Generalized Energy-Based Fragmentation Approach for Computing the Ground-State Energies and Properties of Large Molecules

Wei Li, Shuhua Li,* and Yuansheng Jiang

School of Chemistry and Chemical Engineering, Institute of Theoretical and Computational Chemistry, Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Nanjing University, Nanjing, 210093, P. R. China

Received: November 21, 2006; In Final Form: January 4, 2007

We present a generalized energy-based fragmentation (GEBF) approach for approximately predicting the ground-state energies and molecular properties of large molecules, especially those charged and polar molecules. In this approach, the total energy (or properties) of a large molecule can be approximately obtained from energy (or properties) calculations on various small subsystems, each of which is constructed to contain a certain fragment and its local surroundings within a given distance. In the quantum chemistry calculation of a given subsystem, those distant atoms (outside this subsystem) are modeled as background point charges at the corresponding nuclear centers. This treatment allows long-range electrostatic interaction and polarization effects between distant fragments to be taken into account approximately, which are very important for polar and charged molecules. We also propose a new fragmentation scheme for constructing subsystems. Our test calculations at the Hartree−Fock and second-order Møller−Plesser perturbation theory levels demonstrate that the approach could yield satisfactory ground-state energies, the dipole moments, and static polarizabilities for polar and charged molecules such as water clusters and proteins.

1. Introduction

Linear scaling electronic structure algorithms have been developed for more than a decade, which are expected to be applicable to quantum mechanics calculations of large molecules. Many useful algorithms have been proposed for the Hartree−Fock (HF) and density functional theory (DFT) calculations,[1−11] post-HF calculations such as Møller−Plesser perturbation theory (MP) and coupled cluster theory (CC).[12−32] In addition, some lower order or linear scaling algorithms for energy gradient calculations[7,8,29−32] and molecular properties calculations[33−36] have also been proposed. For instance, the static response properties of molecules are usually computed through solving the coupled-perturbed self-consistent-field (CPSCF) method, whose computational cost scales as the fifth power of the number of basis functions.[37,38] Several approaches have been developed for reducing the scaling of the CPSCF calculation.[33,34] Among them, a promising approach is the linear scaling density matrix perturbation theory developed very recently.[35,36] However, such linear scaling algorithms have not been established as a practical tool for geometry optimizations and molecular properties calculations, because the crossover between these linear scaling algorithms and corresponding conventional methods occurs at quite large molecules.

The development of alternative approaches for performing quantum chemistry calculations for very large molecules, such as molecular fragmentation approaches, is also an active area in the recent years.[39−59] There are mainly two types of fragmentation approaches, one is the density matrix (DM) based approach[39−46] and the other is the energy-based approach.[50−59] In the DM-based approaches, the total energy of a target molecule is calculated from the assembled density matrix, which is constructed from the density matrices or molecular orbitals (MOs) of a series of subsystems. Such DM-based approaches have been shown to be capable for giving quite accurate ground-state energies and some molecular properties. But the use of this type of approach for performing geometry optimizations and frequency calculations of large molecules has not been achieved. However, in various energy-based approaches, the total energy of a target molecule can be directly derived from the energies of all subsystems constructed according to a fragmentation scheme. This type of approach is generally applicable at various theory levels and can be easily implemented for the geometry optimizations, calculations of vibrational frequencies and other properties. Depending on different fragmentation schemes, various energy-based approaches have been proposed.[50,51,53−56,59] These approaches have been shown to be quite successful for nonpolar or less charged molecules.[53−56,59] But test calculations also showed that they might give less accurate results for highly polar and charged molecules.[53] To treat the charged molecules, Jiang and Ma et al. improved our previous work in their electrostatic field-adapted molecular fractionation with conjugated caps (EFA-MFCC) approach by adding point charges on the charge centers of those charged groups (outside a given subsystem).[57] Then, these point charges are incorporated in the quantum chemistry calculations of various subsystems. This modified approach was found to give improved results for some charged biomolecules. However, their approach is not applicable for highly polar molecules such as α-helix polypeptides and water clusters, and also for those charged molecules with delocalized charges in some groups (in which charges may spread over several atoms rather than a specific atom). To treat large cluster systems, Sakai and Morita recently proposed the integrated multicenter molecular orbitals (IMiCMO) method,[50,51] which can also be considered as a energy-based fragmentation approach. In this approach the total electronic energy of a larger cluster system is also expressed as

* Corresponding author. E-mail: shuhua@nju.edu.cn.

10.1021/jp067721q CCC: $37.00 © 2007 American Chemical Society
Published on Web 02/23/2007

the summation of electronic energies of some small clusters. For a given small cluster, those distant molecules (outside this cluster) are also modeled by point charges, which are incorporated in the quantum calculation of this cluster. The method has been shown to give satisfactory ground-state energies and vibrational frequencies for neutral and charged water clusters.

In the present work, we suggest a generalized energy-based fragmentation (GEBF) approach for treating general large molecules (macromolecules or large cluster systems), which may be charged or highly polar. In this approach, those distant atoms of the target molecule, which are not explicitly included in a given subsystem, are represented as point charges. These point charges are then incorporated in the quantum calculations of subsystems. In this way, not only the electrostatic interaction between distant fragments but also the polarization of a given fragment by distant atoms of the target molecule are approximately taken into account. The way we use here for treating the interaction between distant fragments is similar to that in the field-adapted adjustable density matrix assembler (FA-ADMA) approach[41] and the IMiCMO method.[50,51] The main purpose of this paper is to develop a very simple but effective way to compute the total energy of a general large system (this molecule may be a macromolecule or a weakly bonded cluster system). It should be mentioned that the treatment of the present approach for large cluster systems is somewhat similar to the IMiCMO method.[50,51] However, an advantage of the present approach over the IMiCMO method is that the present approach is also applicable for treating general large macromolecules, in which different parts are bonded covalently. In comparison with the previous EFA-MFCC approach,[57] the present approach represents a further improvement because the electrostatic interaction between polar groups is now taken into account. In the present work, we also propose a new fragmentation scheme, which is different from those in all previous works.[50,51,53−56,59] In addition, the GEBF approach is shown to be directly applicable for approximately computing the dipole moment and static polarizability of large molecules.

This paper is organized as follows. In section 2, the theory and computational details of the GEBF approach are introduced. In section 3, we illustrate the accuracy of this approach for some typical macromolecules and larger clusters, and make comparisons with conventional calculations. Finally, a brief summary is given in section 4.

2. Methodology

In previous energy-based fragmentation approaches,[51,53−57,59] the total ground-state energy of the target system can be expressed as

$$E_{\mathrm{Tot}} = \sum_{m}^{M} C_m E_m \tag{1}$$

where E_m and C_m is the ground-state energy and coefficient of the mth subsystem and M is the number of constructed subsystems. The basic procedures of these energy-based fragmentation approaches include (1) divide a target molecule into small fragments, (2) construct subsystems from small fragments according to some rules, (3) perform conventional quantum chemistry calculations on these subsystems, and (4) obtain the total energy of the target molecule using eq 1. It should be pointed out that in constructing subsystems, only the local surroundings of a given fragment within a given distance are explicitly considered. In other words, the interaction between a given fragment and those distant atoms of the target molecule is completely ignored. For nonpolar macromolecules, the neglect of this interaction has little influence on the calculated total energy, as shown in previous test calculations.[53−56,59] However, this approximation will lead to serious errors for charged and highly polar macromolecules such as proteins or water clusters, in which the interaction between two distant fragments may be electrostatic in origin and thus long-range.

Before we introduce a GEBF approach for treating general molecules, we will propose a new way of constructing subsystems for a given target system. The overall procedure follows: (1) Divide a target molecule into fragments of comparable size. (2) Cap each fragment (called central fragment) with its neighboring fragments to construct closed-shell subsystems. For convenience, these subsystems are called primitive subsystems (the total number is equal to the total number of fragments). If the neighboring fragment is bonded to the rest of the system by a covalent single bond X−Y (X belongs to this fragment), we need to replace the X−Y bond with the X−H bond. The position of this added hydrogen atom could be determined as discussed elsewhere.[53] Here, for a given fragment, its neighboring fragments are those fragments that are linked to this fragment through a covalent bond or a hydrogen bond.[60] In some cases such as molecular clusters, each molecule is chosen as a fragment, and its neighboring fragments are defined to be those molecules within a given distance threshold ξ_1. Obviously, a larger ξ_1 value will generate larger subsystems and thus lead to more accurate results. To take a compromise between accuracy and computational cost, we usually choose a moderate ξ_1 value. For example, for water clusters, we set ξ_1 to be 3 Å. If a subsystem is totally embedded in another larger subsystem, then this subsystem is eliminated from the list. Suppose that the maximum number of fragments in all primitive subsystems is $n_{\max}$ (there may exist many such subsystems). (3) Construct m-fragment ($m = n_{\max} - 1$) derivative subsystems. If the net number of a specific m-fragment interaction term occurring in all primitive subsystems is k, we then construct a derivative subsystem containing these m fragments, with its coefficient being chosen as $(1 - k)$. Thus, the net number of each specific m-fragment interaction term in all subsystems (both primitive and derivative) is 1. Similarly, for all other m-fragment interaction terms, we can construct their corresponding derivative subsystems. (4) For m-fragment ($m = n_{\max} - 2, ..., 2, 1$) interaction terms, we repeat the process described above to construct corresponding derivative subsystems. (5) Construct additional two-fragment subsystems. Although in primitive and derivative subsystems, most important two-fragment interactions have been taken into account, many non-negligible two-fragment interactions are not included explicitly. In general, if the distance between two fragments is larger than a threshold ξ_2 (ξ_2 is set to be 8 Å in our calculations), the interaction between these two fragments can be approximately modeled by the Coulomb interaction between point charges of these two fragments. However, when the fragment−fragment distance is less than ξ_2, it is better to compute their interaction by subtracting the energy of the corresponding dimer from the energies of two separated monomers (formed by adding hydrogen atoms to fragments if necessary). As a result, some two-fragment subsystems (the distance between two fragments is less than ξ_2) are added to get more accurate results. To illustrate this new fragmentation scheme, we take mixed-α/3_{10}-acetyl(gly)$_6$NH$_2$ as an example. This molecule can be divided into six fragments, as shown in Figure 1. According to the connectivity of various fragments, we can form six subsystems in the first step, which can be denoted as (1234), (1235), (12346), (13456), (2456), and

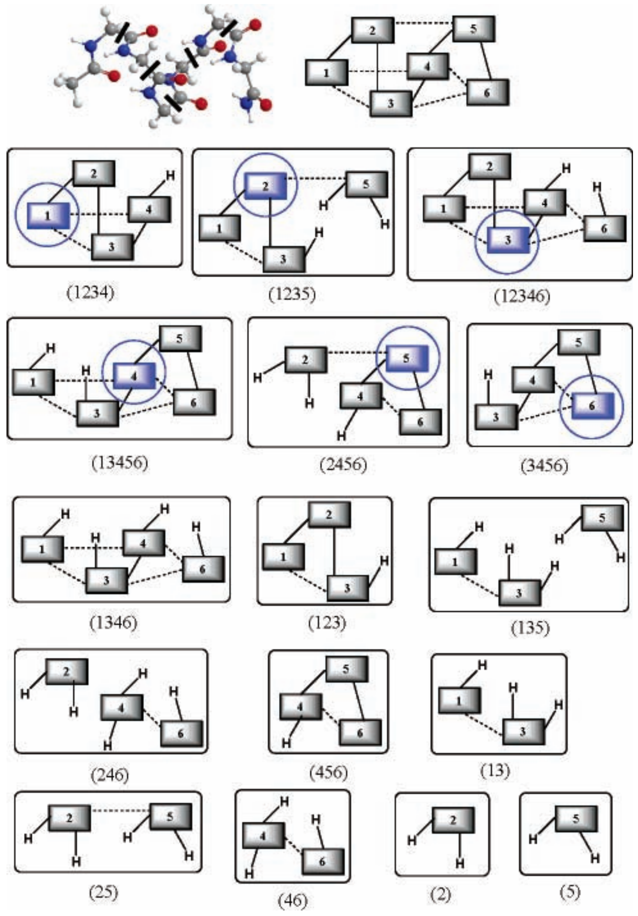

Figure 1. Fragmentation scheme and constructed subsystems for mixed-α/3_{10}-acetyl(gly)$_6$NH$_2$. In the schematic diagram, solid lines between fragments stand for covalent bonds, and dashed lines stand for hydrogen bonds. In the primitive subsystems, the central fragment is denoted with a cycle. For all subsystems, hydrogen atoms added for valence saturation are explicitly displayed.

TABLE 1: All Subsystems Constructed for Mixed-α/3_{10}-acetyl(gly)$_6$NH$_2$[a]

subsystem	coefficient	components	subsystem	coefficient	components
1	1	(1235)	8	−1	(246)
2	1	(12346)	9	−1	(456)
3	1	(13456)	10	1	(13)
4	1	(2456)	11	−1	(25)
5	−1	(1346)	12	1	(46)
6	−1	(123)	13	1	(2)
7	−1	(135)	14	1	(5)

[a] Components of a given subsystem stand for all fragments involved (hydrogen atoms may add).

(3456) (the labels of fragments are combined to specify a subsystem, and hydrogen atoms added for saturating dangling bonds are implicitly assumed to belong to the corresponding fragment; see Figure 1). Clearly, two subsystems (1234) and (3456) can be eliminated because they are included in two larger subsystems (12346) and (13456). Then the retained four primitive subsystems are (1235), (12346), (13456), and (2456), in which the maximum number of fragments in these subsystems is five, as seen in Table 1. Second, by checking four-fragment interaction terms in four subsystems, we construct corresponding derivative subsystems. For example, because the 1-2-3-4 term occurs only once in all four primitive subsystems, it is not necessary to construct a derivative subsystem (1234). But for the 1-3-4-6 term, we can see that it occurs twice in two five-fragment subsystems, so we have to build a derivative subsystem (1346) and set its coefficient to be (−1). In a similar way, we can build three-fragment, two-fragment, and one-fragment derivative subsystems. Totally, for this model system 14 subsystems can be obtained (as shown in Figure 1) and their coefficients are listed in Table 1. By checking this table, one can see that the net number of each intrafragment term or each specific n-fragment term ($n = 2, 3, ..., n_{\max}$) is 1.

It should be pointed out that in the construction of all subsystems, most important three- or four-fragment interaction terms are included, but not all of them are explicitly treated. The neglect of some three- or four-fragment interaction terms may give rise to some errors in the total energy, but our calculations show that the addition of corresponding three- or four-fragment subsystems only brings insignificant improvement. For a target system, the construction of all subsystems can be initiated by manually assigning all fragments, and then the subsequent steps can be completed automatically by running a program. Within this fragmentation scheme, the renumbering of all fragments in the target molecule will lead to identical subsystems. Thus, the same total energy will be obtained, regardless of the labels of various fragments.

An essential requirement for an energy-based fragmentation approach is that the net number of added hydrogen atoms ("link" atoms) must be zero. Let us analyze whether the above-described fragmentation procedure meets this requirement. In general, an arbitrary fragment i is linked to fragment j through a covalent bond (otherwise no link hydrogen atoms are required). According to our construction rules described above, the net number of subsystems including both fragments i and j, λ_{ij}, must be equal to 1. In other subsystems including i but without j, an extra hydrogen atom must be bonded to i. Assume that the net number of such subsystems is denoted as $\lambda_{i(j)}$. In addition, we know that the net number of subsystems including i, λ_i, must be equal to 1. Because $\lambda_i = \lambda_{ij} + \lambda_{i(j)}$, we then deduce $\lambda_{i(j)} = 0$. This result clearly shows that extra hydrogen atoms bonded to fragment i occurring in all subsystems are cancelled.

As shown previously, the total energy of a large molecule can be approximately obtained from eq 1 through the calculations of all subsystems if the target molecule is nonpolar and neutral. However, for highly polar or charged molecules, we have suggested earlier that those atoms not included in a given subsystem should be approximately represented as point charges. Then these point charges located at the corresponding nuclear centers are incorporated into the quantum calculations of all constructed subsystems, which can be done with the Gaussian 03 program[61] (thus no additional programming effort is needed). So within the GEBF approach, each subsystem is computed in the presence of point charges on those distant atoms (outside this subsystem). It should be pointed out that charges on those junction atoms (atoms replaced by extra hydrogen atoms) are fully taken into account in calculations on subsystems (because they are far apart from the central fragment and their influence can be neglected). Now we discuss how to obtain these partial charges used for quantum chemical calculations on subsystems. Although various schemes have been proposed to compute atomic charges,[62−67] we have found that the use of natural charges from natural population analysis (NPA)[66,67] leads to satisfactory results, as shown later in the next section. An iterative way for obtaining these natural charges can be described below. First, we perform a standard HF (or DFT) calculation for each primitive subsystem without including background charges. Then, only natural charges on the central fragment are extracted from the corresponding NPA calculation. Calculations on all primitive subsystems (its number is equal to the number of fragments) will produce an initial guess for partial charges on all atoms. Next, these partial charges are incorporated in the HF (or DFT) calculation of each primitive subsystem (to replace those distant atoms), and partial charges on all atoms of the target molecule can be recalculated as described above. In principle, we could repeat the above step until partial charges on all atoms are converged. In fact, we have found that one iteration is usually enough to produce nearly convergent natural charges. Therefore, the natural charges from the first iteration are employed for the energy and properties calculations of the target molecule in the GEBF approach.

The most important question for the GEBF approach is how to obtain the total energy expression of a target molecule from the energies of its various subsystems. Fortunately, we find that there exists a very simple relationship between the total energy of the target molecule and those of all subsystems constructed within the GEBF approach,

$$E_{\mathrm{Tot}} = \sum_m^M C_m \tilde{E}_m - (\sum_m^M C_m - 1)\sum_{\mathrm{A}} \sum_{\mathrm{B>A}} \frac{Q_{\mathrm{A}} Q_{\mathrm{B}}}{R_{\mathrm{AB}}} \qquad (2)$$

where $\tilde{E}_m$ is the total energy of the mth subsystem including the self-energy of charges (on those distant atoms), C_m is the coefficient of the mth subsystem, and Q_{A} is the partial charge on atom A. We will give a simple analysis for the derivation of this equation. Suppose there are M subsystems constructed for a target molecule. For an arbitrary fragment i, if all atoms of this fragment are treated as background point charges, we denote this fragment as a dummy fragment i'. Because the total net number of fragment i and its dummy counterpart occurring in all subsystems is $\lambda_i + \lambda_{i'} = \sum_m^M C_m$, the net number of the dummy fragment i' is $\lambda_{i'} = \sum_m^M C_m - 1$ ($\lambda_i = 1$ by construction). If we simply calculate the total energy of the target system as the sum of energies of all subsystems, the self-energy of the charges on each fragment will be counted for $(\sum_m^M C_m - 1)$ times and thus should be removed. On the other hand, in all subsystems the two-fragment interaction term between i and j may exist in four different forms, such as $i-j$, $i-j'$, $i'-j$, and $i'-j'$, with their net number denoted as λ_{ij}, $\lambda_{ij'}$, $\lambda_{i'j}$, and $\lambda_{i'j'}$, respectively. For instance, the $i-j'$ term denotes that in a given subsystem the fragment i is explicitly treated, and the fragment j is modeled as point charges. By construction, we have $\lambda_{ij} + \lambda_{ij'} + \lambda_{i'j} + \lambda_{i'j'} = \sum_m^M C_m$. If the distance between fragments i and j is less than ξ_2, we have $\lambda_{ij} = 1$ and $\lambda_{ij'} = \lambda_{i'j} = 0$ ($\lambda_{ij'} = \lambda_i - \lambda_{ij}$). As a result, $\lambda_{i'j'} = \sum_m^M C_m - 1$. Thus the $i'-j'$ interaction term has been counted for $(\sum_m^M C_m - 1)$ times and should be removed (because the $i-j$ interaction term has been incorporated). If the distance between i and j is larger than ξ_2, $\lambda_{ij} = 0$ and $\lambda_{ij'} = \lambda_{i'j} = 1$. Then, $\lambda_{i'j'} = \sum_m^M C_m - 2$. It is well-known that when two fragments are separated from a distance ξ_2 (8 Å in the present work), the two-fragment interaction terms approximately satisfy $E_{ij} \approx E_{ij'} \approx E_{i'j} \approx E_{i'j'}$. Because the interaction terms $i-j'$, $i'-j$, and $i'-j'$ are all included in the energies of all subsystems, we should remove the $i'-j'$ term for $(\sum_m^M C_m - 1)$ times to ensure that the interaction between i and j is only counted once. In summary, a combination of all the intrafragment and two-fragment electrostatic interaction terms required to be deleted is just equal to the second term of eq 2, which completes the proof.

TABLE 2: GEBF-Energy Deviations with Respect to the Conventional Energies for $(H_3O^+)_5(HO^-)_5(H_2O)_{22}$ by Using Different Types of Atomic Charges[a]

	energy deviation (mH)	
charges	HF	MP2
no charges	−61.72	−130.03
Mulliken	−5.44	−8.87
ESP	−7.06	−8.36
Natural	−0.49	0.34

[a] The conventional HF and MP2 energies are −2432.28054 and −2439.75819 au, respectively.

One may wonder whether eq 2 can also be extended for calculations on some molecular properties of large molecules. By adding an external electric field F_i ($i = x, y, z$) to a molecule, one can express the dipole moment and static polarizability as the first and second derivatives of the total energy with respect to the electric field,[68]

$$\mu_i = -\frac{\partial E}{\partial F_i} \qquad (i = x,y,z) \qquad (3)$$

$$\alpha_{ij} = -\frac{\partial^2 E}{\partial F_i \partial F_j} \qquad (i, j = x,y,z) \qquad (4)$$

Then, within the GEBF approach, the dipole moment, and static polarizability of the target molecule can be approximately calculated a

$$\Omega_{\mathrm{Tot}} = \sum_m^M C_m \tilde{\Omega}_m \qquad (\Omega = \mu_i, \alpha_{ij}, ...) \qquad (5)$$

where $\tilde{\Omega}_m$ is the corresponding property of the mth subsystem (with distant atoms represented as point charges). In a recent work by Zhang et al.,[69] a similar formula for estimating the dipole moment was implemented in their molecular fractionation with conjugated caps (MFCC) approach, but the subsystems were not in the presence of background charges. Equation 5 could also be applied to calculate other molecular properties, which will be explored in our future work. In addition, we found that within the current implementation of the GEBF approach, it is difficult to give quantitative predictions on static hyperpolarizability, which can be derived from the third energy derivatives. The reasons are not clear yet.

3. Results and Discussion

In this section, we will illustrate the effectiveness and applicability of the present GEBF approach for various highly polar and charged systems, including small proteins and water clusters. All the conventional quantum chemistry calculations for the target systems and their subsystems are performed with the GAUSSIAN 03 package.[61] For MP2 calculations, all electrons are correlated, and 6d Cartesian functions are used for polarized basis sets. The present GEBF approach has been implemented in the LSQC quantum chemistry package.[70]

First, we want to investigate the influence of different types of point charges on the total energy of the target system. We have taken $(H_3O^+)_5(HO^-)_5(H_2O)_{22}$, a heavily charged water cluster, as a test molecule. Table 2 shows the energy deviations obtained from the GEBF approach with respect to the corresponding conventional values at the HF and MP2 levels (6-311G* basis set is used). Here the point charges are calculated from the conventional HF calculation on the whole system. One can see from Table 2 that if point charges are not considered, both the HF and MP2 energies given by the GEBF approach deviate from the conventional values by −61.72 and −130.03 milliHartree (mH), respectively. When the natural charges[66,67] are used as background charges, the GEBF-HF and GEBF-MP2

TABLE 3: GEBF-Energy Deviations with Respect to the Conventional Energies for $(H_3O^+)_5(HO^-)_5(H_2O)_{22}$ by Using Natural Charges from Different Iteration Steps[a]

iteration step	energy deviation (mH)		mean deviation of NPA charges
	HF	MP2	
0	−1.11	−1.33	0.0160
1	−0.35	0.49	0.0031
2	−0.36	0.47	0.0029
3	−0.36	0.47	0.0029

[a] In the last column, the mean deviation between the NPA charges from the GEBF approach and those from the conventional calculation on the whole system is listed.

energies are different from the conventional values only by less than 0.5 mH, significantly better than those with Mulliken charges[62,63] and electrostatic potential (ESP) charges.[64,65] For other systems, we also obtained similar trends. Thus the natural charges are adopted in the present work. Next, for this molecule, we have investigated how many iterations are required to obtain nearly convergent natural charges, according to the procedure described earlier in the preceding section. The results are listed in Table 3. One can see that the GEBF-HF and GEBF-MP2 energies are almost convergent when the natural charges are obtained from the first iteration. In addition, a comparison of the natural charges from the GEBF approach with those from the conventional calculation on the whole system (listed in the last column of TABLE 3) shows that the mean deviation between them is only 0.003 after the first iteration. This result indicates that the natural charges after the first iteration are quite close to those exact natural charges from the whole system, and thus can be adopted as background charges in subsystem calculations. It should be mentioned that the sum of charges on all atoms of the whole system obtained in this way is not a precise integer value. But the calculated total charges usually deviate from the corresponding integer by a small value. For example, the total natural charge is only −0.034 for $(H_3O^+)_5$-$(HO^-)_5(H_2O)_{22}$.

The systems we have chosen for validating the applicability of the GEBF approach at the HF level are a series of water clusters and proteins. They include ice-like water clusters $(H_2O)_n$ (n = 32, 48, 64, 80, 96), and $(H_3O^+)_5(HO^-)_5(H_2O)_{22}$, three conformers of acetyl(ala)$_n$NH$_2$ peptide (n = 10, 18), and eleven proteins from the protein data bank (PDB).[71] For some medium-sized molecules, we have used the 6-311G** or 6-311G* basis set, and for relatively larger proteins we have used the 6-31G basis set for saving the computational time. The geometries of all studied molecules are taken from other literatures or constructed by us. For these systems, the fragmentation details are described below. For water clusters, each water molecule is selected as a fragment. And for proteins, we cut the C−C bond between α-carbon and the carbonyl group in the central residues, the S−S bond between two residues, and the C−C bond between $\beta-$ and $\gamma-$carbons in five residues with large side chains (Arg, Lys, Phe, Trp, and Tyr).

For selected water clusters and proteins, we have shown their total energies, dipole moments, and static polarizabilities calculated by the conventional HF and GEBF-HF calculations in Table 4 for comparison. For all studied systems, the differences between the GEBF-HF energies and the conventional HF values are less than 14 mH. For ice-like water clusters, the GEBF approach can reproduce the conventional HF energies quite well, with errors less than 4.0 mH. For the charged water cluster $(H_3O^+)_5(HO^-)_5(H_2O)_{22}$, the GEBF-HF energy deviates from the conventional HF value only by −0.4 mH. But if background charges are not included, the corresponding energy

TABLE 4: Ground-State Energies, Dipole Moment, and Static Polarizability Calculated from the Conventional and GEBF Calculations

molecule	basis set and basis functions	total energy		dipole moment (Debye)		polarizability (Bohr^3)		basis functions of the largest subsystem
		conventional (au)	GEBF (mH)[a]	conventional	GEBF	conventional	GEBF	
ice-like $(H_2O)_{32}$	6-311G** (960)	−2433.660 98	−0.06 (27.77)	19.26	19.23	215.54	216.94	210
ice-like $(H_2O)_{48}$	6-311G** (1440)	−3650.537 68	−1.32 (47.11)	12.05	12.02	324.93	327.91	210
ice-like $(H_2O)_{64}$	6-311G** (1920)	−4867.448 88	−2.01 (71.03)	5.75	5.77	435.46	439.52	210
ice-like $(H_2O)_{80}$	6-311G** (2400)	−6084.515 33	−2.49 (122.13)	9.74	9.59	551.87	559.26	210
ice-like $(H_2O)_{96}$	6-311G** (2880)	−7301.522 60	−3.52 (192.21)	3.93	3.79	665.45	673.82	210
$(H_3O^+)_5(HO^-)_5(H_2O)_{22}$	6-311G* (768)	−2432.267 91	−0.40 (−61.88)	46.73	46.70	219.25	219.58	270
β-strand acetyl(ala)$_{18}$NH$_2$	6-311G* (1977)	−4634.234 92	1.07 (1.25)	40.68	38.55	756.57	743.18	378
3_{10}-helix acetyl(ala)$_{18}$NH$_2$	6-311G* (1977)	−4633.988 84	−0.65 (37.03)	90.42	91.46	724.39	712.30	588
α-helix acetyl(ala)$_{18}$NH$_2$	6-311G* (1977)	−4634.295 77	−1.45 (47.76)	93.57	94.78	742.03	729.11	600
α-conotoxin pnib (1AKG, 1^+2^-)[b]	6-311G* (2324)	−6832.988 11	−5.49	150.50	150.28	886.57	903.04	881
α-conotoxin mii (1M2C, 3^+1^-)	6-311G* (2432)	−7108.158 63	−5.23	251.47	250.77	911.58	920.95	1033
α-conotoxin pni1 (1PEN, 1^+2^-)	6-311G* (2291)	−6808.852 81	−2.53	178.44	178.15	865.50	877.63	1014
epidermal growth factor subdomain (1FGD, 1^+7^-)	6-311G* (2896)	−7610.993 01	−2.28	144.78	144.71	1081.53	1088.66	1045
crambin (1CNR, 3^+3^-)	6-31G (3597)	−17994.671 09	−13.38	37.08	36.66	2211.96	2236.43	675
μ-conotoxin giiib (1GIB, 11^+5^-)	6-31G (1749)	−10055.552 46	−4.06	484.77	485.74	1161.99	1176.09	557
Bovine lactoferricin (1LFC, 9^+1^-)	6-31G (2447)	−11335.149 97	−8.44	117.66	117.14	1627.85	1633.50	498
ω-conotoxin mviia (1OMG, 7^+2^-)	6-31G (1973)	−11118.098 34	1.65	128.55	129.54	1295.15	1313.21	503
tertiapin (1TER, 7^+1^-)	6-31G (1894)	−9693.830 82	−4.10	137.87	136.99	1285.34	1307.78	643
vacuolar targeting peptide (1VTP, 4^+7^-)	6-31G (2206)	−10010.378 48	−4.34	230.94	232.03	1424.82	1416.76	668
trypsin inhibitor II (2ETI, 4^+3^-)	6-31G (2164)	−11988.640 78	−0.15	61.93	61.99	1437.23	1460.90	560

[a] The relative energies with respect to corresponding conventional energies, and those without point charges included in parentheses. [b] The number of positive and negative charge centers included in parentheses for eleven proteins.

TABLE 5: HF and MP2 Energies Calculated from the Conventional and GEBF Approaches with the 6-311G* Basis Set

	basis functions		conventional (au)		GEBF (mH)[a]	
molecule	whole system	largest subsystem	HF	MP2	HF	MP2
β-strand acetyl(ala)$_{10}$NH$_2$	1191	396	−2667.042 11	−2676.607 67	0.49	1.41
3$_{10}$-helix acetyl(ala)$_{10}$NH$_2$	1191	616	−2667.064 28	−2676.672 67	−0.32	0.33
α-helix acetyl(ala)$_{10}$NH$_2$	1191	628	−2667.045 58	−2676.659 72	0.27	1.01
$(H_2O)_{32}$	800	300	−2433.466 43	−2440.882 29	0.02	0.88
$(H_2O)_{40}$	1000	300	−3041.730 36	−3051.006 71	0.57	2.23
$(H_2O)_{48}$	1200	300	−3650.122 28	−3661.259 38	0.23	2.24
$(H_3O^+)_5(HO^-)_5(H_2O)_{22}$	800	281	−2432.280 54	−2439.758 19	−0.35	0.49

[a] The relative energies with respect to the conventional HF or MP2 energies.

difference for this molecule will climb up to about 61.9 mH. For three conformers of acetyl(ala)$_{18}$NH$_2$ peptide, the maximum deviation of the GEBF-HF energy from the conventional HF energy is 1.45 mH for the α-helix structure, but 47.76 mH if the background charges are not included. But for the β-strand structure, the GEBF-HF energies differ from the conventional values by about 1.1 mH either with or without background charges. Although both conformers are not charged, it is well-known that the α-helix structure is highly polar but the β-strand structure is a weakly polar molecule. The performance of the GEBF approach for these two conformers shows that the GEBF approach can give satisfactory descriptions for the polarization effect introduced by polar groups. For eleven proteins with at most 16 charge centers, their GEBF-HF energies deviate from conventional HF energies by at most −13.38 mH for crambin (PDB id: 1CNR), which is the largest molecule in our calculations. For crambin, there are 642 atoms with 3597 basis functions (6-31G), and in the GEBF calculations the largest subsystem includes only 124 atoms with 675 basis functions. On the other hand, one can see from Table 4 that for all systems under study the GEBF approach can give satisfactory results on the dipole moment and polarizability properties. For the dipole moment, the largest and mean deviations with respect to the conventional values are 5.2% and 0.9%, respectively. And for the static polarizability, the largest and mean deviations are 1.9% and 1.2%, respectively. Therefore, all the results at the HF level show that the present GEBF approach can predict the ground-state energy, the dipole moment and polarizability fairly well even for highly polar and charged molecules.

To illustrate the performance of the GEBF approach for post-HF calculations, seven molecules are calculated with the conventional and GEBF-MP2 approaches, with their results listed in Table 5. The 6-311G* basis set is used for all MP2 calculations. From TABLE 5, one can see that for all seven molecules, GEBF-MP2 energies deviate from the conventional MP2 energies by at most 2.24 mH. For water clusters $(H_2O)_n$ ($n = 32, 40, 48$), the largest deviation is 2.24 mH for $(H_2O)_{48}$, in which the number of basis functions is 1200 for the whole system but only 300 for the largest subsystem. For the highly charged $(H_3O^+)_5(HO^-)_5(H_2O)_{22}$, the GEBF-MP2 calculation also gives remarkably accurate results. Thus, the GEBF approach at the MP2 level is as successful as that at the HF level.

Although the computational cost of the GEBF approach at the HF or post-HF level is expected to increase linearly with the system size (because the number of subsystems grows linearly with the system size), it is interesting to see where the crossover between the conventional and GEBF calculations occurs. For some systems shown in Table 4, we have found that the computation cost of the GEBF-HF calculations is even larger than that required by the conventional HF calculations. This is because in these systems subsystems are only slightly smaller than the whole system. If subsystems are noticeably smaller than the target molecule, the crossover between conventional HF and GEBF-HF calculations will occur at medium-sized molecules. As shown in Figure 2 for ice-like water clusters, one can see that the crossover point appears at $(H_2O)_{48}$ (all the calculations are carried out on 3.0 GHz Pentium 4 workstations). On the other hand, at the MP2 level the computational advantage of the GEBF approach over the conventional approach is much more obvious. For instance, for the water cluster $(H_2O)_{32}$ with the 6-311G* basis set, the GEBF−MP2 calculation is already 7 times faster than the conventional MP2 calculation. In addition, for large systems conventional MP2 calculations are not feasible also due to the lack of sufficient disk and memory space. However, within the GEBF approach, if all subsystems can be treated at a theoretical level, the target molecule can then be treated at this level no matter how large it is. Furthermore, because the calculation of each subsystem is independent of other subsystems, highly efficient parallel computations can be achieved within the GEBF approach, which are hardly possible within the conventional HF or post-HF methods. Therefore, with a message-passing interface (MPI)[72] parallel technique, one can apply the GEBF approach to perform ab initio quality calculations on systems with thousands of atoms.

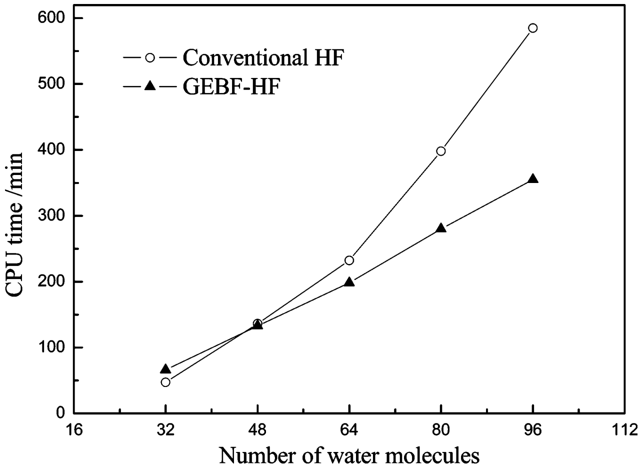

Figure 2. CPU times for conventional HF and GEBF-HF calculations of ice-like water clusters at the 6-311G** basis set.

4. Conclusion

In this work, we have presented a GEBF approach for approximately computing the ground-state energies and some response properties of general large molecules, especially those highly polar and charged molecules. In this approach, each subsystem is placed in the background charges generated by those distant atoms (outside this subsystem). This treatment allows long-range electrostatic interactions and polarization effects to be taken into account approximately, which are very important for highly polar and charged molecules. In addition, we have introduced a new scheme for fractionizing a general molecule. Our test calculations show that the present approach can reproduce the conventional HF and MP2 energies within a few milliHartrees for selected highly polar and charged molecules. Furthermore, some properties, such as the dipole moment and static polarizability, can also be reasonably predicted within the GEBF approach.

It should be mentioned that the GEBF approach has its inherent limitations. For example, it is difficult to extend this approach to molecules with highly delocalized electrons, such as two-dimensional conjugated systems and radicals. We also find that the GEBF approach is less successful for estimating some molecular properties, which are dependent on the energy derivatives of third or higher order, such as static hyperpolarizability. Despite these limitations, the GEBF approach, if appropriately employed, is expected to become a promising theoretical tool for performing ab initio quantum chemistry calculations for very large molecules in the near future.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (Grant Nos. 20373022, 20625309, and 20433020), the National Basic Research Program of China (Grant No. 2004CB719901), the Chinese Ministry of Education (Grant No. NCET-04-0450), and Fok Ying Tong Education Foundation (Grant No. 91014).

Supporting Information Available: The Cartesian coordinates of all systems and schematic fragmentation schemes. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Strout, D. L.; Scuseria, G. E. *J. Chem. Phys.* **1995**, *102*, 8448.
(2) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51.
(3) White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1994**, *101*, 6593.
(4) Scuseria, G. E. *J. Phys. Chem. A* **1999**, *103*, *4782.*
(5) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1996**, *105*, 2726.
(6) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 1663.
(7) Burant, J. C.; Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *248*, 43.
(8) Kudin, K. N.; Scuseria, G. E. *Phys. Rev. B* **2000**, *61*, 16440.
(9) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *257*, 213.
(10) Millam, J. M.; Scuseria, G. E. *J. Chem. Phys.* **1997**, *106*, 5569.
(11) Li, X.; Millam, J. M.; Scuseria, G. E.; Frisch, M. J.; Schlegel, H. B. *J. Chem. Phys.* **2003**, *119*, 7651.
(12) Lecszsynski, J. *Computational Chemistry: Review of Current Trends*; World Scientific Publishing: Singapore, 2002; Vol. 7.
(13) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151.
(14) Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213.
(15) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286.
(16) Schütz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691.
(17) Schütz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661.
(18) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149.
(19) Ayala, P. Y.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 3660.
(20) Scuseria, G. E.; Ayala, P. Y. *J. Chem. Phys.* **1999**, *111*, 8330.
(21) Ayala, P. Y.; Kudin, K. N.; Scuseria, G. E. *J. Chem. Phys.* **2001**, *115*, 9698.
(22) Almlöf, J. *Chem. Phys. Lett.* **1991**, *181*, 319.
(23) Head-Gordon, M.; Maslen, P. E.; White, C. A. *J. Chem. Phys.* **1998**, *108*, 616.
(24) Nakao, Y.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 6375.
(25) Christiansen, O.; Manninen, P.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2006**, *124*, 084103.
(26) Förner, W.; Ladik, J.; Otto, P.; Čižek, J. *Chem. Phys.* **1985**, *97*, 251.
(27) Li, S.; Ma, J.; Jiang, Y. *J. Comput. Chem.* **2002**, *23*, 237.
(28) Li, S.; Shen, J.; Li, W.; Jiang, Y. *J. Chem. Phys.* **2006**, *125*, 074109.
(29) Saebø, S.; Baker, J.; Wolinski, K.; Pulay, P. *J. Chem. Phys.* **2004**, *120*, 11423.
(30) Azhary, A. E.; Rauhut, G.; Pulay, P.; Werner, H.-J. *J. Chem. Phys.* **1998**, *108*, 5185.
(31) Rauhut, G.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 4853.
(32) Schütz, M.; Werner, H.-J.; Lindh, R.; Manby, F. R. *J. Chem. Phys.* **2004**, *121*, 737.
(33) Ochsenfeld, C.; Head-Gordon, M. *Chem. Phys. Lett.* **1997**, *270*, 399.
(34) Larsen, H.; Helgaker, T.; Olsen, J.; Jorgensen, P. *J. Chem. Phys.* **2001**, *115*, 10344.
(35) Niklasson, A. M. N.; Challacombe, M. *Phys. Rev. Lett.* **2004**, *92*, 193001.
(36) Weber, V.; Niklasson, A. M. N.; Challacombe, M. *Phys. Rev. Lett.* **2004**, *92*, 193002.
(37) Stevens, R. M.; Pitzer, R. M.; Lipscomb, W. N. *J. Chem. Phys.* **1963**, *38*, 550.
(38) Stevens, R. M.; Lipscomb, W. N. *J. Chem. Phys.* **1964**, *41*, 3710.
(39) Yang, W. *Phys. Rev. Lett.* **1991**, *66*, 1438.
(40) Yang, W.; Lee, T.-S. *J. Chem. Phys.* **1995**, *103*, 5674.
(41) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2004**, *108*, *4301.*
(42) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 031103.
(43) Chen, X.; Zhang, Y.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105.
(44) Chen, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903.
(45) Li, W.; Li, S. *J. Chem. Phys.* **2005**, *122*, 194109.
(46) Gu, F. L.; Aoki, Y.; Korchowiec, J.; Imamura, A.; Kirtman, B. *J. Chem. Phys.* **2004**, *121*, 10385.
(47) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832.
(48) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *122*, 134103.
(49) Hirata, S.; Valiev, M.; Dupuis, M.; Xantheas, S. S.; Sugiki, S.; Sekino, H. *Mol. Phys.* **2005**, *103*, 2255.
(50) Morita, S.; Sakai, S. *J. Comput. Chem.* **2001**, *22*, 1107.
(51) Sakai, S.; Morita, S. *J. Phys. Chem. A* **2005**, *109*, 8424.
(52) Li, W.; Li, S. *J. Chem. Phys.* **2004**, *121*, 6649.
(53) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.
(54) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.
(55) Collins, M. A.; Deev, V. *J. Chem. Phys.* **2006**, *125*, 104104.
(56) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, *8777.*
(57) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.
(58) Li, W.; Fang, T.; Li, S. *J. Chem. Phys.* **2006**, *124*, 154102.
(59) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.
(60) The criteria for hydrogen bonds X−H⋯Y in our calculations is $r_{\mathrm{H}\cdots\mathrm{Y}} \le 2.9$ Å, $r_{\mathrm{X}\cdots\mathrm{Y}} \le 3.5$ Å, and ∠X−H⋯Y ≥ 120°.
(61) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.04. Gaussian, Inc.: Wallingford, CT, 2004.
(62) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.
(63) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1841.
(64) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, *129.*
(65) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.
(66) Foster, J. P.; Weinhold, F. *J. Am. Chem. Soc.* **1980**, *102*, 7211.
(67) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735.
(68) Cao, Y.; Friesner, R. A. *J. Chem. Phys.* **2005**, *122*, 104102.
(69) Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Phys. Chem. A* **2005**, *109*, 2.
(70) Li, S.; Li, W.; Fang, T.; Ma, J.; Jiang, Y. LSQC; Version 1.1. Nanjing University: Nanjing, 2006.
(71) 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.
(72) http://www.mpi.org/.