JCTC Journal of Chemical Theory and Computation

Multilevel Fragment-Based Approach (MFBA): A Novel Hybrid Computational Method for the Study of Large Molecules

Jan Řezáč*,†,‡ and Dennis R. Salahub†

Department of Chemistry, IBI—Institute for Biocomplexity and Informatics and ISEEE—Institute for Sustainable Energy, Environment and Economy, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4, Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Systems, 166 10 Prague 6, Czech Republic

Received August 07, 2009

Abstract: We present a novel method for the calculation of large molecules and systems, the multilevel fragment-based approach. It is based on dividing the system into small fragments followed by separate calculations of these fragments and the interactions between them. Unlike previous fragmentation-based methods, we use multiple computational methods for the individual calculations. Using an accurate method only to calculate local interactions and more approximate methods for interactions over larger distances, it is possible to achieve results very close to a more demanding fragmented calculation using the higher level method only. The number of calculations performed at the higher level scales linearly with the size of the system, which significantly improves the efficiency and allows this scheme to be used for very large systems. In this work, we have combined density functional theory with the more approximate density functional tight binding method and applied this method to the calculation of model peptides. Formulation of first derivatives of the total energy within this fragmentation scheme is also presented and tested.

Introduction

One of the important trends in modern computational chemistry is the application of quantum-mechanical (QM) methods to molecules and molecular systems of increasing size while achieving the greatest accuracy possible. This is not an easy task, because of unfavorable scaling of these methods with the system size. Even the most efficient ab initio methods, Hartree−Fock (HF) and density functional theory (DFT), in their most efficient implementations have O(N^3) complexity. This algorithm complexity arises from a matrix diagonalization that is an indispensable part of all these calculations. Post-HF methods are even more demanding.

The ultimate goal is linear scaling. It is obvious that it can be achieved only by introducing some approximations, which in turn affect the accuracy of the results. Many approaches have been developed and published. All these methods can be divided into two very general groups: The first one is to set up the calculation of the whole system and then use approximations within the algorithms used to solve it (for a recent review of linear scaling methods, see refs 1, 2). The second group of methods relies on partitioning of the system into small fragments that are calculated separately and then composing the total energy and properties of the system from results obtained for the fragments. There is some overlap between these two groups, because some methods use the electronic structure of the fragments to build the electronic structure of the whole system, which is then used to obtain its properties. The adjustable density matrix assembler[3,4] (ADMA) and fragment molecular orbital[5−9] (FMO) methods are examples of this hybrid approach.

* Corresponding author. E-mail: rezac@uochb.cas.cz.

† University of Calgary.

‡ Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Systems.

10.1021/ct900413s © 2010 American Chemical Society
Published on Web 12/16/2009

In this work, we focus only on the pure fragment-based approach (FBA), exploiting the advantages of truly independent calculations of the fragments. Using only the final results of the calculations, such as the energy and its derivatives, this approach is independent of the method used and the calculations can be done using readily available software. This strategy also offers the possibility of lossless parallelization, because the calculations of the fragments can be distributed to multiple processors without any need of communication between them during the calculation.

There exist multiple methods based on the principle of fragmentation and independent calculation of the fragments: the molecular tailoring approach[10−13] (MTA) of Gadre et al., the molecular fractionation with conjugated caps[14−16] (MFCC), the kernel energy method[17,18] (KEM) devised by Huang, Massa and Karle, the generalized energy-based fragmentation (GEBF) by Li et al.[19] and others.[20−22] Most of these methods are designed to achieve linear scaling, and they do it by calculating interactions of a fragment with only a limited number of nearest surrounding fragments, selected either on the basis of their connectivity or by a distance cutoff. In large systems, the number of nearest neighbors of a fragment is practically constant and only kN interactions have to be calculated in a system consisting of N fragments. This incomplete description is an approximation that works well in systems where long-distance noncovalent interactions are negligible. This limits the use of these approaches to application to mostly neutral, nonpolar systems, because electrostatics is the strongest of these interactions. Noncovalent interactions play a crucial role in biomolecules, such as DNA and proteins, and their neglect can lead to serious errors. This issue will be explored systematically later in this paper.

This drawback has been addressed within the MFCC method by Li et al.,[14] who added all pairwise interactions between fragments into the scheme. The kernel energy method also uses all pairwise terms and was later extended to use higher-order terms up to four-body effects.[17]

In general, the linear scaling fragment-based methods rely on the calculation of a limited number of interactions between the fragments, neglecting the long-distance ones. Including all interactions is of course possible, but it brings quadratic scaling with the number of fragments. We present a novel approach that goes beyond the binary logic of selection of interaction and combines the advantage of calculating all pairwise terms with practically linear scaling. To do so, we combine multiple methods of calculation where they are appropriate. Only the local interactions are treated by the most accurate method, while all other interactions are also included but calculated at a more approximate and thus more efficient level. The scaling of the higher-level part of the calculation is linear, because we calculate only a limited (on average constant) number of interactions of each fragment with its neighbors. The lower-level part still scales quadratically, but its total cost is negligible compared to the higher-level part, because a more efficient method is used.

This multilevel fragment-based approach (MFBA) is in some respects related to previously reported MFCC calculations at the MP2 level,[14] where distinct distance cutoffs were used for the calculation of the correlation energy and the underlying HF calculations. On the other hand, the present MFBA scheme is, to the best of our knowledge, the first fragment-based method that combines completely independent methods, which opens up many new possibilities. The MFBA scheme is very general and its possible applications reach far beyond the single example described in this work. It can be applied to systems of moderate size, where we combine accurate ab initio calculations with more efficient ab initio calculations, semiempirical methods, or molecular mechanics. On the other end of the scale, we can combine more approximate atomistic methods, such as molecular mechanics, with a coarse-grained method used to calculate long-distance interactions (for example, as the interaction of multipoles representing the whole fragments). In this paper, we will focus on the first possibility, with the aim of achieving high accuracy by combining DFT with the more efficient self-consistent charge density functional based tight-binding (SCC-DFTB) method.[23]

The MFBA method is not limited to two levels of calculation. We can define more than one distance threshold and combine multiple methods in the calculation, utilizing different distance-dependent behavior of different types of interactions.

Our method of course shares some limitations with other fragment-based approaches. First, it can be applied only to molecules or systems that can be divided into fragments. The electronic structure of the fragments should represent the respective part of the molecule as closely as possible. It is not possible to cut a molecule with delocalized electrons, such as conjugated or aromatic systems. To limit the perturbation of the electronic structure, the fragments should be either independent molecules or connected by a chemical bond that can be cut and replaced by a suitable cap. We use hydrogen link atoms as the caps to treat C−C single bonds. Our scheme includes all pairwise interactions but neglects three-body and higher-order effects. Finally, apart from perturbation of the electronic structure of the molecule, adding link atoms to fragments introduces an additional error due to interaction of these atoms with the rest of the system. This problem has not been discussed before, and we will provide a detailed analysis and suggestions for how to minimize it.

The use of fragment-based methods is not limited to the calculation of the energy. Schemes for the calculation of first[13,14,24] and second[24,25] derivatives or electrostatic properties[26−28] have been proposed in the literature. We have formulated the calculation of first derivatives of the MFBA energy, which allows one to use the method for geometry optimizations. The simple scheme of composing the gradients reported for other fragment-based methods does not work in the MFBA scheme, and a more elaborate solution had to be developed.

Finally, we would like to discuss the relationship between MFBA and QM/MM or QM/QM′ methods.[29,30] Both approaches have in common the use of different computational methods combined to take advantage of the accuracy of the higher-level method and the efficiency of the other, and both have their specific uses. QM/MM methods are useful when

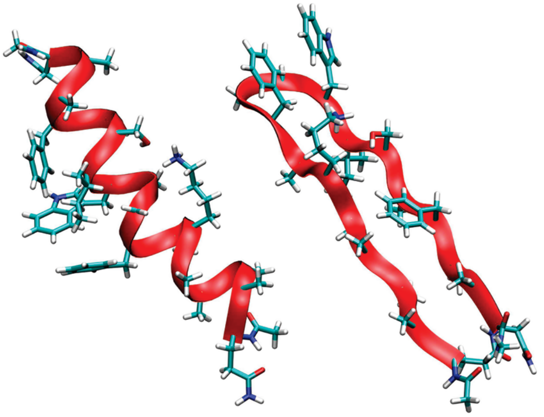

Figure 1. Model 18-peptide in α-helix and β-sheet conformations. The sequence QAAAAKAFGGWISAFAAN features various noncovalent interactions between the side chains.

there is a localized region of interest, while the rest of the system acts only as an environment. MFBA, on the other hand, aims to achieve the accuracy of the high-level method for the whole system.

In the past, several hybrid schemes using fragment-based QM calculation have been proposed. It is worth mentioning that fragment-based calculations can be used as the QM part in a QM/MM calculation,[31] or it can be used to replace some terms in the MM forcefield.[32] Also, an equivalent of a QM/QM′ scheme has been set up directly within one FMO calculation by separating the fragments into layers treated by different methods.[33]

The accuracy of the MFBA calculations and its comparison to a scheme neglecting long-distance interaction are demonstrated on a model system designed to be challenging. To make our conclusions relevant to applications in biochemistry, we have chosen a peptide in which electrostatic forces are important. We systematically study different states of the peptide, ranging from neutral to zwitterion forms, as well as an artificial hydrocarbon structure mimicking the same peptide. For each state, we calculate the energy difference between α-helix and β-sheet conformations (Figure 1). Although the MFBA approach can be applied to much larger molecules, the peptide was chosen to be relatively small (250 atoms), so that a full DFT calculation is still feasible and is used for comparison.

Methods

MFBA Formalism. Our fragmentation scheme is based on dividing the system into N fragments followed by calculations of these fragments and all of their pairs. In analogy with the kernel energy method, the total energy of the system is then expressed as a sum of energies of the fragments E_i and pairwise interaction energies ΔE_{ij} between them

$$E = \sum_{i=1}^{N} E_i + \sum_{i=1}^{N} \sum_{j=1}^{i-1} \Delta E_{ij} \qquad (1)$$

where interaction energy is defined as the difference between the energy of a pair of fragments E_{ij} and the energies of the isolated fragments

$$\Delta E_{ij} = E_{ij} - E_i - E_j \qquad (2)$$

This formulation is straightforward for noncovalent pairs, but the same can be applied to pairs of fragments connected by a covalent bond. To construct the fragments, the original chemical bond is replaced by a cap. For reasons described later, we are limited to the smallest caps possible, and we systematically use hydrogen atoms. This is the same approach as using hydrogen link atoms in QM/MM to terminate covalent bonds at the QM−MM boundary. In pairs (dimers) of fragments, the original bond between the fragments is conserved, and in the final summation, the added caps and the contribution of the new C−H bonds cancel out, leaving the contribution of the original C−C bonds that are conserved in the dimers.

In the MFBA method, the level of calculation for each fragment is decided using the following rules: All covalent dimers are calculated at the high level in order to describe the bond between the fragments accurately. For the rest of the pairs, the level is selected according to the distance of the fragments, evaluated as a minimal distance of their atoms. Pairs with distance below an arbitrary cutoff are calculated at the higher level, and other pairs are calculated at the lower level. Multiple cutoffs can be introduced if more than two methods are used for the calculations. The cutoff distance should be adjusted to the methods used, and it also represents the parameter balancing accuracy and efficiency. One important goal of this study is to show how the results change when the cutoff is varied in systems of different types.

Caps. The way that the system is divided into fragments may affect the accuracy of the results in multiple ways. To introduce as little perturbation as possible, we break only single C−C bonds, which are nonpolar, and replace them with C−H bonds, conserving well the original electronic structure of the fragment. The length of the newly formed C−H bond is calculated from the length of the C−C bond it replaces by multiplying it by a ratio of their average equilibrium distances. This value is not critical, since the contribution of these bonds cancels out in the summation.

Regarding only the electronic structure, the use of larger functional groups for the caps would be beneficial, but there is another reason why we use only hydrogen atoms. The added caps of course interact both with the rest of the fragment and with the other fragments. The major part of these interactions cancels out in the summation, but some interactions between the caps themselves do not and are an inherent source of error in this type of fragmentation schemes. Detailed analysis of an example is provided in the Appendix. To minimize this error, we have to use caps that interact as little as possible, hydrogen atoms being the best solution.

Gradient Calculation. The existing implementations[13,14] of gradient calculations in fragmentation-based methods compose the final gradients from gradients on the respective atoms in fragments using a scheme analogous to the energy calculation; the gradients on the caps are discarded. This may work in schemes where large caps are used and the original bonds at the boundary of each fragment are conserved, but it does not work when single atoms are used as the caps.

First, we compose the gradients on all but the cap atoms using a formula analogous to eq 1. Then, we go through all cap atoms in all fragments and apply the following correction to the atom in the original bond (with the appropriate sign from eqs 1 and 2). Since the position of the cap atom $\mathbf{r}_c$ is determined by the position of the atoms in the bond it replaces (atom in the fragment $\mathbf{r}_f$, the original atom it replaces $\mathbf{r}_o$) and the scaling factor g,

$$\mathbf{r}_c = \mathbf{r}_f + g(\mathbf{r}_o - \mathbf{r}_f) \qquad (3)$$

the gradient on the cap atom has to be projected on those two atoms:

$$\frac{\partial E}{\partial r_f} = \frac{\partial E}{\partial r_f} + (1 - g)\frac{\partial E}{\partial r_c} \qquad (4)$$

$$\frac{\partial E}{\partial r_o} = \frac{\partial E}{\partial r_o} + g\frac{\partial E}{\partial r_c} \qquad (5)$$

These equations are analogous to the ones used in subtractive QM/MM schemes.[34] This treatment works not only with MFBA, but it can be also applied to obtain gradients in the kernel energy methods, as it can be considered as a special case of MFBA calculation (with infinite cutoff distance for the high-level calculation of interactions).

In the current implementation, the selection of the methods is done once prior to the calculation. The use of this simplification in geometry optimization can be justified only when we do not expect substantial changes of the geometry. Updating the lists of fragment interactions during the calculation was not tested but is in principle possible. However, it might introduce problems caused by discontinuities of the potential energy surface.

Fragmentation. In this paper, we are working with peptides. Due to their polymeric nature, it is easy to break them into fragments. There is only one C−C bond in the peptidic backbone, the bond between C and C^{α}. To automate the fragmentation, we have developed a tool that identifies these bonds in a PDB file and prepares the input for fragmented calculation.

Implementation. For practical use, the MFBA calculation has to be automated. The implementation itself should be independent of the methods used for the particular calculation in order to take advantage of combining different methods. Our implementation is written in a high-level, object oriented programming language, Ruby.[35] It uses the Cuby (Chemistry in Ruby) framework, developed recently by one of the authors (J.R.). The framework provides unified access to external software packages used for the calculations and allows convenient manipulation of the results. Multiple simulation protocols, including geometry optimization and molecular dynamics algorithms, are available in Cuby and can be used with MFBA calculations. The resulting code is easy to use, as it needs only the geometry of the complete system and a simple definition of fragments (in the form of a list of bonds that are to be cut) as input. Currently, the following software packages and programs can be used for the calculations: AMBER,[36] deMon,[37] DFTB+,[38] Gaussian 03,[39] MNDO99,[40] MOPAC,[41] and Turbomole.[42]

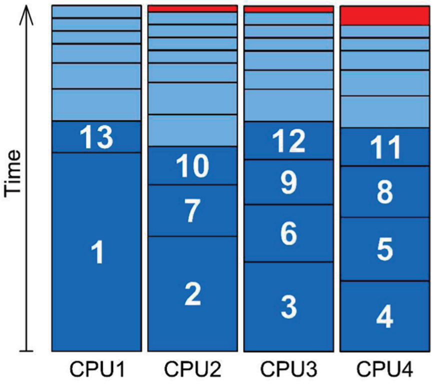

Figure 2. Calculations of fragments distributed to four processors using sorted queue. High-level calculations (darker blue, numbered) run first, faster low-level calculations fill the remaining time. This procedure minimizes the CPU time wasted while waiting for the last calculation (red).

Parallelization. The fragmented calculation can be made very efficient by simultaneous calculation of the fragments and their interactions. We have implemented a parallelization strategy that distributes these calculations to multiple nodes in a cluster and multiple processors on these nodes. The communication between the master node and other nodes, the initiation of the calculation, and retrieving the results use the widely available SSH protocol. When the calculation is run on a single multiprocessor computer, multiple threads are used to run the calculations in parallel, one at each processor.

To minimize the time of waiting for the last calculation to finish, we have implemented a simple queue system used to run the calculations on a given numbers of processors. This approach has been used previously in fragment-based methodology.[13] However, distributing the calculation to multiple nodes can lead to a situation where everything but the last calculation is finished and most allocated nodes have to wait for it. To minimize this waiting time, the calculation can be sorted by their expected length, as demonstrated previously for the FMO method.[43] In the MFBA scheme, multiple methods with different efficiency are used, but all the calculations can be put into one common pool because they are fully independent. The calculations in the pool are then sorted by the expected time of calculation, which is approximated from the method used and the number of atoms in the systems, and passed to a queue system balancing the load on multiple processors. The sorting ensures that the longest calculation will start first, and the shortest ones will then fill the remaining time with as little wasted time as possible (see Figure 2). This is particularly efficient in the case of the MFBA scheme, where there is only a limited number of demanding calculations.

Methods and Software Used. In this paper, we present MFBA calculations combining DFT and SCC-DFTB[23] calculations. All DFT calculations and the DFT part of MFBA calculations were carried out with the Turbomole package,[42] using the TPSS functional[44] and the TZVP basis set.[45] The resolution of identity approximation[46,47] was used to accelerate these calculations. In the case of the zwitterionic

Table 1. Energies of Model Structures Calculated Using DFT without and with Fragmentation and the Energy Difference between α-Helix and β-Sheet Conformations of the Peptide

		E (au)		Δ*E*(α−β) (kcal/mol)	
model	conformation	full	fragmented	full	fragmented
hydrocarbon	α	−4456.902	−4456.906		
	β	−4456.961	−4456.969	36.9	39.6
neutral	α	−6192.697	−6192.708		
	β	−6192.711	−6192.719	8.7	7.0
capped	α	−6193.111	−6193.145		
	β	−6193.098	−6193.027	−8.5	−11.2
zwitterion	α	−6020.733	−6020.782		
	β	−6020.903	−6020.950	106.9	105.2

peptide, a level shift of 0.5 au was used to improve convergence of the calculations.

The SCC-DFTB calculations were done in the DFTB+ code,[38] using the parameters set "mio". For the geometry optimization of model structures, empirical dispersion[48] available in the DFTB+ code was used (method abbreviated as SCC-DFTB-D).

Model Structures. The basis of all our test molecules is an 18-peptide of the sequence QAAAAKAFGGWISAFA-AN. It has net charge +1 due to the presence of a lysine (K) residue. Two conformers of this peptide were build using the program Ribosome,[49] an α-helix and an antiparallel β-sheet with the two glycine residues in the middle forming a β-turn. The orientation of the side chains, especially the aromatic groups, was adjusted manually to ensure favorable interactions. From this starting structure, four derivatives were prepared and their geometries optimized using the SCC-DFTB-D method.

The first structure is the peptide itself in zwitterionic form that would exist in solution. This serves as a model of a molecule with very strong intramolecular electrostatic interactions.

The second model is the same peptide with acetyl and N-methyl caps. The ends of the peptide chain are thus neutral, and the only charged group in the molecule is the lysine side chain. This model presents a structure where there are important electrostatic interactions between this charge and the other neutral, but polar, residues.

The third model is build from the second one by removing the proton from lysine, making it neutral, to further attenuate the effect of long-distance electrostatics.

Finally, we need a nonpolar model where there will be only a very small contribution of electrostatics. It was prepared from the previous model by replacement of all the peptide bonds (−CO−NH−) with the *trans*-alkene analog, −CH=CH−. The result is a hydrocarbon with functional groups originating from amino acid side chains. The optimized structures differ slightly from the geometry of the original peptide, but the overall conformation (helix and straight, parallel strands) is well conserved, the hydrogen bonds of the peptide replaced by weaker dispersion interactions. Interaction of the side chains stabilizes the structure further.

These four models are denoted zwitterion, capped, neutral, and hydrocarbon in the following text.

Performance of geometry optimizations was tested on a smaller model, a hexapeptide of the sequence FAGGAF with acetyl and N-methyl caps in the α-helix conformation. The structure was prepared analogously to the large model.

Results

Comparison to Full Calculations. First, let us compare fragment-based calculations to full DFT ones. Here, the fragmented calculation uses only one level, the same DFT method as in the full calculation, and all pairwise interactions are included in the scheme. The results (Table 1) show the accuracy of the fragment-based approach itself. The error in relative energies caused by the fragmentation reaches 2.7 kcal/mol in the hydrocarbon and capped models. Errors in absolute energies are substantially larger, but the major part of them is systematic and is eliminated by working with the relative energies only. These errors have multiple sources: most important are the missing three-body terms; other sources of the error are differences in electronic structure of the fragments and the complete molecule, interactions between the caps, and the lack of higher-order terms. These problems arise from the approximate nature of the method, and their further investigation is beyond the scope of this paper.

For the following discussion of MFBA results, the fragmented DFT calculations will serve as the benchmark, because they represent the upper limit of its accuracy, because no interaction are calculated at the lower level.

Neglected Interactions and MFBA. The most important parameter of both MFBA and the linear scaling scheme neglecting long-distance interactions is the cutoff distance above which the interaction energy is calculated by the more approximate method or not calculated at all. To show the advantage of MFBA over neglect of the interactions and to determine the optimal value of the cutoff for further calculations, we have performed a set of calculations with varying cutoff on each model. The results are summarized in Table 2 and plotted in Figure 3. In addition to the final energy differences, the number of interactions calculated at the high level (average for both conformers) is listed, because it is an important measure of the efficiency of the method.

It is clear that the MFBA energies converge much faster to the benchmark fragmented DFT value. In all four cases, the results are a good approximation of the benchmark energy even when the cutoff is set to zero, which implies that only the covalent pairs are calculated at the high level.

When the interactions above the cutoff distance are neglected, reasonable results can be achieved only with very

Table 2. Energy Difference between Conformers as a Function of Cutoff Distance and Number of Pairs Calculated at the High Level, Calculated by MFBA and Fragmentation with Neglected Long-Distance Interactions[a]

		$\Delta E(\alpha-\beta)$ (kcal/mol)				number of pairs			
method	cutoff (Å)	hydrocarbon	neutral	capped	zwitterion	hydrocarbon	neutral	capped	zwitterion
frag. DFT	–	39.6	7.0	−11.2	105.2	171	171	171	153
neglected	0	26.9	89.0	73.6	27.4	18	18	18	17
	2	34.4	56.7	62.5	117.5	20	31	24	22
	3	41.1	43.7	19.7	129.7	49	47	47	42
	4	39.5	27.0	1.7	118.3	70	65	66	60
	5	39.6	27.5	2.3	112.9	79	70	70	65
	8	39.5	12.8	−7.9	138.0	110	99	101	92
	12	39.5	8.7	−10.5	114.2	138	125	127	116
	16	39.6	7.6	−11.2	116.7	159	146	149	135
	20	39.6	7.2	−11.0	117.0	167	161	162	147
MFBA	0	34.4	2.6	−9.6	114.5	18	18	18	17
	2	32.8	0.3	−9.8	105.3	20	31	24	22
	3	40.7	9.1	−10.0	106.2	49	47	47	42
	4	39.4	8.7	−8.9	106.9	70	65	66	60
	5	39.7	9.1	−9.3	106.1	79	70	70	65
	8	39.5	7.5	−10.9	105.2	110	99	101	92
	12	39.6	7.1	−11.2	105.4	138	125	127	116
	16	39.6	7.1	−11.2	104.4	159	146	149	135
	20	39.6	7.1	−11.2	102.7	167	161	162	147

[a] The results of fragmented DFT calculation including all the interactions are provided as a reference.

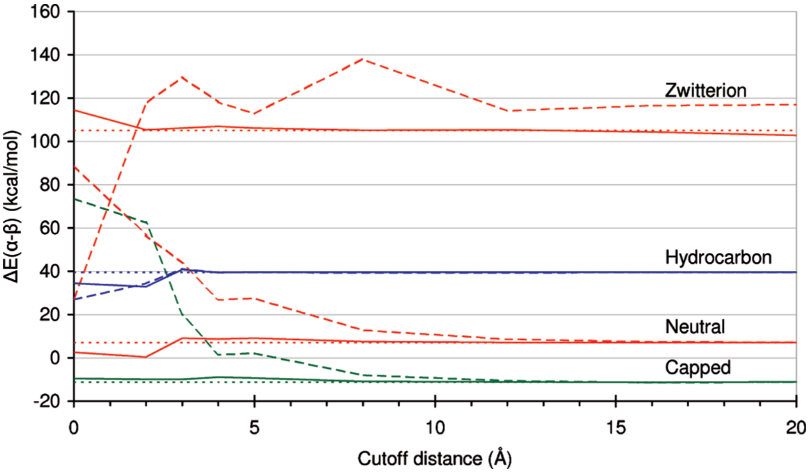

Figure 3. Energy difference between conformers as a function of cutoff distance. For each model (zwitterions, orange; hydrocarbon, blue; neutral, red; capped, green), the MFBA calculations (solid line) are compared to fragmentation scheme neglecting long-distance interactions (dashed line). The benchmark value, fragmented DFT calculation including all the pairwise interaction, is shown as dots.

large cutoff values. This behavior is more pronounced in the polar and charged systems, while in the nonpolar hydrocarbon model, the results are similar to the MFBA values, because the long-range interactions here are very weak. In the zwitterionic model, it is impossible to get good agreement with the benchmark energy when some of the interactions are neglected.

Generally, a cutoff of 5 Å is enough for consistent agreement with fragmented DFT calculation. This cutoff corresponds to about four calculations of pairs per fragment on average. Above 8 Å, the MFBA results are within 0.1 kcal/mol from the benchmark value.

Geometry Optimization. This work introduces the procedure for the calculation of derivatives of the total energy in the MFBA scheme and related methods, which allows one to apply this class of fragment-based methods to geometry optimizations.

We compare three variants of the method with full DFT geometry optimization: Fragmented DFT calculation (the KEM method), MFBA calculation combining DFT and SCC-DFTB, and the linear scaling scheme neglecting the long-distance interactions. The model system, an α-helical hexapeptide, is smaller than in the case of energy calculations. The peptide was divided into six fragments using the procedure described above. To put the method in a test, we use a small cutoff of 3 Å to select the interaction calculated at the higher level, which means that all pairs of fragments not connected by a covalent or hydrogen bond are calculated at the lower level or neglected. The discussion of the cutoff presented above suggests that using this cutoff distance should clearly reveal the differences between the methods.

The following convergence criteria were used: energy difference in the optimization step <0.006 kcal/mol, largest element of the gradient vector <1.2 kcal/mol/Å, and absolute

Table 3. Comparison of Results of MFBA Geometry Optimization[a]

		MFBA		
	full DFT	DFT	DFTB	neglect int
rmsd (Å)	–	0.04	0.06	0.10
ΔE (kcal/mol)	–	0.06	0.07	0.30
chain end–end distance (Å)	9.79	9.86	9.90	10.00
ALA2–ALA5 side chain distance (Å)	6.12	6.16	6.21	6.18
PHE1–PHE6 side chain distance (Å)	14.24	14.28	14.33	14.35
H-bonds in helix dist (Å)	7.17	7.25	7.29	7.38

[a] All noncovalent interactions in the model hexapeptide were calculated using the method listed in the table. Resulting geometries are compared to the geometry obtained by full DFT calculation by root mean square deviation (rmsd) of the Cartesian coordinates and by the energy difference calculated at the DFT level. Selected intramolecular distances are listed for illustration.

value of the gradient vector <1.2 kcal/mol/Å. The energy decreased smoothly during the optimizations. All four geometries obtained were very similar and the original structure of the peptide was conserved. The quality of the geometries was evaluated in two ways (the results are summarized in Table 3): First, we calculated the root-mean-square deviation (rmsd) between the DFT geometry and the one obtained by the fragment-based method (minimizing rmsd by translation and rotation of the geometries). Second, a full DFT calculation was performed on all the geometries and the energy is compared to that of the DFT geometry. Both the fragmented DFT calculation and MFBA combining DFT with SCC-DFTB yield results very close (0.04 and 0.06 Å, 0.06 and 0.07 kcal/mol) to the reference values. When the noncovalent interactions are neglected, the agreement is worse: rmsd 0.10 Å, energy difference 0.30 kcal/mol.

Conclusions

We have developed a novel fragmentation-based method combining multiple levels of calculations to maximize the accuracy and efficiency of calculations of large molecular systems: the multilevel fragment-based approach (MFBA).

This approach is not limited to the DFT/SCC-DFTB combination presented in this paper. While this combination allows very accurate calculations to be performed on molecules and systems containing hundreds of atoms, the same scheme can be used with both more accurate and more approximate methods.

The MFBA scheme is not limited to two levels of calculation. Multiple distance cutoffs can be used to select from multiple methods for each interaction.

The MFBA method was implemented in our modular framework, Cuby, which allows one to combine different methods. External software is employed to perform the calculations, multiple commercial or free computational packages and programs are supported now and other interfaces can be added easily.

Fragment-based calculations have an important advantage of efficient parallelization, because independent calculations of fragments can be distributed to multiple computers or processors. Our implementation allows this parallelization and uses an internal queue system and sorting of the calculations to use the allocated computational resources as efficiently as possible.

The performance of the MFBA method was demonstrated on model molecules derived from an 18-amino acid peptide. MFBA is superior to the fragmented calculation neglecting long-distance interactions while almost the same scaling and efficiency is conserved. MFBA works well even for charged systems where the scheme neglecting long-distance interactions fails. The accuracy of MFBA calculations is very close to the fragmented calculation using the higher level only that served as a benchmark for our calculations. However, further improvements are required to improve the absolute accuracy of the fragmentation-based method when compared to a full calculation.

We have derived the expressions for the calculation of gradients in the MFBA scheme, which opens applications of the method in geometry optimization and molecular dynamics. The same scheme can be applied to the kernel energy method.

Geometry optimization of a model hexapeptide has shown that even very approximate MFBA calculation can yield a geometry almost identical to that of a fragmented DFT calculation, and both these geometries are very close to the one from full DFT optimization. The agreement is better than in the case of energy calculations and makes MFBA an interesting choice for geometry optimizations of larger molecules.

Acknowledgment. We are grateful to NSERC, Canada for ongoing support. This work is a part of Research Project Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and was supported by Grant LC 512 from the Ministry of Education, Youth and Sports of the Czech Republic.

Appendix

Error Introduced by Interaction of the Caps. The addition of the caps to the fragments is a source of error not only because they may not represent exactly the original electronic structure of the fragment, but also because they interact with each other and with the fragments themselves. A minor part of these interactions is not canceled in the final summation.

To demonstrate this issue, we will analyze the simplest system that can be treated with this fragmentation scheme, a linear system consisting of three fragments. The original system, the capped fragments, and their pairs are schematically depicted in Figure 4. Here, we identify all the interactions between the caps themselves and between the caps and the fragments and between the fragments in the pairs and express the energy of each of these subsystems as the sum of these interactions ΔE_{xy}, energies of the fragments without the caps $E_{f(x)}$ and energies of the caps alone $E_{c(x)}$.

The total summation (eq 1) for this system can be simplified using eq 2 to the following sum of the subsystems (the capped fragments and their pairs):

$$E = E_{AB} + E_{BC} + E_{AC} - E_A - E_B - E_C \qquad (6)$$

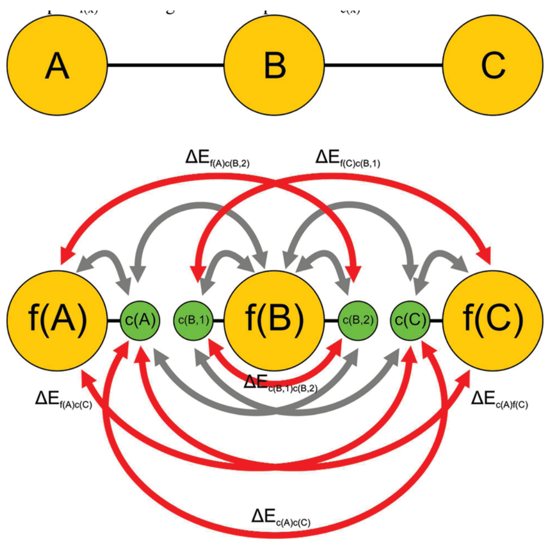

Figure 4. Schematic representation of a system consisting of three fragments, before and after introduction of the caps. Interactions between the caps and other caps or fragments that are present in the calculation are shown, and those that are present in the final energy expression are highlighted red and labeled.

By substituting the subsystem energies by the sums of the contributions outlined above and simplifying the expression, we obtain the total energy as

$$E = E_{\mathrm{f(A)}} + E_{\mathrm{f(B)}} + E_{\mathrm{f(C)}} + \Delta E_{\mathrm{f(A)f(B)}} + \Delta E_{\mathrm{f(B)f(C)}} + \Delta E_{\mathrm{f(A)f(C)}} - \Delta E_{\mathrm{c(B,1)c(B,2)}} + \Delta E_{\mathrm{c(A)c(C)}} + \Delta E_{\mathrm{f(A)c(C)}} + \Delta E_{\mathrm{c(C)f(A)}} + \Delta E_{\mathrm{f(A)c(B,2)}} + \Delta E_{\mathrm{f(C)c(B,1)}} \quad (7)$$

while the desired total energy is only the sum of energies of the fragments and their interactions

$$E = E_{\mathrm{f(A)}} + E_{\mathrm{f(B)}} + E_{\mathrm{f(C)}} + \Delta E_{\mathrm{f(A)f(B)}} + \Delta E_{\mathrm{f(B)f(C)}} + \Delta E_{\mathrm{f(A)f(C)}} \quad (8)$$

and the remaining terms (eq 7 minus eq 8) are the artifacts introduced by interactions of the caps.

For this reason, the nature of the caps should be chosen in a way that these interactions are minimized. We use the smallest possible caps, hydrogen atoms. Another factor that affects this error is the size of the fragments, which determines also the distance between the caps. When larger fragments are used, this error will not be as pronounced as when the caps are close to each other and to the other fragment in a pair.

References

(1) Saebo, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213–236.

(2) Wu, S. Y.; Jayanthi, C. S. *Phys. Rep.* **2002**, *358*, 1–74.

(3) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2002**, *106*, 11791–11800.

(4) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980–1986.

(5) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483–2490.

(6) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2004**, *389*, 129–134.

(7) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832–6840.

(8) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

(9) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2002**, *351*, 475–480.

(10) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484–495.

(11) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. *J. Phys. Chem.* **1994**, *98*, 9165–9169.

(12) Gadre, S. R.; Ganesh, V. *J. Theor. Comput. Chem.* **2006**, *5*, 835–855.

(13) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.

(14) Li, S. H.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215–7226.

(15) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599–3605.

(16) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039–12041.

(17) Huang, L.; Massa, L.; Karle, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1849–1854.

(18) Huang, L. L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2005**, *103*, 808–817.

(19) Li, W.; Li, S. H.; Jiang, Y. S. *J. Phys. Chem. A* **2007**, *111*, 2193–2199.

(20) Collins, M. A.; Deev, V. A *J. Chem. Phys.* **2006**, *125*, 104104.

(21) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.

(22) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777–8785.

(23) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(24) Hua, W. J.; Fang, T.; Li, W.; Yu, J. G.; Li, S. H. *J. Phys. Chem. A* **2008**, *112*, 10864–10872.

(25) Rahalkar, A. P.; Ganesh, V.; Gadre, S. R. *J. Chem. Phys.* **2008**, *129*, 234101.

(26) Gao, A. M.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. K. *Chem. Phys. Lett.* **2004**, *394*, 293–297.

(27) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. *J. Phys. Chem.* **1994**, *98*, 9165–9169.

(28) Mueller, T. J.; Roy, S.; Zhao, W.; Maass, A.; Reith, D. *Fluid Phase Equilib.* **2008**, *274*, 27–35.

(29) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185–199.

(30) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(31) Li, H.; Li, W.; Li, S. H.; Ma, J. *J. Phys. Chem. B* **2008**, *112*, 7061–7070.

(32) Soderhjelm, P.; Ryde, U. *J. Phys. Chem. A* **2009**, *113*, 617–627.

(33) Fedorov, D. G.; Ishida, T.; Kitaura, K. *J. Phys. Chem. A* **2005**, *109*, 2638–2646.

(34) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *462*, 1–21.

(35) Ruby programming language. www.ruby-lang.org (accessed Aug 15, 2009).

(36) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(37) Köster, A. M.; Calaminici, P.; Casida, M. E.; Flores-Moreno, R.; Geudtner, G.; Goursot, A.; Heine, T.; Ipatov, A.; Janetzko, F.; del Campo, J. M.; Patchkovskij, S.; Reveles, J. U.; Salahub, D. R.; Vela, A. *deMon2k*; 2006 (http://www.demon-software.com/public_html/program.html).

(38) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J.; 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.; Bakken, V.; 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 C.02*; 2004.

(40) Thiel, W. *MND099*; Max-Planck-Institut für Kohlenforschung: Mülheim, Germany, 1999.

(41) Stewart, J. P. *MOPAC2009*; Stewart Computational Chemistry: Colorado Springs, CO, 2008.

(42) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(43) Fedorov, D. G.; Olson, R. M.; Kitaura, K.; Gordon, M. S.; Koseki, S. *J. Comput. Chem.* **2004**, *25*, 872–880.

(44) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(45) Schafer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(46) Eichkorn, K.; Treutler, O.; Ohm, H.; Haser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

(47) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

(48) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

(49) Srinivasan, R. *Ribosome*; John Hopkins University: Baltimore, MD, 1997.

CT900413S