ACCOUNTS of chemical research

Article

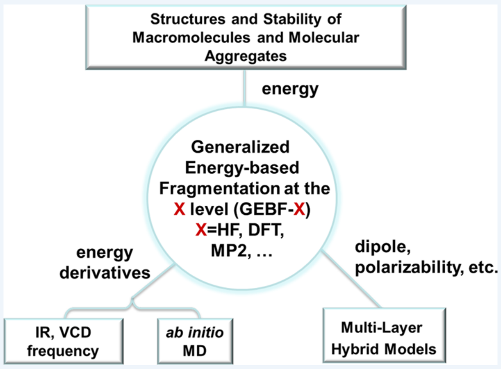
pubs.acs.org/accounts

Generalized Energy-Based Fragmentation Approach and Its Applications to Macromolecules and Molecular Aggregates

Shuhua Li,* Wei Li, and Jing Ma

School of Chemistry and Chemical Engineering, Key Laboratory of Mesoscopic Chemistry of MOE, Institute of Theoretical and Computational Chemistry, Nanjing University, Nanjing 210093, People's Republic of China

Supporting Information


Structures and Stability of Macromolecules and Molecular Aggregates
energy
Generalized Energy-based Fragmentation at the X level (GEBF-X) X=HF, DFT, MP2, ...
energy derivatives
dipole, polarizability, etc.
IR, VCD frequency
ab initio MD
Multi-Layer Hybrid Models

CONSPECTUS: The generalized energy-based fragmentation (GEBF) approach provides a very simple way of approximately evaluating the ground-state energy or properties of a large system in terms of ground-state energies of various small "electrostatically embedded" subsystems, which can be calculated with any traditional *ab initio* quantum chemistry (X) method (X = Hartree−Fock, density functional theory, and so on). Due to its excellent parallel efficiency, the GEBF approach at the X theory level (GEBF-X) allows full quantum mechanical (QM) calculations to be accessible for systems with hundreds and even thousands of atoms on ordinary workstations. The implementation of the GEBF approach at various theoretical levels can be easily done with existing quantum chemistry programs.

This Account reviews the methodology, implementation, and applications of the GEBF-X approach. This method has been successfully applied to optimize the structures of various large systems including molecular clusters, polypeptides, proteins, and foldamers. Such investigations could allow us to elucidate the origin and nature of the cooperative interaction in secondary structures of long peptides or the driving force of the self-assembly processes of aromatic oligoamides. These GEBF-based QM calculations reveal that the structures and stability of various complex systems result from a subtle balance of many types of noncovalent interactions such as hydrogen bonding and van der Waals interactions. The GEBF-based *ab initio* molecular dynamics (AIMD) method also allows the investigation of dynamic behaviors of large systems on the order of tens of picoseconds. It was demonstrated that the conformational dynamics of two model peptides predicted by GEBF-based AIMD are noticeably different from those predicted by the classical force field MD method.

With the target of extending QM calculations to molecular aggregates in the condensed phase, we have implemented the GEBF-based multilayer hybrid models, which could provide satisfactory descriptions of the binding energies between a solute molecule and its surrounding waters and the chain-length dependence of the conformational changes of oligomers in aqueous solutions. A coarse-grained polarizable molecular mechanics model, furnished with GEBF-X dipole moments of subsystems, exhibits some advantages of treating the electrostatic polarization with reduced computational costs. We anticipate that the GEBF approach will continue to develop with the ultimate goal of studying complicated phenomena at mesoscopic scales and serve as a practical tool to elucidate the structure and dynamics of chemical and biological systems.

1. INTRODUCTION

The rapid development of chemistry and related sciences calls for the application of accurate quantum mechanical (QM) methods and molecular dynamics techniques to more sophisticated systems. At the present stage, QM/MM (molecular mechanics) hybrid methods[1,2] are the most popular theoretical methods to describe electronic structures of a wide range of complex systems. In QM/MM methods, only a small active region of the system that is of major interest is treated with QM methods, while the remainder is treated classically using MM methods. With QM/MM methods, people now can understand many types of chemical processes in solution and in proteins. However, many interesting chemical problems require full QM treatments for quantitative descriptions, the assembly processes of molecular aggregates, the structures and dynamics of biomolecules, etc. The applications of QM calculations to the large-sized systems are usually prohibited by the high computational scaling of conventional QM methods. In general, only linear scaling methods (whose computational cost scales linearly with the system) can serve as powerful tools for complex systems. A variety of linear scaling approaches have been proposed for various quantum chemistry methods.[3−41] Especially, fragment-based approaches based on chemical intuition have emerged as practical tools for QM calculations of very large systems.[12−41] Among these approaches, the first fragment-based method is the explicit polarization method[16,17] developed for molecular clusters (or liquids), and more

Special Issue: Beyond QM/MM: Fragment Quantum Mechanical Methods

Received: January 28, 2014
Published: May 29, 2014

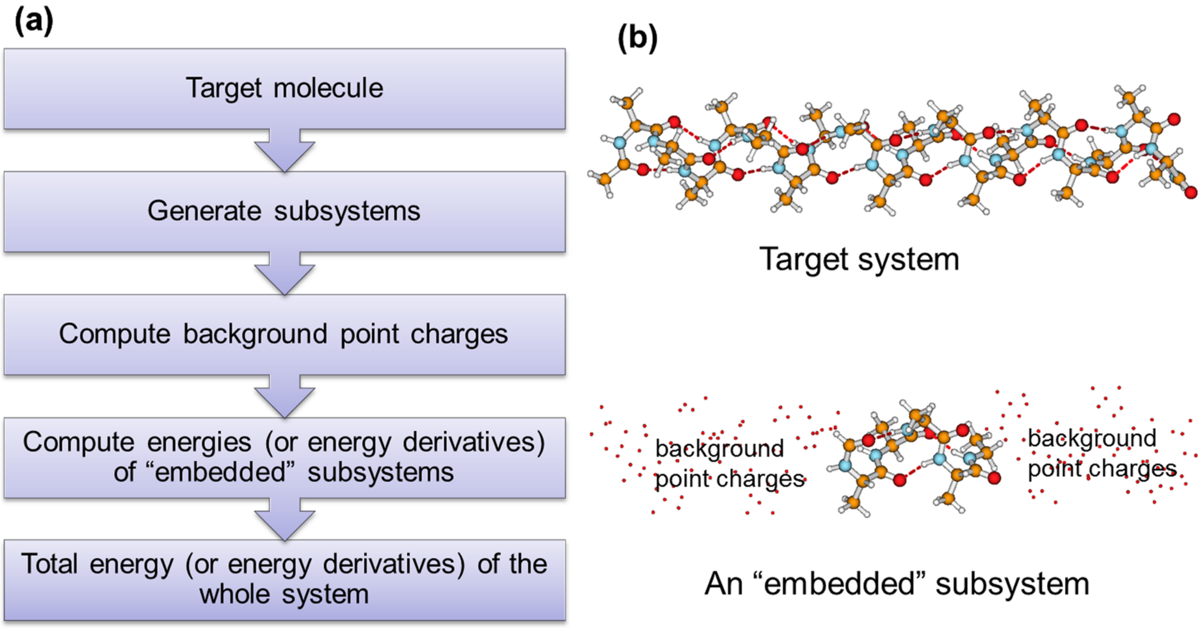

(a)
Target molecule
Generate subsystems
Compute background point charges
Compute energies (or energy derivatives) of "embedded" subsystems
Total energy (or energy derivatives) of the whole system
(b)
Target system
background point charges
background point charges
An "embedded" subsystem

Figure 1. GEBF-X (X = HF, DFT, MP2, etc.) procedure (a) and an illustrative picture of a target molecule and one of its "embedded" subsystems (b). For each subsystem, point charges are placed on all atoms outside of the subsystem.

advanced fragment-based approaches[22−41] have been developed to treat general large molecules and molecular clusters.

For general macromolecules or molecular aggregates, we have developed the generalized energy-based fragmentation (GEBF) approach.[28] This approach is an extension of the energy-corrected molecular fractionation with conjugated caps (EC-MFCC) approach developed by us for neutral systems[26] and its refined approach proposed by Ma and co-workers[27] for systems with charged groups. In the GEBF approach,[28−36] the ground-state energy of a large system (macromolecule or molecular cluster) can be directly estimated with energies of small "electrostatically embedded" subsystems, which can be computed easily with existing quantum chemistry programs. Thus, the GEBF approach has a much wider range of applicability, since it works well not only for neutral systems but also for systems with charged and polar groups. Our recent works[42−47] have demonstrated that the GEBF approach can provide satisfactory descriptions at various *ab initio* levels for a wide range of complex systems, which are beyond the capability of traditional QM methods. In this Account, we will describe briefly the principles and implementation details of the GEBF method and provide some examples to show how the GEBF approach can serve as a powerful tool in predicting structures, stability, and dynamics of various complex systems.

2. GENERALIZED ENERGY-BASED FRAGMENTATION APPROACH AND ITS IMPLEMENTATION

2.1. Ground-State Energies

Within the GEBF approach, the ground-state energy of a large molecule can be directly assembled from ground-state energies of various subsystems, each of which is embedded in the background point charges at all distant atoms (outside this subsystem). With such "embedded" subsystems, the electrostatic interaction between any two distant fragments in the target system is approximately taken into account. The total energy of a large system can be evaluated with the following expression:[28]

$$E_{\mathrm{tot}} = \sum_{m}^{M} C_m \tilde{E}_m - \left[\left(\sum_{m}^{M} C_m\right) - 1\right] \sum_{A} \sum_{B>A} \frac{Q_A Q_B}{R_{AB}} \qquad (1)$$

where $\tilde{E}_m$ stands for the total energy of the mth subsystem (including the self-energy of background point charges),C_m represents the coefficient of the mth subsystem, Q_A is the net atomic charge on atom A, and M is the total number of subsystems. The GEBF approach has been demonstrated to work at various theoretical levels, such as Hartree−Fock (HF), second-order Møller−Plesset perturbation theory (MP2), density functional theory (DFT), and couple cluster singles and doubles (CCSD). For convenience, we use the abbreviation GEBF-X to represent a GEBF calculation at the X level (X = HF, DFT, etc.). The general procedure for performing a single-point GEBF-X calculation is illustrated in Figure 1a, which includes the following steps: (1) The fragmentation of a large molecule into N fragments by cutting single bonds or hydrogen bonds. With a number of functional groups stored in the database, one can achieve the automatic fragmentation of a general molecule.[32] (2) The construction of various subsystems in terms of fragments, and the derivation of the coefficients (C_m) occurring in eq 1. For each fragment, a primitive subsystem centered on this fragment can be formed by adding environmental fragments within a given distance threshold (ξ) (usually taken as 4.0 Å). Hydrogen atoms are added as capping atoms for valence saturation if necessary. The coefficients of these primitive subsystems are all set to +1. Assume that the maximum number of fragments in all primitive subsystems is M. By decomposing the total energy of each primitive subsystem in terms of n-fragment terms ($n \leq M$) and summing up the identical terms, one can find that some terms may have the coefficients different from +1 (due to the overlapping of some primitive subsystems). Thus, to eliminate the overcounting of some multifragment (or one-fragment) terms, one should construct a series of smaller subsystems (called derivative subsystems). The coefficients of derivative subsystems are determined so that the net number of any specific n-fragment term is +1. The details of this procedure have been described elsewhere.[32] An example for illustrating this procedure is given in Supporting Information. (3) The computation of the net atomic charges on all atoms in the target system. In the first step, by performing HF (or DFT) calculations on isolated primitive subsystems, one could obtain initial atomic charges on all atoms (we put one point charge per atom at the nuclear center). Then, refined atomic charges are extracted from HF or DFT calculations on "electrostatically embedded" subsytems. An iterative procedure[28] can be used to obtain converged atomic charges. (4) Conventional *ab initio* calculations at the X level on all "embedded" subsystems (an illustrative picture of one subsystem is given in Figure 1b). Then, one can employ the energies of all subsystems to compute the total energy of the target system.

2.2. Energy Derivatives and Molecular Properties

Molecular properties describe the response of the molecular system to an external perturbation. If the external electrical field is weak (the usual case), the dipole moment and static polarizability of a large molecule can be approximately evaluated within the GEBF approach as[28]

$$\Omega_{\mathrm{tot}} = \sum_{m}^{M} C_m \tilde{\Omega}_m \qquad (\Omega = \mu_i, \alpha_{ij}, \ldots) \tag{2}$$

where $\tilde{\Omega}_m$ is the corresponding property of the mth "embedded" subsystem. Other properties of chemical interests are the geometrical derivatives of the total energy with respect to nuclear displacements. Within the GEBF approach, the fully analytic energy gradients of the target system can be expressed as follows:

$$\frac{\partial E_{\mathrm{tot}}}{\partial \mathbf{q}_A} = \sum_{n} C_n \left(\frac{\partial \tilde{E}_n}{\partial \mathbf{q}_A} - \mathbf{F}_{n,a} Q_a - \sum_{b} \mathbf{f}_{ab} \right) + \left[\left(\sum_{m}^{M} C_m \right) - 1 \right] \sum_{b \in \mathrm{all}} \mathbf{f}_{ab} \tag{3}$$

where A denotes a real atom in a given subsystem, a and b denote the point-charge centers, $\mathbf{F}_{n,a}$ is the electric field generated by the nth subsystem on the center a (which can be calculated with some existing *ab initio* programs), and $\mathbf{f}_{ab}$ represents the Coulomb force between charge on b and charge on a,

$$\mathbf{f}_{ab} = \frac{Q_a Q_b}{|\mathbf{q}_a - \mathbf{q}_b|^3} (\mathbf{q}_a - \mathbf{q}_b) \qquad (b \neq a) \tag{4}$$

In previous studies,[32] we demonstrated that the formula listed below is an excellent approximation to the GEBF energy gradients described above:

$$\frac{\partial E_{\mathrm{tot}}}{\partial \mathbf{q}_A} \approx \sum_{n} C_n \frac{\partial \tilde{E}_n}{\partial \mathbf{q}_A} \tag{5}$$

Hence, the gradients of the GEBF energy on a given atom can be evaluated with the corresponding gradients on this atom in some subsystems, which include this atom as a real atom. The details of the derivation of the above GEBF energy gradient can be found in ref 32. To illustrate the accuracy of the GEBF energy gradients, we have compared the HF and GEBF-HF energy gradients for 10 randomly selected structures during the optimization of a hydrogelator (shown in Figure S1, Supporting Information) at the 6-31G(d) basis set. The results (Table S1, Supporting Information) show that the root-mean-squared deviations (RMSDs) between the HF and GEBF-HF gradients at all atoms are only around 0.0005 au/bohr with the formula in eq 3 or 0.0006 au/bohr with the formula in eq 5. In addition, the GEBF-HF optimized geometries for this system with both fully analytic gradients and approximate gradients are very close to each other. The RMSD between the full system HF-optimized structure and the GEBF-HF optimized structures (obtained with both gradient formulas) is less than 0.08 Å. The results presented here (and reported previously[32,43,45]) indicate that the GEBF energy gradients are very good approximations to the full system energy gradients. For many other systems, our recent calculations also show that the calculations with both formulas, eq 3 or eq 5, usually give almost identical results. In a very similar way, the second derivatives of the total energy (or the Hessian matrix) can also be computed from the second derivatives of the total energies of various "embedded" subsystems. Thus, with the GEBF approach, one can easily determine stationary points and compute vibrational frequencies and intensities of normal modes.[29] In the following, the energy gradients from eq 5 are employed for all geometry optimizations and molecular dynamics (MD) simulations. In addition, various thermochemistry data (such as enthalpy, free energy, etc.) at a given temperature and pressure can also be obtained.

The implementation of the GEBF approach is easy and straightforward. We have developed the LSQC[48] program to perform GEBF-based *ab initio* calculations. The main functions of the LSQC program include (1) the construction of various subsystems, (2) the generation of input files for subsystem calculations, and (3) the calculation of the total energy or energy derivatives from the corresponding quantities of various subsystems. The calculations of all subsystems are carried out with some popular quantum chemistry packages such as Gaussian 09.[49]

3. STRUCTURES AND STABILITY OF COMPLEX SYSTEMS

3.1. Water Clusters

Exploring the structures and properties of various molecular aggregates is fundamentally important in physical, chemical, and biological fields. The GEBF approach can greatly extend the applications of *ab initio* calculations to the complex molecular clusters. For example, the combination of *ab initio*-based GEBF method and the polarizable AMOEBA potential[50] has been employed to investigate low-energy structures and stability of water clusters $(H_2O)_n$ with n = 20–30.[43] First, a large database of low-lying isomers is generated by using a modified basin-hopping optimization method with the AMOEBA potential. Then, for selected dozens of low-lying structures, we optimized their structures at the GEBF-B3LYP/6-311++G(d,p) level. Our calculations indicate that for water clusters a transition from one-centered to two-centered cage structure first appears at n = 26, and the number of hydrogen bonds per water molecule in the lowest-energy structure tends to increase gradually with increasing the cluster size. The lowest-energy structures predicted by GEBF-B3LYP and GEBF-MP2 are different in some water clusters (the lowest-energy structures of three water clusters at the GEBF-MP2/6-311++G(3df,2p) level are shown in Figure 2a). On the other hand, the combination of the GEBF approach with the explicitly correlated F12 methods[51] can be applied to obtain

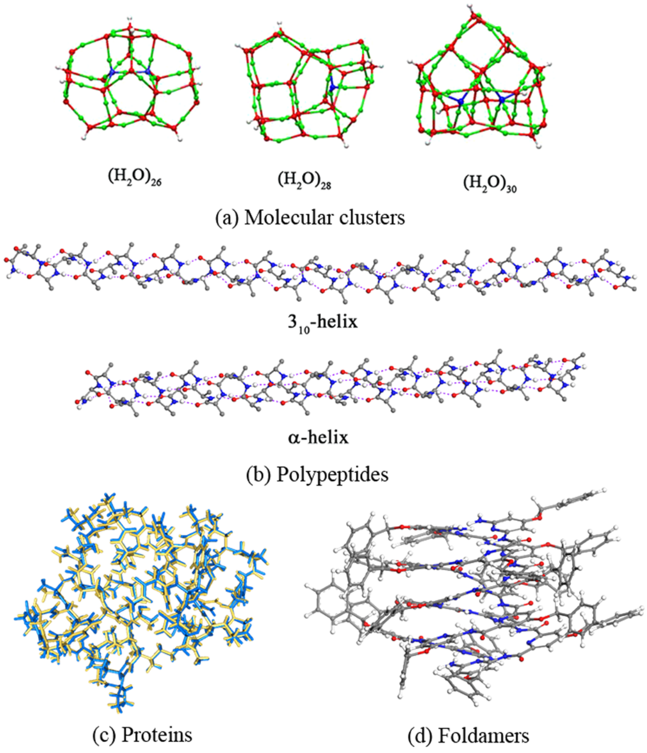


Figure 2. Optimized structures obtained with the GEBF approach: (a) structures of the lowest-energy water clusters $(H_2O)_n$ (n = 26, 28, 30) at the GEBF-MP2/6-311++G(3df,2p) level; (b) optimized α-, and 3_{10}-helical structures of the polypeptide acetyl$(Ala)_{40}NH_2$, obtained at the GEBF-M06-2X/6-31G** level; (c) superposition between the optimized structure (yellow) of crambin at the GEBF-M06-2X/6-31G level and the corresponding crystal structure (blue); (d) double helical foldamer formed from two tridecameric strands optimized at the GEBF-B3LYP(vdW)/6-31G** level.

very accurate relative energies of molecular clusters. With the GEBF-CCSD(T)-F12a/HF method,[35] we computed the average binding energies (ABEs) for four water clusters $(H_2O)_n$ (n = 15−18) at the aug-cc-pVDZ basis set. With the X3LYP global minimum structures,[52] we have reoptimized the structures of these clusters with the PWB6K functional.[53] The GEBF-CCSD(T)-F12a/HF results show that the ABE per water molecule in the four clusters increases gradually from 10.6 to 10.9 kcal/mol with increasing the cluster size (Figure S2, Supporting Information). The calculated ABEs here should be very close to the complete basis set CCSD(T) values. It could be expected that the GEBF-based explicitly correlated methods can provide highly accurate information on structures and stabilities of general molecular clusters.

3.2. Secondary Structures of Polypeptides

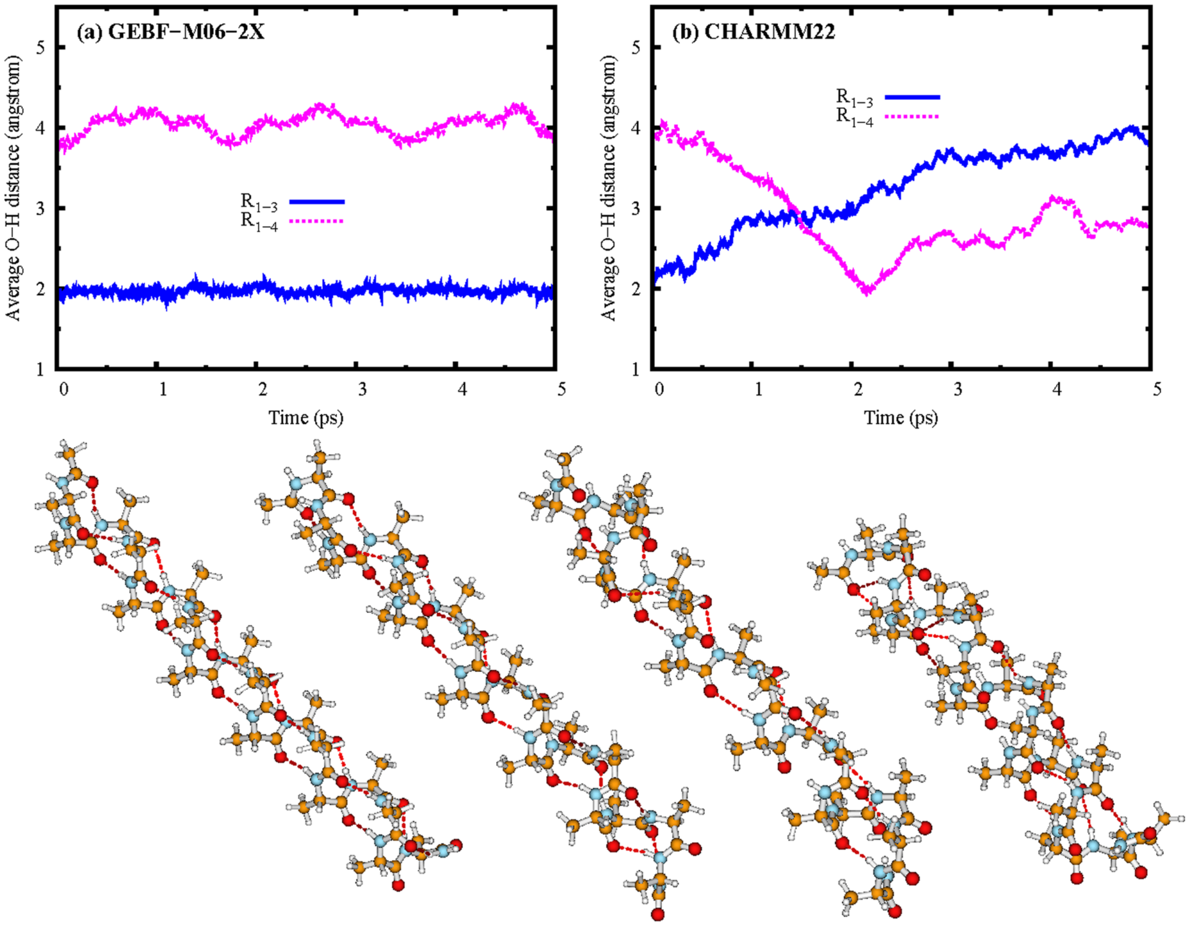
Understanding the factors that control the interconversion among secondary structures of polypeptides remains an active subject in peptide chemistry. Previous *ab initio* studies are limited to relatively small polypeptides. Relative to the β-strand structures, both α- and 3_{10}-helical structures exhibit significant amounts of cooperativity, due to the collective interaction of inter-residue hydrogen bonds. Until recently, the origin and nature of the cooperative interaction in long helical structures have not been well recognized. To elucidate this issue, we have carried out a computational study[45] on three typical secondary structures of a series of polyalanines, acetyl$(Ala)_NNH_2$. With GEBF-MP2 energies as reference data, we find that the M06-2X functional[54] can provide satisfactory results, while the popular B3LYP functional is not reliable for long polyalanines, due to the lack of van der Waals (vdW) interactions.[55] The optimized structures obtained with GEBF-M06-2X/6-31G** for two helical conformations of the polypeptide with 40 residues (with 409 atoms) are shown in Figure 2b. One can see that in α-helix the average HB length is somewhat longer than that in 3_{10}-helix and the backbone length is shorter than that in 3_{10}-helix. In general, α-helices are more stable than 3_{10}-helices for systems with more than 10 residues, and the relative stability of α-helices (relative to the β-strand counterpart) decreases more steeply than those of 3_{10}-helices with increasing the number of residues. Detailed analysis shows that the long-range electrostatic interaction tends to stabilize 3_{10}-helices more than α-helices, and the stronger cooperativity in α-helices over 3_{10}-helices is mainly from the dispersion-like interaction. This result is in contradiction with the traditional viewpoint that the long-range electrostatic interaction stabilizes α-helices over 3_{10}-helices.[56]

3.3. Proteins

With the GEBF-DFT approach, now we can routinely obtain optimized structures for proteins with hundreds or thousands of atoms. For example, the optimized structure of crambin (with 642 atoms) obtained at the GEBF-M06-2X/6-31G level (with the crystal structure as the initial structure) is displayed in Figure2c, together with the experimentally measured X-ray structure for comparison. The RMSD between the two structures is only 0.46 Å. Thus, the GEBF-DFT approach can be used to predict the three-dimensional structures of some proteins, whose structures are experimentally unknown. In such cases, one may use threading and homology modeling methods[57] to build initial three-dimensional structures for subsequent geometry optimizations.

3.4. The Self-Assembly Process of Aromatic Oligoamides

The folding or assembling processes of artificial oligomers into foldamers are usually dominated by many types of noncovalent interactions. Quantitative understanding of the factors that control these processes is still lacking, since the artificial oligomers synthesized by experimental chemists are usually beyond the capability of conventional QM calculations. We have applied the GEBF-DFT(vdW) (DFT with empirical vdW correction[55]) approach to investigate the energetics of the self-assembly processes of several aromatic oligoamides (based on 2,6-diaminopyridine and 2,6-pyridine dicarboxylic acids).[42] For two oligomeric species, our geometry optimizations at the 6-31G** level led to two double-helical structures. In both cases, the RMSD value between the crystal and optimized structures is small. To understand why two single helices will assemble into a double-helical structure, we have calculated the dimerization energy ($\Delta E = E_{\text{double}} - 2E_{\text{single}}$) to measure the driving force for the self-assembly process. Our calculations show that for both species the formation of the dimeric species from two single helical strands is an energetically favorable process, and the interstrand vdW interaction offers the main driving force for this self-assembly process. Furthermore, for one compound, we also optimized the structures of single-helical strands and double-helical foldamers for two longer homologues (nonamer and tridecamer). The optimized structure of the tridecameric dimer (with 490 atoms) is displayed in Figure 2d. The dimerization energy is estimated to be −5.1 kcal/mol for the nonameric strand and +8.8 kcal/mol for the tridecameric strand, respectively. Thus, the double-helix



(c) GEBF-M06-2X: 1ps (d) GEBF-M06-2X: 3ps (e) CHARMM22: 1ps (f) CHARMM22: 3ps

Figure 3. Average values of 1−3 type and 1−4 type O−H distances (except those involving terminal residues), which can characterize 3_{10}- and α-helical structures, respectively, as functions of time in MD simulations of $(Ala)_{18}$ (with the 3_{10}-helical structure as the initial structure) with (a) GEBF-M06-2X and (b) CHARMM22 force field at 300 K. Here the 1−3 (or 1−4) O−H distance indicates the distance between the O atom in the *j*th residue and the H atom in the (*j*+3)th [or (*j*+4)th] residue. Four snapshots at times of 1.0 and 3.0 ps from GEBF-M06-2X and CHARMM22 MD simulations are shown in parts c−f, respectively.

dimer from the tridecameric strand is predicted to be thermodynamically unstable. This result is in qualitative agreement with the related experimental facts that short aromatic oligoamides can form double-helical dimers, but their longer analogues cannot.[58]

4. STRUCTURAL DYNAMICS OF BIOMOLECULES FROM GEBF-BASED MOLECULAR DYNAMICS

The dynamic properties of large systems can be investigated by GEBF-based *ab initio* MD (AIMD) method, in which forces on nuclei are obtained by "on the fly" GEBF-based QM calculations. To investigate the behavior of the GEBF-based AIMD, we first perform a microcanonical (NVE) simulation at the GEBF-M06-2X/STO-3G level with a time step of 0.25 fs for a 3_{10}-helical model peptide, acetyl$(Ala)_{18}NH_2$ (abbreviated as $(Ala)_{18}$) to check whether the total energy is conserved. The change of energy in the 3000 time steps (0.75 ps) is displayed in Figure S3 (Supporting Information). One can see from Figure S3 that the energy change is about 0.004 au, being comparable to the energy change (0.003 au) in full system AIMD simulations with the same conditions. This result shows that GEBF-AIMD is applicable for short time (approximately picoseconds) AIMD simulation since the deviations are on the order of about 0.0001% with respect to the total energy of the system (4601 au). To illustrate the performance of the GEBF-based AIMD method, we have investigated the conformational dynamics of two model peptides, $(Ala)_8$ and $(Ala)_{18}$. The AIMD simulations were carried out at the GEBF-M06-2X/6-31G level in the canonical (NVT) ensemble for both systems at 300 K. The trajectories were propagated with the modified Beeman algorithm with a time step of 1 fs. In addition, under the same conditions, we also perform AIMD simulations with the conventional M06-2X method for $(Ala)_8$ (with the α-helical structure as the initial structure) and classical MD simulations with CHARMM22 force field[59] method for comparison. For $(Ala)_{18}$ (with the 3_{10}-helical structure as the initial structure), we only perform GEBF-M06-2X/6-31G AIMD and classical MD simulations, since the corresponding M06-2X AIMD simulations are very time-consuming.

By monitoring the variation of the average 3_{10}-type and α-type O−H distances in hydrogen bonds between residues (except those involving terminal residues) in each structure, one can get some insight into the conformational changes between helical structures for peptides under study. As shown in Figure S4 in Supporting Information, both M06-2X and GEBF-M06-02X AIMD calculations predict that at 300 K the conformation change from α-helical to 3_{10}-helical structure does occur in the 5 ps simulation in $(Ala)_8$. However, classical

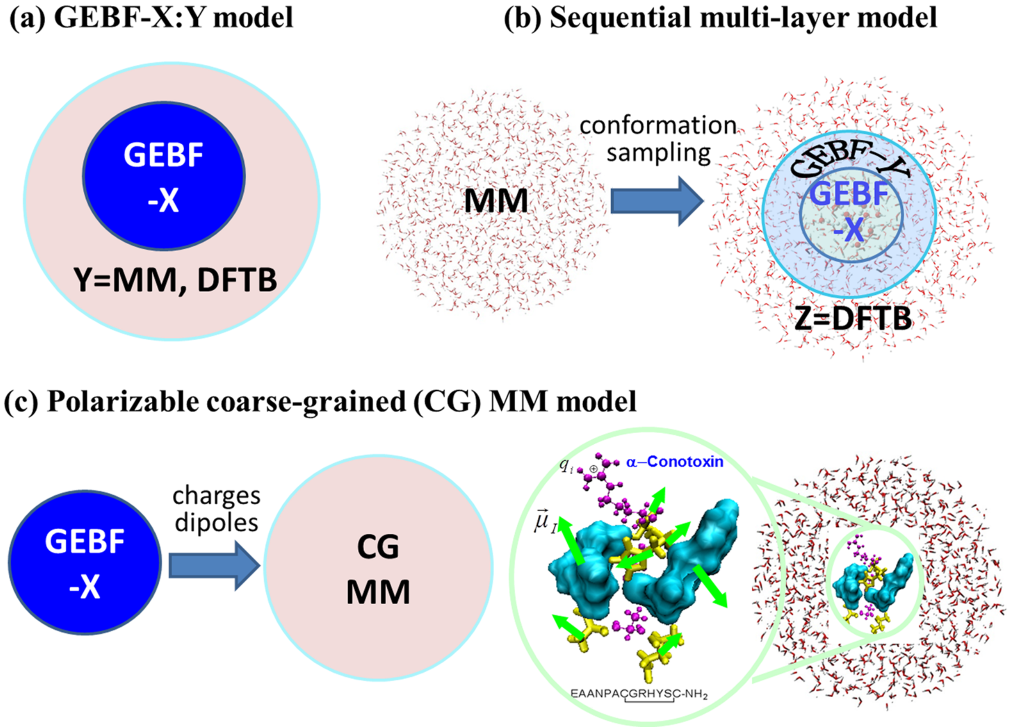


Figure 4. GEBF-based multilayer hybrid models: (a) two-layer GEBF-X/Y ONIOM model; (b) sequential multilayer model; first, conformational sampling is performed at the MM level, and then average properties are calculated with the three-layer GEBF-X/GEBF-Y/Z (X = MP2-F12, Y = MP2, Z = DFTB) ONIOM model; (c) GEBF-based polarizable coarse-grained (CG) MM model. The fragment-based charges and dipoles required in the CG MM model are obtained from GEBF-based QM calculations at some snapshots.

MD simulations with the CHARMM22 force field fail to describe the conformation change between two helical structures during the simulation time. This result is in accord with the fact that the optimized α-helical structure at the GEBF-M06-2X/6-31G level is higher in energy (by 2.3 kcal/mol) than the corresponding 3_{10}-helical structure, whereas a reverse order is predicted by the CHARMM22 force field (Table S2, Supporting Information). For $(\mathrm{Ala})_{18}$, one can see from Figure 3 that at 300 K the 3_{10}-helical structure remains almost unchanged during the 5 ps GEBF-M06-2X AIMD simulations, but it quickly transforms into the α-helical structure (at the time of about 2 ps) in the classical MD simulations with the CHARMM22 force field. For this system, GEBF-M06-2X and CHARMM22 predict that the α-helical structure is more stable than the 3_{10}-helical structure by 16.9 and 52.7 kcal/mol, respectively. The very different dynamic behaviors of $(\mathrm{Ala})_{18}$ at two theoretical levels should be ascribed to the big energy difference predicted by two methods. From MD simulations on both peptides, we can conclude that the force field methods may give incorrect descriptions on dynamics of large biomolecules in some cases. Our illustrative applications show that GEBF-based AIMD may provide a practical tool to investigate the dynamic behaviors of many interesting chemical systems on the order of tens of picoseconds.

5. GEBF-BASED MULTILAYER MODELS FOR COMPLEX SYSTEMS IN CONDENSED PHASE

The extension of the GEBF approach to the multilayer hybrid energy models is necessary to treat even more complicated systems in condensed phase. The essence of the multilayer models is to treat different parts of a large system at different theoretical levels, spanning from *ab initio* electronic structure methods to MM methods. The GEBF-based multilayer models may be implemented in either simultaneous or sequential ways, as shown in Figure 4. In simultaneous two-layer GEBF-based ONIOM model (Figure 4a), the small active part is treated with the high-level GEBF-X (X = MP2, CCSD, etc.) and the remaining part is treated with a low-level Y method (Y = HF, DFT, MM, etc.). The only difference between the GEBF-based ONIOM models and conventional ONIOM models is that the conventional X calculation is replaced with the GEBF-X calculation. The advantage of this GEBF-based ONIOM model over the conventional ONIOM models is that a much larger active part can be treated with the GEBF-X method.

With the GEBF-based ONIOM models, we can perform MD or Monte Carlo (MC) simulations to sample the conformational space of large systems in liquid and solutions. For example, we have carried out the GEBF-QM/MM MD simulations on two polymers, poly(ethylene oxide) (PEO) and polyethylene (PE), with various chain lengths in aqueous solutions to investigate their dynamic curling behaviors.[31] In such MD simulations, oligomers with up to 30 repeat units are treated with the GEBF-HF method, and water solvents are treated with the MM method. The time-dependence of calculated end-to-end distances indicates that the curling rate of both polymers will first increase and then decrease with increasing number of atoms in the main chain, but PE tends to have a larger curling rate than PEO (due to the strong hydrogen bonding interactions between water solvents and PEO).

On the other hand, sequential multilayer models are also adopted in studying complex systems, in which several calculations are carried out at different theoretical levels. For instance, one may sample the conformational space at the MM level and calculate the average energies or properties with the high-level GEBF-based ONIOM models. Recently, we have employed a three-layer GEBF-based ONIOM model (Figure 4b) to investigate the binding energies between a methanol molecule and its surrounding waters molecules in dilute methanol aqueous solutions.[35] In a large cluster model (with 3351 atoms), a methanol and its neighboring water molecules within 4 Å (first solvent shell), the water molecules between 4 and 9 Å, and the remaining part (up to 20 Å) are treated with the GEBF-MP2-F12, GEBF-MP2, and DFTB (density functional tight-binding), respectively. We found that the long-range interaction between a methanol molecule and water molecules within about 9 Å is essential to obtain the accurate binding energy. This study suggests that very large cluster models should be used in studying the properties of polar molecules in polar solvents.

In another type of hybrid model, we use the GEBF-based QM calculations to get atomic charges and dipole moments (and other properties) for predefined fragments, which may provide input parameters for polarizable MM methods. In most polarizable MM models, the electrostatic contributions are computed with atom-based charges and dipole moments, whose computational cost scales quadratically with the total number of atoms. Recently, we suggested an alternative polarizable coarse-grained (CG) MM model (Figure 4c),[46,47] in which a set of fragment-based electrostatic parameters are used to dramatically reduce the computational cost. Here a fragment may be chosen as a secondary structure, a residue, or even an atom. The selection of these fragments can be achieved according to the variation of fragment-based charges or dipoles in response to the electrostatic environment. These parameters may be obtained from GEBF-based QM calculations at some snapshots. Our test calculations show that this CG polarization model within the FF03 framework can provide reasonably reliable descriptions on the relative stabilities of various α-conotoxin peptides, in good agreement with other polarizable force field methods. This polarizable CG MM model is useful in studying some biologically interesting but very complicated phenomena (such as protein−protein interactions).

6. CONCLUSIONS AND OUTLOOK

The GEBF approach has been demonstrated to be very successful in predicting structures, stabilities, and properties of a wide range of complex systems such as molecular aggregates, polypeptides, proteins, and supramolecules. GEBF-based *ab initio* MD simulations are now routinely available for exploring the gas-phase dynamic behaviors of many chemically interesting large systems. For large systems in condensed phase, one may employ GEBF-based multilayer hybrid models to understand their conformational preferences and dynamic properties. However, the extension of the GEBF approach to periodic systems is highly desirable to treat solid state materials and liquids. With this development, the applications of GEBF-based simulations will be dramatically expanded. On the other hand, we should mention that the GEBF approach is not sufficiently accurate for systems with strongly delocalized electrons (like metallic compounds). Despite this limitation, we expect that the GEBF approach will play an increasingly important role in interpreting and understanding existing experimental facts on various complex systems.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

An example for illustrating the construction of subsystems, the comparison between the HF and GEBF-HF energy gradients for an example, energy differences between two helical structures of $(Ala)_8$ and $(Ala)_{18}$ calculated with different methods, the PWB6K-optimized lowest-energy structures of water clusters $(H_2O)_n$ (n = 15−18), energy changes as functions of time in AIMD simulations of $(Ala)_{18}$, and the average O−H distances in $(Ala)_8$ as a function of time in MD simulations. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Shuhua Li. E-mail: shuhua@nju.edu.cn.

Notes

The authors declare no competing financial interest.

Biographies

Shuhua Li was born in 1969 in Hunan, China. He received his Ph.D. from Nanjing University in 1996 and is currently professor of theoretical chemistry at Nanjing University. His research interests focus on the development of novel electron correlation methods and linear scaling electronic structure algorithms.

Wei Li was born in 1979 in Jiangsu, China. He is currently associate professor of theoretical chemistry at Nanjing University. His research includes the development and implementation of electron correlation methods for complex systems.

Jing Ma was born in 1971 in Jiangsu, China. She received her Ph.D. in Chemistry from Nanjing University and is currently professor of theoretical chemistry at Nanjing University. Her research applies multiscale simulation models to molecules in condensed phase or at interfaces.

■ ACKNOWLEDGMENTS

This work was supported by Grant 2011CB808501 (S. Li) from the National Basic Research Program of China and Grants 21333004 (S. Li), 21103086 (W. Li), and 21290190 (J. Ma) from the National Natural Science Foundation of China. We are thankful to Dr. Tao Fang for his contribution in developing the GEBF energy gradients.

■ REFERENCES

(1) Warshel, A.; Karplus, M. Calculation of Ground and Excited State Potential Surfaces of Conjugated Molecules. I. Formulation and Parametrization. *J. Am. Chem. Soc.* **1972**, *94*, 5612−5625.

(2) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227−249.

(3) Yang, W. Direct Calculation of Electron Density in Density-Functional Theory. *Phys. Rev. Lett.* **1991**, *66*, 1438−1441.

(4) Millam, J. M.; Scuseria, G. E. Linear Scaling Conjugate Gradient Density Matrix Search as an Alternative to Diagonalization for First Principles Electronic Structure Calculations. *J. Chem. Phys.* **1997**, *106*, 5569−5577.

(5) Li, X.-P.; Nunes, R. W.; Vanderbilt, D. Density-Matrix Electronic-Structure Method with Linear System-Size Scaling. *Phys. Rev. B* **1993**, *47*, 10891−10894.

(6) Pulay, P. Localizability of Dynamic Electron Correlation. *Chem. Phys. Lett.* **1983**, *100*, 151−154.

(7) Hampel, C.; Werner, H.-J. Local Treatment of Electron Correlation in Coupled Cluster Theory. *J. Chem. Phys.* **1996**, *104*, 6286−6297.

(8) Ayala, P. Y.; Scuseria, G. E. Linear Scaling Second-Order Møller-Plesset Theory in the Atomic Orbital Basis for Large Molecular Systems. *J. Chem. Phys.* **1999**, *110*, 3660−3671.

(9) Li, S.; Ma, J.; Jiang, Y. Linear Scaling Local Correlation Approach for Solving the Coupled Cluster Equations of Large Systems. *J. Comput. Chem.* **2002**, *23*, 237−244.

(10) Li, S.; Shen, J.; Li, W.; Jiang, Y. An Efficient Implementation of the "Cluster-in-Molecule" Approach for Local Electron Correlation Calculations. *J. Chem. Phys.* **2006**, *125*, No. 074109.

(11) Li, W.; Piecuch, P.; Gour, J. R.; Li, S. Local Correlation Calculations using Standard and Renormalized Coupled-Cluster Approaches. *J. Chem. Phys.* **2009**, *131*, No. 114109.

(12) He, X.; Zhang, J. Z. H. A New Method for Direct Calculation of Total Energy of Protein. *J. Chem. Phys.* **2005**, *122*, No. 031103.

(13) Li, W.; Li, S. A Localized Molecular-Orbital Assembler Approach for Hartree-Fock Calculations of Large Molecules. *J. Chem. Phys.* **2005**, *122*, No. 194109.

(14) Gu, F. L.; Aoki, Y.; Korchowiec, J.; Imamura, A.; Kirtman, B. A New Localization Scheme for the Elongation Method. *J. Chem. Phys.* **2004**, *121*, 10385−10391.

(15) Kobayashi, M.; Imamura, Y.; Nakai, H. Alternative Linear-Scaling Methodology for the Second-Order Møller-Plesset Perturbation Calculation Based on the Divide-and-Conquer Method. *J. Chem. Phys.* **2007**, *127*, No. 074103.

(16) Gao, J. Toward a Molecular Orbital Derived Empirical Potential for Liquid Simulations. *J. Phys. Chem. B* **1997**, *101*, 657−663.

(17) Gao, J. A Molecular-Orbital Derived Polarization Potential for Liquid Water. *J. Chem. Phys.* **1998**, *109*, 2346−2354.

(18) Gao, J.; Wang, Y. Variational Many-Body Expansion: Accounting for Exchange Repulsion, Charge Delocalization, and Dispersion in the Fragment-Based Explicit Polarization Method. *J. Chem. Phys.* **2012**, *136*, No. 071101.

(19) Hirata, S.; Valiev, M.; Dupuis, M.; Xantheas, S. S.; Sugiki, S.; Sekino, H. Fast Electron Correlation Methods for Molecular Clusters in the Ground and Excited States. *Mol. Phys.* **2005**, *103*, 2255−2265.

(20) Sakai, S.; Morita, S. Ab Initio Integrated Multi-Center Molecular Orbitals Method for Large Cluster Systems: Total Energy and Normal Vibration. *J. Phys. Chem. A* **2005**, *109*, 8424−8429.

(21) Dahlke, E. E.; Truhlar, D. G. Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters. *J. Chem. Theory Comput.* **2007**, *3*, 46−53.

(22) Fedorov, D. G.; Kitaura, K. Second Order Møller-Plesset Perturbation Theory Based upon the Fragment Molecular Orbital Method. *J. Chem. Phys.* **2004**, *121*, 2483−2490.

(23) Nagata, T.; Fedorov, D. G.; Kitaura, K.; Gordon, M. S. A Combined Effective Fragment Potential-Fragment Molecular Orbital Method. I. The Energy Expression and Initial Applications. *J. Chem. Phys.* **2009**, *131*, No. 024101.

(24) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632−672.

(25) Li, W.; Li, S. Divide-and-Conquer Local Correlation Approach to the Correlation Energy of Large Molecules. *J. Chem. Phys.* **2004**, *121*, 6649−6657.

(26) Li, S.; Li, W.; Fang, T. An Efficient Fragment-Based Approach for Predicting the Ground-State Energies and Structures of Large Molecules. *J. Am. Chem. Soc.* **2005**, *127*, 7215−7226.

(27) Jiang, N.; Ma, J.; Jiang, Y. Electrostatic Field-adapted Molecular Fractionation with Conjugated Caps for Energy Calculations of Charged Biomolecules. *J. Chem. Phys.* **2006**, *124*, No. 114112.

(28) Li, W.; Li, S.; Jiang, Y. Generalized Energy-Based Fragmentation Approach for Computing the Ground-State Energies and Properties of Large Molecules. *J. Phys. Chem. A* **2007**, *111*, 2193−2199.

(29) Hua, W.; Fang, T.; Li, W.; Yu, J.-G.; Li, S. Geometry Optimizations and Vibrational Spectra of Large Molecules from a Generalized Energy-Based Fragmentation Approach. *J. Phys. Chem. A* **2008**, *112*, 10864−10872.

(30) Li, S.; Li, W. Fragment Energy Approach to Hartree-Fock Calculations of Macromolecules. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2008**, *104*, 256−271.

(31) Li, H.; Li, W.; Li, S.; Ma, J. Fragmentation-Based QM/MM Simulations: Length Dependence of Chain Dynamics and Hydrogen Bonding of Polyethylene Oxide and Polyethylene in Aqueous Solutions. *J. Phys. Chem. B* **2008**, *112*, 7061−7070.

(32) Hua, S.; Hua, W.; Li, S. An Efficient Implementation of the Generalized Energy-Based Fragmentation Approach for General Large Molecules. *J. Phys. Chem. A* **2010**, *114*, 8126−8134.

(33) Jiang, N.; Tan, R. X.; Ma, J. Simulations of Solid-State Vibrational Circular Dichroism Spectroscopy of (S)-Alternarlactam by Using Fragmentation Quantum Chemical Calculations. *J. Phys. Chem. B* **2011**, *115*, 2801−2813.

(34) Hua, S.; Li, W.; Li, S. The Generalized Energy-Based Fragmentation Approach with an Improved Fragmentation Scheme: Benchmark Results and Illustrative Applications. *ChemPhysChem* **2013**, *14*, 108−115.

(35) Li, W. Linear Scaling Explicitly Correlated MP2-F12 and ONIOM Methods for the Long-Range Interactions of the Nanoscale Clusters in Methanol Aqueous Solutions. *J. Chem. Phys.* **2013**, *138*, No. 014106.

(36) Wang, K.; Li, W.; Li, S. Generalized Energy-Based Fragmentation CCSD(T)-F12a Method and Application to the Relative Energies of Water Clusters $(H_2O)_{20}$. *J. Chem. Theory Comput.* **2014**, *10*, 1546−1553.

(37) Deev, V.; Collins, M. A. Approximate Ab Initio Energies by Systematic Molecular Fragmentation. *J. Chem. Phys.* **2005**, *122*, No. 154102.

(38) Mullin, J. M.; Roskop, L. B.; Pruitt, S. R.; Collins, M. A.; Gordon, M. S. Systematic Fragmentation Method and the Effective Fragment Potential: An Efficient Method for Capturing Molecular Energies. *J. Phys. Chem. A* **2009**, *113*, 10040−10049.

(39) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular Tailoring Approach for Geometry Optimization of Large Molecules: Energy Evaluation and Parallelization Strategies. *J. Chem. Phys.* **2006**, *125*, No. 104109.

(40) Isegawa, M.; Wang, B.; Truhlar, D. G. Electrostatically Embedded Molecular Tailoring Approach and Validation for Peptides. *J. Chem. Theory Comput.* **2013**, *9*, 1381−1393.

(41) Mayhall, N. J.; Raghavachari, K. Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.* **2011**, *7*, 1336−1343.

(42) Dong, H.; Hua, S.; Li, S. Understanding the Role of Intra- and Intermolecular Interactions inthe Formation of Single- and Double-Helical Structures of Aromatic Oligoamides: A Computational Study. *J. Phys. Chem. A* **2009**, *113*, 1335−1342.

(43) Yang, Z.; Hua, S.; Hua, W.; Li, S. Low-Lying Structures and Stabilities of Large Water Clusters: Investigation Based on the Combination of the AMOEBA Potential and Generalized Energy-Based Fragmentation Approach. *J. Phys. Chem. A* **2010**, *114*, 9253−9261.

(44) Yang, Z.; Hua, S.; Hua, W.; Li, S. Structures of Neutral and Protonated Water Clusters Confined in Predesigned Hosts: A Quantum Mechanical/Molecular Mechanical Study. *J. Phys. Chem. B* **2011**, *115*, 8249−8256.

(45) Hua, S.; Xu, L.; Li, W.; Li, S. Cooperativity in Long α- and 3_{10}-Helical Polyalanines: Both Electrostatic and van der Waals Interactions Are Essential. *J. Phys. Chem. B* **2011**, *115*, 11462−11469.

(46) Jiang, N.; Ma, J. Conformational Simulations of Aqueous Solvated α-Conotoxin GI and Its Single Disulfide Analogues Using a Polarizable Force Field Model. *J. Phys. Chem. A* **2008**, *112*, 9854−9867.

(47) Jiang, N.; Ma, J. Multi-Layer Coarse-Graining Polarization Model for Treating Electrostatic Interactions of Solvated α-Conotoxin Peptides. *J. Chem. Phys.* **2012**, *136*, No. 134105.

(48) Li, S.; Li, W.; Fang, T.; Ma, J.; Hua, W.; Hua, S.; Jiang, Y. *LSQC (low-scaling quantum chemistry) Program*, version 2.2; Nanjing University, Nanjing, 2012; see http://itcc.nju.edu.cn/lsqc.

(49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; et al.. *Gaussian 09*, revision B.01. Gaussian Inc.: Wallingford CT, 2009. See Supporting Information for complete reference.

(50) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549−2564.

(51) Ten-no, S. Initiation of Explicitly Correlated Slater-type Geminal Theory. *Chem. Phys. Lett.* **2004**, *398*, 56−61.

(52) Su, J. T.; Xu, X.; Goddard, W. A. Accurate Energies and Structures for Large Water Clusters Using the X3LYP Hybrid Density Functional. *J. Phys. Chem. A* **2004**, *108*, 10518−10526.

(53) Zhao, Y.; Truhlar, D. G. Design of Density Functionals That Are Broadly Accurate for Thermochemistry, Thermochemical Kinetics, and Nonbonded Interactions. *J. Phys. Chem. A* **2005**, *109*, 5656−5667.

(54) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.

(55) Wu, Q.; Yang, W. Empirical Correction to Density Functional Theory for van der Waals Interactions. *J. Chem. Phys.* **2002**, *116*, 515−524.

(56) Wu, Y.-D.; Zhao, Y.-L. A Theoretical Study on the Origin of Cooperativity in the Formation of 3_{10}- and α-Helices. *J. Am. Chem. Soc.* **2001**, *123*, 5313−5319.

(57) Flohil, J.; Vriend, G.; Berendsen, H. Completion and Refinement of 3-D Homology Models with Restricted Molecular Dynamics: Application to Targets 47, 58, and 111 in the CASP Modeling Competition and Posterior Analysis. *Proteins: Struct., Funct., Bioinf.* **2002**, *48*, 593−604.

(58) Jiang, H.; Maurizot, V.; Huc, I. Double versus Single Helical Structures of Oligopyridine Dicarboxamide Strands. Part 1: Effect of Oligomer Length. *Tetrahedron* **2004**, *60*, 10029−10038.

(59) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; 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., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.