprotonation. The repulsion in the protonated case is not unexpected, since there are more positive residues in the immediate environment of Lys55 than negative. For example, the region within 14 Å, discussed above, has a net positive charge and results in a PA decrease (to 233.78 kcal/mol; cf. Table 2). However, the entire EFP is neutral, so the more distant negative residues attenuate the repulsion in the protonated state. In the neutral state this long-range attraction will be lost so that shorter range repulsive interactions dominate. Indeed, further calculations show that the ab initio/Pro22 interaction changes from slightly attractive (−0.82 kcal/mol)

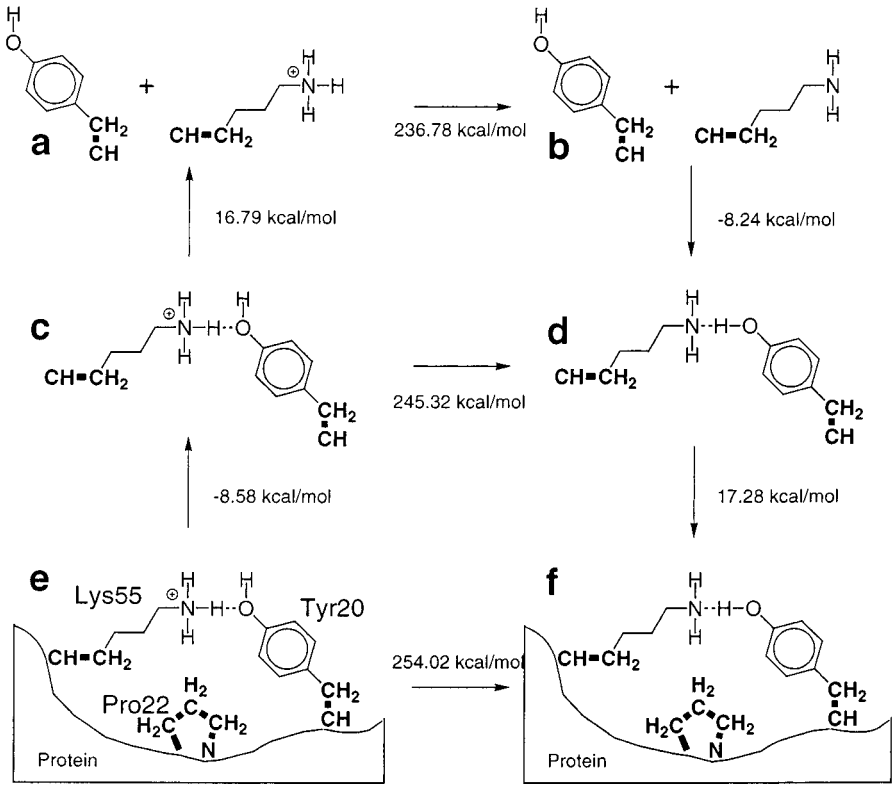


Figure 5. Thermodynamic cycle used to analyze the proteon affinity of Lys55 (e is a schematic depiction of Figure 3c). See Section III.D. for more detail.

to repulsive (3.58 kcal/mol), so this short-range interaction contributes 4.40 kcal/mol to the 8.70 kcal/mol PA shift. The main change in the ab initio geometry upon deprotonation is the position of the hydrogen and lone pairs on the Tyr20 oxygen. The increased repulsion in the deprotonated state is likely due to increased repulsion between these lone pairs and (1) a nearby $C^{\delta+}$-$H^{\delta-}$ bond in the Pro22 ring and (2) the negative charge on the neighboring glutamate residue (Glu19). Future computational studies will address these questions in more detail using site-directed mutagenesis.

E. The PA of Tyr20 in OMTKY3. Deprotonation of Tyr20, followed by a geometry optimization, results in a *spontaneous* proton transfer from Lys55 to Tyr20, to yield the same $-OH\cdot\cdot\cdot NH_2-$ hydrogen-bonded structure that resulted from Lys55 deprotonation. Thus, due to the intramolecular hydrogen bond between the two residues, their gas-phase PAs cannot be separated and the PA of Tyr20 equals that of Lys55. Interestingly, the experimentally measured solution pK_as of Lys55 and Tyr20 are also identical (both are 11.1). The pK_as are measured by monitoring the deprotonation event through the change in chemical shifts of the CH protons in $-CH_2NH_3^+$ and $(CH)_2$-COH vs pH. Our calculations show that the electronic structures, and therefore presumably the chemical shifts, of these four CH protons are coupled due to the intramolecular hydrogen bond. For example, the Mulliken charges of the CH protons change by up to 0.1 in $CH_2NH_3^+$ and 0.03 in $(CH)_2COH$ upon deprotonation of *either* group. It is therefore possible that the measured pK_as correspond to a single pK_a of the whole $-OH\cdot\cdot\cdot NH_3^+$ unit.

Our calculations are done in the gas phase using one of 50 NMR structures. The effect of solvent and protein dynamics could conceivably lead to different conclusions, so until we include both effects in our model, we cannot unequivocally verify our prediction computationally. However, the prediction can be tested experimentally by determining whether 1 or 2 equivalents of protons are released at pH ~11.1.

IV. Summary and Future Directions

This paper presents a divide-and-conquer method by which an accurate static and induced multipole representation of the eletrostatic potential of a protein can be generated using ab initio electronic structure theory. The method is used within the context of the effective fragment potential (EFP) method, a hybrid method in which only the active part of a molecular system is treated with ab initio quantum mechanics while the rest is replaced by an EFP (charges through octupoles at all atomic centers and bond midpoints and dipole polarizability tensors for each localized molecular orbital, LMO).

The proton affinities (PAs) of Lys55 and Tyr20 in the protein turkey ovomucoid third domain are calculated by treating the electronic and geometric structures of the Lys55 and Tyr20 side chains quantum mechanically [RHF/6-31G(d)]. The ab initio region is separated from the EFP by a buffer region comprised of the $C_\alpha-C_\beta$ bond of Lys55 and Tyr20 and the associated CH and core LMOs, as well part of the Pro22 ring. The Pro22 buffer is needed to describe its short-range interactions with Tyr20. The buffer is generated by an RHF/6-31G(d) calculation on a subset of the system (Figure 2).

The EFP describing the rest of the protein is generated by nine separate ab initio calculations on overlapping pieces of the protein truncated by methyl groups. These large calculations were made possible by the development of a new and computationally efficient method for calculating LMO polarizability tensors. Two different regions of overlap where used, depending on whether it occurred on the protein backbone or on a disulfide bridge. These regions of overlap are demonstrated to be sufficient by calculations on the protein environment within a 14 Å radius of Lys55 (see Figure 1b).

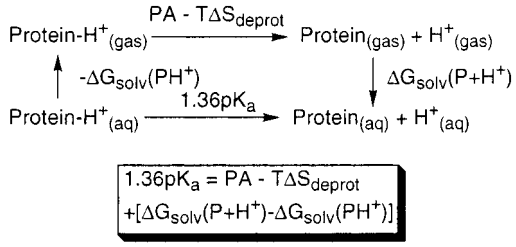


Figure 6. Thermodynamic cycle relating the pK_a to the gas-phase proton affinity (PA) via the solvation energies of the products and reactants. The value 1.36 corresponds to *RT* ln(10) at 298 K in kcal/mol.

On the basis of these and previous results, the Lys55 PA calculated using the EFP representation of the entire protein (see Figure 1b), 254.02 kcal/mol, should therefore be within about 1.0 kcal/mol of the fully ab initio RHF/6-31G(d) value. Dipoles and induced dipoles are necessary to obtain this accuracy. The PA value is used as a sensitive measure of the accuracy with which the molecular environment is modeled and is not meant to be a prediction of a gas-phase measurement, since electron correlation, nonelectronic energies and entropies, and protein dynamics are neglected.

The PA can be related to the pK_a via a thermodynamic cycle by calculating the change in solvation energy due to deprotonation, as shown in Figure 6. This can be accomplished by using discrete water EFPs to solvate the system or by a dielectric continuum solvation model (or a combination of the two). The former approach requires the implementation of computationally efficient short-range interactions, as well as the use of Monte Carlo[37] and molecular dynamics[38] techniques, which have been interfaced with the EFP method. Work on the EFP/continuum interface is also in progress.[39] We note that the use of the united atom implementation of the PCM method[40] has yielded accurate absolute pK_a values for small molecules using RHF/6-31+G(d).[41] Since the EFP method can reproduce the molecular electrostatic potential at this level, similar accuracy can be expected for proteins. We will address the solvation issue in future studies. Other calculations will study the effect of protein dynamics by averaging over other NMR structures as suggested by McCammon and Gilson.[40]

Even without the inclusion of these effects, this first-principles method can already provide new insight, since it predicts that the PA of Tyr20 will be identical to that of Lys55, consistent with experimental measurements of the solution-phase pK_a. The reason for the equal PAs is that the geometry optimization following deprotonation from either the NH_3^+ group on Lys55 or the OH group on Tyr20 yields the same structure, with a $-OH\cdots NH_2-$ hydrogen bond. Thus, deprotonation of Tyr20 is followed by a *spontaneous* proton transfer from Lys55, an event that cannot be predicted using a classical model.

Acknowledgment. This work was supported by the University of Iowa, the University of Iowa Biosciences Initiative Pilot Program, a Research Innovation Award from the Research Corporation, and a type G starter grant from the Petroleum Research Fund. J.H.J. gratefully acknowledges an Old Gold fellowship from the University of Iowa. The calculations were performed on IBM RS/6000 workstations obtained through a CRIF grant from the NSF (CHE-9974502) and on supercomputers at the Maui High Performance Computing Center and the National Center for Supercomputer Applications at Urbana−Champaign. The authors are indebted to Profs. Michael Gilson, Andrew Robertson, and Daniel Quinn, for careful readings of the manuscript, and to Dr. William Kearney, for drawing our attention to this interesting system.

References and Notes

(1) (a) Sharp, K. A.; Honig, B.. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301. (b) Warshel, A.; Åqvist, J. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 267. (c) Bashford, D. *Curr. Opin. Struct. Biol.* **1991**, *1*, 175. (d) Gilson, M. K. *Curr. Opin. Struct. Biol.* **1995**, *5*, 216. (e) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211. (f) Ullmann, G. M.; Knapp, E.-W. *Eur. Biophys. J.* **1999**, *28*, 533.

(2) See for example: Rappe, A. K.; Casewit, C. J. *Molecular Mechanics Across Chemistry;* University Science Books: Sausalito, CA, 1997.

(3) (a) Weiner, S. J.; Kollman, PA.; Case, D. A.; Singh, U. C.; Ghio. C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765. (b) Weiner, S. J.; Kollman, PA.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230.

(4) (a) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J. *J. Comput. Chem.* **1983**, *4*, 187. (b) Nilsson, L. Karplus, M. *J. Comput. Chem.* 7, 591. (c) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *2*, 3586.

(5) Van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992.

(6) Wiberg, K. B.; Rablen, P. R. *J. Comput. Chem.* **1993**, *14*, 1504−1518.

(7) Sokalski, W. A.; Maruszewski, K.; Harihan, P. C.; Kaufman, J. J. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1989**, *16*, 119.

(8) Stone, A. J.; Price, S. L. *J. Phys. Chem.* **1988**, *92*, 3325.

(9) (a) Faerman, C. H.; Price, S. L. *J. Am. Chem. Soc.* **1990**, *112*, 1990. (b) Price, S. L.; Richards, N. G. J. *J. Comput. Aided Mol. Des.* **1991**, *5*, 41.

(10) Hunter, C. A.; Singh, J.; Thornton, J. M. *J. Mol. Biol.* **1991**, *218*, 837.

(11) Mitchel, J. B. O.; Nandi, C. L.; Thornton, J. M.; Price, S. L.; Singh, J.; Snarey, M. *J. Chem. Soc., Faraday Trans.* **1993**, *89*, 2619.

(12) Dixon, R. W.; Kollman, P. A. *J. Comput. Chem.* **1997**, *18*, 1632.

(13) (a) Buckingham, A. D.; Fowler, P. W. *J. Chem. Phys.* **1983**, *79*, 6426. (b) Buckingham, A. D.; Fowler, P. W. *Can. J. Chem.* **1985**, *63*, 2018.

(14) See for example the following reviews: (a) Price, S. L. *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 2997. (b) Engkvist, O.; Åstrand, P.-O.; Karlstrom, G. *Chem. Rev.* **2000**, *100*, 4087.

(15) (a) Jensen, J. H.; Day, P. N.; Gordon, M. S.; Basch, H.; Cohen, D.; Garmer, D. R.; Kraus, M.; Stevens, W. J. In *Modeling the Hydrogen Bond*; ACS Symposium Series 569; Smith, D. A., Ed.; American Chemical Society: Washington, DC, 1994; Chapter 9. (b) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Kraus, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968. (c) Gordon, M. S.; Freitag, M.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293.

(16) Stone, A. J. *Chem. Phys. Lett.*, **1981**, 83, 233.

(17) Garmer, D. R.; Stevens, W. J. . *Phys. Chem.* **1989**, *93*, 8263.

(18) Bader, R. F. W. *Atom in Molecules. A Quantum Theory*; Clarendon Press: Oxford, 1994.

(19) Koch, U.; Stone, A. J *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 1701.

(20) Price, S. L.; Stone, A. J *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 1755.

(21) Matta, C. F.; Bader, R. F. W. *Proteins: Struct. Func, Genet.* **2000**, *40*, 310.

(22) We note that once transferable mMEPs are available for protein residues, they can easily be used in conjunction with the EFP method.

(23) Kairys, V.; Jensen, J. H. *J. Phys. Chem. A* **2000**, *104*, 6656.

(24) see for example (a) Zhang, Y.; Lee, T.-S.; Yang, W. J. Chem. Phys. **1999**, *110*, 46. (b) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *Chem. Phys. Lett.* **2000**, *321*, 113.

(25) Hoogstraten, C. G.; Choe, S.; Westler, W. M.; Markley, J. L. *Protein Sci.* **1995**, *4*, 2289.

(26) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.***1972**, *56*, 2257.

(27) King, H. F.; Stanton, R. E.; King, H.; Wyatt, R. E.; Parr R. G. *J. Chem. Phys.* **1967**, *47*, 1936.

(28) Stevens, W. J.; Fink, W. H. *Chem. Phys. Lett.* **1987**, *139*, 15.

(29) Bagus, P. S.; Hermann, K.; Bauschlicher, C. W., Jr. *J. Chem. Phys.* **1984**, *80*, 4378.

(30) (a) Boys, S. F. *Quantum Science of Atoms, Molecules and Solids*, Lowdin, P. O. Ed.; Academic Press: New York, 1966. (b) Edmiston, C.; Ruedenberg, K. *Rev. Mod. Phys.* **1963**, *35*, 457.

(31) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(32) Webb, S. P. Ph.D. Thesis, Iowa State University, 1998.

(33) Yamaguchi, Y.; Masamura, Y.; Goddard, J. D.; Schaefer, H. F. *A New Dimension, A New Dimension to Quantum Chemistry;* Oxford University Press: Oxford, 1994.

(34) Jeziorski, B.; Moszynski, R.; Ratkiewicz, A.; Rybak, S.; Szalewicz, K.; Williams, H. L. In *Methods and techniques in computational chemistry;* Clementi, E., Ed.; Cagliari, 1993; Metecc-94 Vol B, Chapter 3.

(35) Day, P. N.; Pachter, R.; Gordon, M. S.; Merrill, G. N. *J. Chem. Phys.* **2000**, *112*, 2063−2073.

(36) Gordon, M. S. and co-workers, work in progress.

(37) (a) Bandyopadhyay, P.; Gordon, M. S. *J. Chem. Phys.* **2000**, *113*, 1104. (b) Bandyopadhyay, P.; Gordon, M. S., work in progress.

(38) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(39) da Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem. A* **2000**, *103*, 11194.

(40) Antosiewicz, A.; McCammon, J. A.; Gilson, M. K. *Biochemistry* **1996**, *35*, 7819.