JCTC Journal of Chemical Theory and Computation

Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide

Ernesto Suárez, Natalia Díaz, and Dimas Suárez*

Departamento de Química Física y Analítica, Universidad de Oviedo, 33006 Oviedo (Asturias), Spain

Received November 17, 2008

Abstract: Herein, we first review different methodologies that have been proposed for computing the quantum mechanical (QM) energy and other molecular properties of large systems through a linear combination of subsystem (fragment) energies, which can be computed using conventional QM packages. Particularly, we emphasize the similarities among the different methods that can be considered as variants of the multibody expansion technique. Nevertheless, on the basis of thermochemical arguments, we propose yet another variant of the fragment energy methods, which could be useful for, and readily applicable to, biomolecules using either QM or hybrid quantum mechanical/molecular mechanics methods. The proposed computational scheme is applied to investigate the stability of a triple-helical collagen model peptide. To better address the actual applicability of the fragment QM method and to properly compare with experimental data, we compute average energies by carrying out single-point fragment QM calculations on structures generated by a classical molecular dynamics simulation. The QM calculations are done using a density functional level of theory combined with an implicit solvent model. Other free-energy terms such as attractive dispersion interactions or thermal contributions are included using molecular mechanics. The importance of correcting both the intermolecular and intramolecular basis set superposition error (BSSE) in the QM calculations is also discussed in detail. On the basis of the favorable comparison of our fragment-based energies with experimental data and former theoretical results, we conclude that the fragment QM energy strategy could be an interesting addition to the multimethod toolbox for biomolecular simulations in order to investigate those situations (e.g., interactions with metal clusters) that are beyond the range of applicability of common molecular mechanics methods.

Introduction

The idea of representing the total energy of a large molecule as a combination of fragment energies has been considered for decades. To better appreciate their similarities and differences, we will first review several computational approaches for combining fragment energies that have been developed during recent years. We note, however, that other linear-scaling methodologies[1,2] aimed at construction of the full density matrix of a large system from the fragment density submatrices are beyond the scope of this paper. Thus, we will discuss first the methods based on the multibody expansion approach and other closely related methods that include implicitly high-order many-body effects into fragment energies using various approximations. We will also comment on the so-called kernel energy method that turns out to be essentially a multibody expansion method. Subsequently, we will review other methods that approximate the quantum mechanical energy of large systems by combining fragment energies on the basis of intuitive and/or thermochemical argumentations. Although we will see that these thermochemically based protocols can be considered as truncated forms of the more general multibody expansion method, they are conceptually simpler and can be readily

* Corresponding author phone: +34-985103689; fax: +34-985103125; e-mail: dimas@uniovi.es.

10.1021/ct8005002 CCC: $40.75
Published on Web 04/29/2009

applicable using many computational tools at a moderate computational cost. In fact, we will formulate yet another variant of the thermochemical fragment energy methods that could be particularly useful to compute the energies of large biomolecular systems. Finally, as a real case application of the proposed method, we will combine fragment-based quantum chemical energies with molecular mechanics and standard quantum chemical calculations in order to compute the relative free energy of the triple-helical form of a collagen model peptide with respect to its monomer state.

Multibody Expansion Method. The so-called cluster expansion method[3] has been developed in the framework of solid-state chemistry in order to represent the total energy of an atomic crystal as a linear combination of the characteristic energies of clusters of atoms over a fixed lattice. The coefficients in the cluster expansion are computed using quantum mechanical energy calculations of a few prototype structures. However, the so-constructed functions are not transferable, i.e., they cannot be used for each conceivable configuration of the system. Subsequently, the multibody expansion (MBE) method, also called *N*-body potentials, or otherwise, cluster potentials, has been developed as a more refined version of the cluster expansion technique.[4] The MBE method evaluates the total energy as a summation of energies corresponding to isolated atomic clusters extracted from the global structure so that they include systematically two-, three-, and *N*-body effects. More recently, it has been demonstrated that the MBE approach can be generalized for an *arbitrary* system, whose energy can be uniquely evaluated using series of structure-independent, perfectly transferable, many-body potentials.[5] In this general MBE formalism, the total energy of an *M*-particle system (composed of atoms, molecules, or molecular fragments linked covalently) can be expressed as $E_M(A_1, A_2, ..., A_M)$, where $A_i = \{\mathbf{R}_i, \sigma_i\}$ has the information about the coordinates ($\mathbf{R}_i$) and the type (σ_i) of the i particle. Since the ordering of the M particles is arbitrary, the functional form of E_M must be such that E_M is invariant to any permutation $A_i \leftrightarrow A_j$.

Representing the total energy by an expansion of a series of *N*-order (or *N*-body or *N*-fragment) energy contributions $E^{(N)}$, we have

$$E_M(A_1, A_2, ..., A_M) = \sum_{N=1}^{M} E^{(N)}(A_1, A_2, ..., A_M) \quad (1)$$

where, in turn, the $E^{(N)}$ terms can be computed from a multiple summation of *N*-order interaction potentials

$$E^{(N)} = \sum_{m_1<...<m_N}^{M} V^{(N)}(A_{m_1}, A_{m_2}, ..., A_{m_N}) \quad (2)$$

where the sum $\sum_{m_1<...<m_N}^{M} V^{(N)}$ runs over all possible combinations $\{m_1, ..., m_N\} \in \{1, ..., M\}$.

Note that eqs 1 and 2 express the total energy E in terms of *N*-order potentials. In practice, however, one needs to compute the $V^{(N)}$ potentials from energy calculations performed on different subsystems. The general relationship between $V^{(N)}$ and subsystem energies can be obtained through a Möbius inversion as defined in number theory.[5] The general result is

$$V^{(N)}(A_1, A_2, ..., A_N) = \sum_{L=1}^{N} (-1)^{N-L} \sum_{m_1<...<m_L}^{N} E(A_{m_1}, A_{m_2}, ..., A_{m_L}) \quad (3)$$

In the above equation, $E(A_{m_1}, A_{m_2}, ..., A_{m_L})$ stands for the energy of a cluster composed by L fragments labeled by the $(m_1, m_2, ..., m_L)$ indices. In fact, eq 3 constitutes a unique definition of the *N*-order interaction potential $V^{(N)}$, which is structure independent because this equation does not carry any information about the environment of the subsystems.[5] The actual significance of eq 3 can be more easily grasped by deriving the first terms of the *N*-order expansion leading to the total energy. Thus, the sum of the first-order potentials is just the sum of the energies of the isolated fragments

$$E^{(1)} = \sum_{m_1=1}^{M} V^{(1)}(A_{m_1}) = \sum_{m_1=1}^{M} E(A_{m_1}) \quad (4)$$

For the second-order contribution, which can be interpreted as the *excess* energy due to pair interactions, we obtain

$$E^{(2)} = \sum_{m_1<m_2}^{M} V^{(2)}(A_{m_1}, A_{m_2}) = \sum_{m_1<m_2}^{M} [E(A_{m_1}, A_{m_2}) - E(A_{m_1}) - E(A_{m_2})] \quad (5)$$

and, of course, $E_M \approx E^{(1)} + E^{(2)}$ defines the well-known pairwise additive approximation to the total energy. Analogously, the three-body $E^{(3)}$ contribution, which collects the $V^{(3)}$ potentials, is the additional energy due to three-body effects, and that cannot be assessed from a two-body representation

$$V^{(3)} = \sum_{m_1<m_2<m_3}^{M} [E(A_{m_1}, A_{m_2}, A_{m_3}) - E(A_{m_1}) - E(A_{m_2}) - E(A_{m_3}) - V^{(2)}(A_{m_1}, A_{m_2}) - V^{(2)}(A_{m_1}, A_{m_3}) - V^{(2)}(A_{m_2}, A_{m_3})] \quad (6)$$

Finally, it may be interesting to note that the MBE equations can be rewritten in terms of the so-called mutual information functions (MIFs),[6] which have been used to compute the configurational entropy of flexible molecules. Thus, the MIF expansion approaches the full-dimensional configurational probability distribution by including systematically *N*-order correlations among the internal degrees of freedom; likewise, the successive $V^{(N)}$ potentials include the *N*-order effects on the total energy. Similarly, the energy of a system composed of M arbitrary fragments can be expanded using the MIFs in the following form

$$E_M(A_1, A_2, ..., A_M) = \sum_{i=1}^{M} E(A_i) - \sum_{m_1<m_2}^{M} I_2(A_{m_1}, A_{m_2}) + ... + (-1)^{N-1} \sum_{m_1<...<m_N}^{M} I_N(A_{m_1}, ..., A_{m_N}) \quad (7)$$

where the mutual information function $I_N(A_{m_1}, ..., A_{m_N})$ combines the energies of all the clusters formed by N fragments

$$I_N(A_{m_1}, ..., A_{m_N}) = \sum_{L=1}^{N} (-1)^{L+1} \sum_{m_1<...<m_L}^{N} E(A_{m_1}, ..., A_{m_L}) \quad (8)$$

Note that the mathematical form of the MBE and MIF expressions are identical due to the fact that $(-1)^{N-L} \equiv (-1)^{N+L}$.

Kernel Energy Method is an MBE Method. At this point, it is convenient to simplify the notation used in the MBE equations by replacing $E(A_{m_1}, A_{m_2}, ..., A_{m_L})$ (the energy of the subsystem with L fragments) with $E_{ijk...}$ (the energy of the subsystem composed of the *i, j, k, ...* particles or fragments). In this way, the pairwise additive approximation for a system composed of a total of M fragments can be written as

$$E_M = \sum_{i=1}^{M} E_i + \sum_{i=1}^{M} \sum_{j=i+1}^{M} (E_{ij} - E_i - E_j) \quad (9)$$

In recent years, the so-called kernel energy method (KEM) has been utilized to compute the quantum mechanical (QM) energy of large biomolecules[7−11] by representing a full molecule by smaller *kernels* of atoms (i.e., fragments A_i). The majority of the KEM applications that have been reported to date compute the total energy "by summation over the energy contributions of all *double* kernels reduced by those of any *single* kernels, which have been overcounted in the sum over double kernels",[8] that is, by means of the following expression

$$E_M = \sum_{m=1}^{M-1} \left(\sum_{i=1}^{M-m} E_{i,i+m} \right) - (M - 2) \sum_{i=1}^{M} E_i \quad (10)$$

However, it can be easily demonstrated (see Supporting Information) that the original KEM energy formula is equivalent to the MBE pairwise additive approximation.

Several KEM applications on biomolecules have been reported in which the dangling bonds of the molecular fragments are saturated with hydrogen atoms before carrying out the corresponding fragment energy calculations. However, the presence of the H-link atoms introduces an error in the computation of the total energy given that the validity of the MBE equations requires that only the actual fragments are considered in the calculations. Nevertheless, if the fragments are large enough and the total number of fragments is relatively low, the associated error can be reasonably small. Of course, the H-link error can be further reduced by including higher order MBE terms given that these terms progressively account for the environment of each fragment by considering larger and larger clusters of fragments. This has been done in a recent article in which the KEM equation is expanded up to fourth order[11] through a cumbersome derivation that follows an MBE recipe employed in a former study of water clusters.[12]

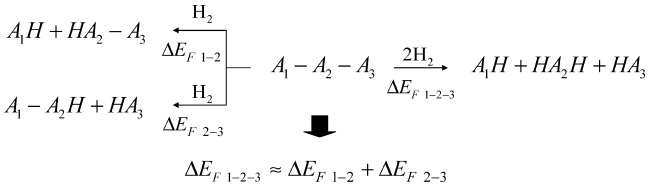
Electrostatically Embedded MBE Methods. In principle, the pairwise additive approximation defined by eq 9 is not enough to accurately compute the total energy of complex systems. Unfortunately, the calculation of higher order MBE terms is extremely expensive in terms of computer time. In order to overcome the limitations of second-order methodologies at a reasonable computational cost, some authors proposed to compute the energies of the individual fragments (E_i) and fragment pairs (E_{ij}) taking into account the electrostatic field of the rest of the system.[13−18] For example, in the fragment molecular orbital (FMO) method, the energies of the different fragments are computed by iteratively solving *effective* fragment Hamiltonians that include the electrostatic effects from the electrons in the surrounding $(M - 1)$ fragments as well as from all nuclei in the total molecule.[14,19] The resulting FMO energies are then combined using MBE equations of order 2 or 3 to derive the total energy. A similar alternative for noncovalently connected fragments is the electrostatically embedded many-body expansion (EE-MBE).[16−18] The energy of each cluster is calculated in the presence of the electric field due to the fixed partial atomic charges of the surrounding fragments. A significant improvement in the electrostatically embedded second- and third-order energies for a series of water clusters is found when compared with the results of standard MBE calculations.[16]

Molecular Tailoring Approach. The so-called molecular tailoring approach (MTA)[20] divides the total system into *overlapping* fragments and subsequently estimates the total energy by summing the fragment contributions and then subtracting the energies of fragment *intersections*. This means that interactions between nonoverlapping fragments are neglected in the MTA method and that each fragment intersection formally accounts for N-body effects to the total energy, with N being the number of overlapping fragments at the particular intersection. This strategy is somehow equivalent to employing localized multibody expansions, and therefore, the MTA approach can be considered as a *flexible* MBE method. The MTA method can also compute one-electron properties of the full system by combining the fragment density matrices into a single density matrix for the whole system.[21]

Molecular Fractionation with Conjugate Caps. The so-called molecular fractionation with conjugate caps (MFCC) scheme also estimates the total energy of large systems from calculations performed on fragments. The MFCC method was originally designed to compute the QM interaction energy between a protein and a small ligand,[22] but this method has been expanded to predict the total energy of protein molecules.[23] In this approach, the protein is divided into fragments $A_i = (-C_\alpha HR_i-CO-N_{i+1}H-)$, with R_i being the side chain of the i amino acid residue and N_{i+1} is the backbone N atom of the $(i + 1)$ amino acid. Instead of H-link atoms, two "conjugate caps", NH_2- and $-C_\alpha H_2R_{i+1}$, are placed at the corresponding $C_{\alpha,i}/N_{i+1}$ atoms to saturate the exposed valence sites of each fragment A_i. The total energy of an M-residue protein molecule is first approximated by summing the energies of the (capped) fragments and then subtracting the energies of the $NH_2-C_\alpha H_2R_{i+1}$ conjugate caps. This first-order approximation is then corrected ad hoc by adding a second-order term ($\delta E^{(2)}$) that accounts for the pairwise interaction energy between non-neighboring fragments. The final MFCC expression is

$$E_M = [E(A_1 - C_\alpha H_2R_2) + \sum_{i=2}^{M-1} E(NH_2 - A_i - C_\alpha H_2R_{i+1}) +$$

Scheme 1



$$E(\mathrm{NH_2} - A_M)] - [\sum_{i=1}^{M-1} E(\mathrm{NH_2} - \mathrm{C_\alpha H_2 R}_{i+1})] + \delta E^{(2)} \quad (11)$$

To compute the $\delta E^{(2)}$ contribution, the fragments are capped with H-link atoms as in the KEM scheme. Alternatively, another variant of the MFCC method has been proposed that uses only fragment energies, which are computed in the presence of the electrostatic field created by point charges representing the non-neighboring residues.[24]

Systematic Molecular Fragmentation. As we will see later, the MFCC expression[11] can be justified by means of simple thermochemical arguments on the basis of formal fragmentation processes of the protein system. In fact, the thermochemical approach for computing the fragment-based energy of large molecules has already been explored systematically by Collins et al.[25] The basic reasoning behind the generalization proposed by Collins et al. is summarized in Scheme 1, which shows a generic molecular system composed of three fragments ($A_1-A_2-A_3$) that can be formally broken through three different fragmentation processes.

The key approximation in the protocol of Collins et al. is that the reaction energy for the total fragmentation of $A_1-A_2-A_3$ (ΔE_{F1-2-3}) is estimated as the sum of the reaction energies corresponding to the two single-fragmentation processes (i.e., $\Delta E_{F1-2} + \Delta E_{F2-3}$). The straightforward consequence of this approximation is that the energy of the total system can be expressed as a combination of the energies of the three smaller subsystems

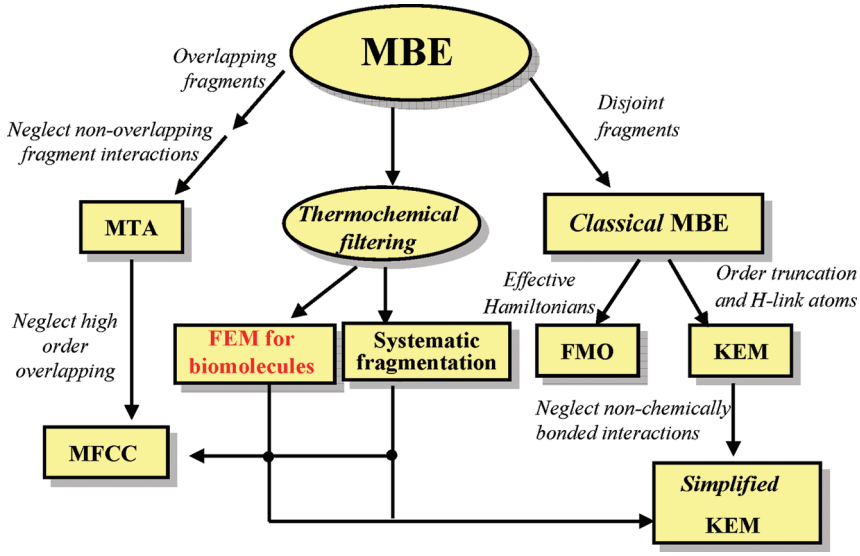
$$E_{123} = E_{12} + E_{23} - E_2 \quad (12)$$

In principle, Collins et al. employ both chemical topology and computer cost considerations in order to choose the best site at which a large molecule is cut so that the resulting A_2 fragment is (a) large enough to reasonably neglect the interaction between A_1 and A_3 and (b) simultaneously small enough to compute the energy of the A_1-A_2H fragment using high-level QM methods. If the accompanying HA_2-A_3 fragment is too large, the fragmentation protocol defined in Scheme 1 is then applied iteratively until all the produced fragments can be described quantum mechanically. Ultimately, this thermochemical approach results in the total energy being approximated by a linear combination of fragment energies, whose precise form depends on the nature of the chemical system and on the chemical topology and computer cost considerations. Like in the MFCC method, the systematic fragmentation technique can be augmented with a nonbonded energy correction by computing the interaction energy between two nonchemically bonded fragments if their separation is below a certain threshold.[25]

Comparison of the Different Methods. Although largely unnoticed in some of the previous works, the MBE formalism provides the general framework for developing computational strategies aimed at the evaluation of the total energy of large systems from subsystem (fragment) energies (see Scheme 2). Thus, the FMO method, the various KEM formulas, and the MFCC expression with pairwise interactions can be classified as MBE techniques that include *N*-body effects through fragment energy calculations. Similarly, the systematic fragmentation method of Collins et al. can be generated directly from the MBE expansion by neglecting all the MBE interaction potentials beyond second order and using an additional chemical topology criterion to neglect a large number of second-order contributions. We can also see in Scheme 2 that inclusion of the H-link atoms to cap the exposed valence sites of the fragments extracted from a covalent system makes the Collins' fragmentation method nearly identical to the simplified version of the KEM method in which only the chemically bonded *double kernels* are considered.[8] Thus, once a fragmentation scheme has been applied, the same energy terms are actually computed in the two methods. Similarly, the systematic fragmentation proposed by Collins et al. encompasses the effective MFCC in which only fragment energies are considered. On the other hand, the MFCC method can be considered as a particular case of the MTA formalism given that the MFCC-capped fragments are equivalent to the MTA overlapping fragments and the MFCC conjugate caps would correspond to fragment intersections in the MTA approach. However, while the MFCC fragments are built to make simple overlaps (i.e., each atom can only be part of one or two fragments), the MTA method admits more complex fragment overlaps among *N* fragments. These and other interrelationships show that in general fragment energy methods assume a similar *ansatz*.

Goals of the Present Work. In principle, the ability to perform on a routine basis fragment energy calculations on large biomolecules could be very useful to predict their energetic properties using high-level QM methodologies. Fortunately, previous test applications have shown that high-order MBE contributions contain many more energetic terms than those that are actually required to derive the total energy from fragment energies within a reasonable accuracy. In this way and taking into account that proteins and nucleic acids are linear polymers that exhibit many repetitive secondary structural motifs, we believe that a thermochemical approach complemented with a distance-based criterion is probably the best option to formulate a linear scaling fragment-based energy method for biological molecules. This approach, which can be considered as a thermochemical truncation of the multibody expansion, is also computationally advantageous because the required energetic terms can be easily computed using standard methodologies. Another advantage of the thermochemical framework is that the successive fragmentation energies involved in the formal degradation of the biomolecule can be computed taking into account the effect of a solvent continuum in the QM Hamiltonian. Thus, in this work, assuming a simple fragmentation process, we will derive a fragment energy formula for estimating the total energy of a biomolecule as function of a cutoff criterion. On one hand, we will show that our fragment energy method (FEM in Scheme 2) can have a broader applicability

Scheme 2



than the MFCC equation, which in turn can be derived from our approach as a particular case. On the other hand, with respect to the more general thermochemical scheme of Collins et al., our expression will be more readily applicable to (and limited to) large biomolecular systems in which a natural choice for the formal fragmentation processes can be easily made. In addition, more emphasis will be placed upon the consistent use of a cutoff criterion in the fragment energy calculations, the inclusion of solvent effects, the mixing of QM and molecular mechanical calculations, and the potential implementation of the fragment-based energy methods within the context of QM/MM methodologies.

Theory

For the sake of simplicity, we will consider a macromolecule ***P*** that is a linear chain of M fragments A_i interconnected through covalent bonds ($A_1-A_2-\ldots-A_M$). For example, if ***P*** is a protein, A_i could be a single amino acid or a secondary structure element. We do note, however, that the same equations based on fragment energies would result for more complex topological patterns connecting the A_i fragments like in cyclic or branched macromolecules.

The total fragmentation of ***P*** can be achieved through the following formal reaction

$$A_1 - A_2 - \ldots A_M + \sum_{i=1}^{M-1}(X_i - Y_i) \rightarrow A_1 - X_1 + \sum_{i=2}^{M-1}(Y_{i-1} - A_i - X_i) + Y_{M-1} - A_M \quad \text{(R1)}$$

Note that every fragment linkage in the ***P*** molecule is broken through insertion of a specific X_i-Y_i molecule(s) into the A_i-A_{i+1} bond. If ***P*** is not a linear chain, then X_i and Y_i would stand for all the molecular caps that are required to saturate the exposed bonds after having removed the A_i fragment from the rest of the ***P*** molecule. In any case, the total energy change corresponding to the above formal reaction is

$$\Delta E = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + E(Y_{M-1} - A_M) - \sum_{i=1}^{M-1} E(X_i - Y_i) - E(\mathbf{P}) \quad (13)$$

The thermochemical approximation to compute ΔE can be introduced as follows: we compute first the reaction energy for the fragmentation step in which the A_1 fragment is removed. However, we assume that the reactants involved in the first fragmentation process are subsystems of ***P*** that are defined on the basis of some geometric and/or chemical-structure criterion. The same criterion, denoted onward as the R criterion, should be applied consistently along the ***P*** backbone structure. Perhaps the simplest criterion for defining the reactants could be to impose a layer cutoff around the leaving A_1 fragment, but other choices like sequence proximity could be used. Thus, assuming that a well-defined R criterion is used, the first fragmentation reaction can be written as

$$A_1 - B_1^R + X_1 - Y_1 \rightarrow A_1 - X_1 + Y_1 - B_1^R \quad \text{(R2)}$$

where B_1^R represents a buffer region, which includes all the neighboring atoms (or fragments A_i) that are around A_1 in the ***P*** structure depending on the R criterion being used. Similarly, the fragmentation process for the A_i-A_{i+1} bond can be represented by the following chemical equation

$$Y_{i-1} - A_i - B_i^R + X_i - Y_i \rightarrow Y_{i-1} - A_i - X_i + Y_i - B_i^R \quad \text{(R3)}$$

where the closer atoms or fragments around A_i excepting those in A_{i-1}, A_{i-2}, ..., A_1 are included in the buffer B_i^R. The sum of the $M-1$ fragmentation processes defined in this manner leads to the following chemical equation

$$A_1 - B_1^R + \sum_{i=2}^{M-1}(Y_{i-1} - A_i - B_i^R) + \sum_{i=1}^{M-1}(X_i - Y_i) \rightarrow A_1 - X_1 + \sum_{i=2}^{M-1}(Y_{i-1} - A_i - X_i) + \sum_{i=1}^{M-1}(Y_i - B_i^R) \quad \text{(R4)}$$

In this way, the energy change for the total fragmentation of $\boldsymbol{P}$ through the R-dependent fragmentation processes (ΔE^R) is given by

$$\Delta E^R = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + \sum_{i=1}^{M-1} E(Y_i - B_i^R) - [E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + \sum_{i=1}^{M-1} E(X_i - Y_i)] \quad (14)$$

Extracting the exact fragmentation energy ΔE from eq 13 and defining $\delta E = \Delta E^R - \Delta E$, we can combine eqs 13 and 14 in order to exactly express the total energy of the system $E(\boldsymbol{P})$ in terms of the fragment energies and the δE difference

$$E(P) = \left[E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + E(Y_{M-1} - A_M)\right] - \left[\sum_{i=1}^{M-1} E(Y_i - B_i^R)\right] + \delta E(\mathbf{B}^R, \mathbf{Y}) \quad (15)$$

where the δE difference is expressed as a function of $\mathbf{B}^R = \{B_i^R\}$ and $\mathbf{Y} = \{Y_i\}$. This is a consequence of the fact that $E(\mathbf{P})$ is rigorously independent of $\mathbf{B}^R$, $\mathbf{X} = \{X_i\}$, and $\mathbf{Y}$ and that the terms in the square brackets are independent of $\mathbf{X}$ (i.e., the identity of the X_i moieties is irrelevant).

For practical applications of the thermochemical fragment energy eq 15, the δE term must be neglected. To increase the accuracy of the fragment-based energy calculations, one straightforward solution would be to systematically increase the R criterion in order to include larger portions of the remaining $\boldsymbol{P}$ molecule in the B_i^R buffer regions until reaching a reasonable compromise between accuracy and computational cost. The best systems for which we can efficiently apply this simple strategy would be *linear* structures like carbon nanotubes, DNA segments, collagen molecules, etc. Of course, in the case of more compact systems like globular proteins, a larger computational cost and a lower accuracy can be expected for the same R criterion because the buffer regions would contain many more atoms and truncation effects would be more important. However, we could also use the well-known QM/MM methodologies in order to calculate the reaction energies of the fragmentation steps using the same settings as those that are typically employed in routine QM/MM calculations. In this case, the R criterion would be applied to select the size of the QM region while the rest of the system would be treated classically. Thus, like in the electrostatically embedded variants of the MBE methodologies, we expect that QM/MM calculations of fragmentation energies could account for high-order effects within the thermochemical approach.

As above mentioned, we can particularize the general eq 15 to obtain the MFCC equation for a protein system. This can be done by matching Y_i by $-NH_2$ and B_i^R by $-R_{i+1}C_\alpha H_2$, which are the "conjugate caps" adopted in the MFCC scheme. In our thermochemical terminology, these choices are equivalent to consider X_i-NH_2 as the capping dimers as well as to adopt a minimum sequence proximity R criterion for defining the B_i^R groups. Then eq 15 becomes

$$E(\mathbf{P}) = [E(A_1 - C_\alpha H_2 R_2) + \sum_{i=2}^{M-1} E(NH_2 - A_i - C_\alpha H_2 R_{i+1}) + E(NH_2 - A_M)] - [\sum_{i=1}^{M-1} E(NH_2 - C_\alpha H_2 R_{i+1})] + \delta E \quad (16)$$

If we compare this equation with eq 11, we see that the "non-neighboring interactions" ($\delta E^{(2)}$) in the MFCC approach[23] constitutes an approximation to the actual error (δE) committed in the calculation of the global fragmentation energy. We note in passing that the same energy contributions collected in eq 16 can be associated to other formal fragmentation processes by changing accordingly the definition of the A_i fragments and the corresponding conjugated caps. For example, expression 16 also results if the A_i fragment corresponds to the i residue and $Y_i = H$.

Finally, it may be interesting to note that our approach, like with all the MBE-like methods, computes the total energy as a linear combination of fragment energies. As gradient is a linear operator, its application over the fragment energy expression would be straightforward as previously noticed in other works.[20,25] In this way, both energy and gradient values for the total system could be obtained from fragment calculations using similar approximations and techniques as those typically used by the QM/MM methodologies.[26,27]

Results and Discussion

In many of the previous works, the viability of fragment-based energy methods has been assessed by means of proof of principle applications, that is, by carrying out single-point calculations and using relatively low QM levels of theory. However, most of the biomolecules are flexible molecular systems in aqueous solution, and therefore, in actual applications, structures for performing fragment-based QM calculations should be provided by Monte Carlo or molecular dynamics (MD) simulations using either explicit or implicit solvent models. In this respect, we think that classical MD simulations still constitute the most reasonable alternative to generate the biomolecular structures for the subsequent fragment QM calculations. This approach would be similar to the molecular mechanics Poisson−Boltzmann method,[28] which predicts mean values of free energies of biomolecules in solution as estimated over a series of representative snapshots extracted from classical MD simulations. Moreover, we also note that various levels of approximation could be required in the fragment energy calculations. For example, a standard density functional level of theory combined with an implicit solvent model can take into account both the intramolecular electronic effects and the solute−solvent electrostatic interactions. Other free-energy terms such as attractive dispersion interactions or thermal contributions could be calculated using molecular mechanics (MM). We believe that this and other technical issues like the counterpoise correction of the basis set superposition error (BSSE) in the QM calculations should be explicitly considered in the test calculations in order to assess the actual performance of the fragment QM energy calculations in the context of multimethod approaches to simulating biomolecules. There-

Scheme 3

$$\begin{array}{lcccc} \alpha\text{–chain} & \text{Pro}-\text{Hyp}-\text{Gly}-\text{Pro}- & (\text{Hyp}-\text{Gly}-\text{Pro} & -\text{Hyp}-\text{Gly} \\ \beta\text{–chain} & \text{Pro}-\text{Hyp}-\text{Gly}- & \text{Pro}-\text{Hyp}-\text{Gly} & -\text{Pro}-\text{Hyp}-\text{Gly} \\ \gamma\text{–chain} & \text{Pro}-\text{Hyp}- & \text{Gly}-\text{Pro}-\text{Hyp})_{i=2,9} & -\text{Gly-Pro}-\text{Hyp}-\text{Gly} \\ & \underbrace{\qquad\qquad}_{A_1} & \underbrace{\qquad\qquad\qquad}_{A_i} & \underbrace{\qquad\qquad}_{A_{10}} \end{array}$$

fore, we decided to reexamine in this work the problem of the stability of triple-helical collagen model peptides by combining our fragment energy expression with previous MD and MM data that have been reported by us recently.[29]

Many collagen model peptides with 30−45 amino acids have been synthesized to investigate the thermal stability and folding of the triple-helix domain of natural collagen. These peptides, which are also known as triple-helical peptides (THPs), assemble spontaneously to form a triple-helix complex that can be characterized using a wide array of experimental techniques.[30] The THP molecules present a characteristic triple-helix structure composed of three peptide chains, each in an extended, left-handed polyproline II-like helix, which are staggered by one residue and then supercoiled about a common axis in a right-handed manner. The close packing of the three chains requires the presence of a sterically small glycine residue at every third position. The test calculations reported in this work were performed on the prototypical $[(\text{Pro-Hyp-Gly})_{10}]_3$ system (labeled as **POG10**), which contains many proline and 4(*R*)-hydroxiproline (Hyp) residues that largely stabilize the triple-helix conformation.[31,32]

Selection of a Fragmentation Process. The collagen model for our test calculations, **POG10**, contains three peptide chains (labeled α, β, and γ) with 30 amino acids per chain. As mentioned above, the fragment energy expression, eq 15, that has been derived by assuming that the ***P*** macromolecule is a linear chain, is also applicable for more complex macromolecules like **POG10**. To this end, we describe the triple helix as a *linear* arrangement of 10 fragments comprising each of three *triplets* of residues from the α, β, and γ chains (see Scheme 3). The resulting building blocks or fragments A_i will be termed as *triplets*. A pair of consecutive *triplets*, A_i-A_j, is interconnected through three peptide linkages corresponding to the α, β, and γ chains. We chose this mode of partitioning because it minimizes the interactions between nonconsecutive *triplets* and maximizes the number of interactions among the three peptide chains within each *triplet*.

After having chosen a structurally and computationally convenient partitioning of **POG10**, we can define more precisely the formal fragmentation processes required for the fragment-energy calculations based on eq 15. More specifically, we see in Figure 1 how the terminating Y_i group attached to the *N*-terminal end of the A_i triplet comprises three acetyl groups for the α, β, and γ peptide chains, whose coordinates are extracted from the *C* end of the previous A_{i-1} triplet and augmented with the required H-link atoms. Similarly, the buffer group B_i^R attached to the *C* end of the A_i triplet includes the adjacent A_{i+1} triplet plus three *N*-methyl moieties extracted from the A_{i+2} fragment (this choice of B_i^R is equivalent to a ∼9 Å cutoff around the leaving A_i fragment). This formal fragmentation process can also be applied straightforwardly to obtain the energy of the individual peptide chains α, β, and γ. In this case, the corresponding A_i, B_i^R, and Y_i groups include residues located in the same chain.

Comparison between Conventional and Fragment-Based QM Energies. Before computing the energy of the full **POG10** system, we assessed the combined quality of the fragment energy calculations and the collagen partitioning in order to reproduce the energetic properties of a relatively large collagen subsystem. The size of the selected subsystem, $[\text{Ace-(Pro-Hyp-Gly)}_4\text{-Nme}]_3$ (456 atoms), still allowed us to carry out full QM calculations. Following similar prescriptions to those represented in Figure 1, four different fragments (A_i) can be distinguished in this model. We computed both the interaction energy among the three peptide chains and the absolute energy of the THP model. The calculations were performed on 25 structures that were built using the coordinates of the central region of POG10 extracted from MD snapshots (see Table S1 in the Supporting Information).[29] As described in the Computational Section, the energy calculations were carried out using a density functional level of theory (PBE/SVP) combined with the COSMO solvent model. The intramolecular dispersion energy is included via an empirical method. The BSSE arising from the interchain interactions is corrected using the standard counterpoise (CP) method. In the case of the fragment energy calculations, the CP correction was applied to the fragment electronic energies, that is, the electronic energies of the $A_1-B_1^R$, $Y_1-A_2-B_2^R$, ..., fragments extracted from one peptide chain (e.g., α) were computed in the presence of the ghost basis functions located in the equivalent fragments from the other two chains (e.g., β and γ). For the full QM calculations, the CP recipe was used to correct the BSSE of the electronic energies of the full peptide chains.

The total interaction energy of $[\text{Ace-(Pro-Hyp-Gly)}_4\text{-Nme}]_3$ can be estimated from the combination of five energy terms using eq 15 (see Table 1). Similarly, the energy of each Ace-(Pro-Hyp-Gly)$_4$-Nme peptide chain can be computed from the corresponding fragment energies. In this way, we derived an average interaction energy (ΔE_{int}) of -29.4 ± 0.2 kcal/mol that matches perfectly the *exact* value (-29.5 ± 0.2 kcal/mol) according to conventional QM calculations.

Since ΔE_{int} is a relative quantity, it can be expected that the fragment energy calculations would benefit from partial cancelation of errors. However, we see in Table 1 that the total energy *E* of the whole system in aqueous solution can be computed accurately using the fragment energies given that the error in the mean value of the fragment-based energies with respect to the exact full QM value is rather small, 0.0001 au (∼0.1 kcal/mol). Table S1 (Supporting Information) shows that small errors arise also in each of the individual structures considered in the calculations. We also see in Table 1 that the observed accuracy in the total energy benefits from a partial cancelation of errors in the computation of the individual energetic components, which result in energy differences of +1.4 (gas-phase energy) and −1.5 kcal/mol (solvation energy) between the fragment-based and the exact values. Although the accuracy in the gas-phase

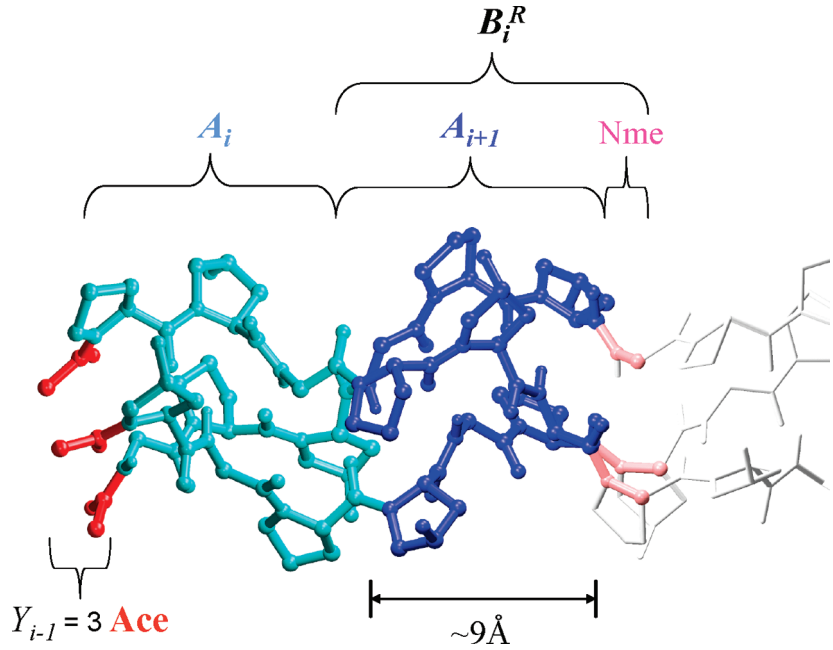


Figure 1. Ball and stick model of the **POG10** triple helix. The various moieties of **POG10** involved in the formal *i*-fragmentation step ($i \geq 2$) are shown in different colors. See text for details.

Table 1. Average Values and Standard Deviations of the Interchain Interaction Energies (ΔE_{int}, in kcal/mol of peptide) for the [Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ System[a]

	$A_1-B_1^R$	$Y_1-A_2-B_2^R$	$Y_2-A_3-B_3^R$	$Y_1-B_1^R$	$Y_2-B_2^R$	[Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ FRAG	[Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ CONV	$\Delta_{FRAG-CONV}$
$\Delta\bar{E}_{int}$	−14.9 ± 0.1	−15.1 ± 0.1	−15.2 ± 0.1	−7.8 ± 0.1	−7.9 ± 0.1	−29.4 ± 0.2	−29.5 ± 0.2	0.1
$\bar{E}$	−6329.7247 (0.0025)	−6329.7250 (0.0021)	−6329.7254 (0.0021)	−3536.8993 (0.0016)	−3536.8983 (0.0014)	−11915.3775 (0.0032)	−11915.3776 (0.0032)	0.1
$\bar{E}_{gas}$	−6329.5131 (0.0023)	−6329.5137 (0.0023)	−6329.5131 (0.0022)	−3536.7781 (0.0015)	−3536.7769 (0.0015)	−11914.9849 (0.0032)	−11914.9872 (0.0033)	1.4
$\Delta\bar{G}^{elec}_{COSMO}$	−87.1 (0.2)	−87.2 (0.3)	−87.7 (0.3)	−54.8 (0.2)	−55.4 (0.3)	−151.9 (0.4)	−150.4 (0.4)	−1.5
$\bar{E}_{disp}$	−105.3 (0.3)	−105.4 (0.4)	−105.3 (0.3)	−53.7 (0.2)	−53.8 (0.2)	−208.4 (0.5)	−208.5 (0.5)	0.1

[a] Average values and standard errors (in parentheses) of the various energy components for the THP fragments: total energy in solution, E, in au; gas-phase energy, E_{gas}, in au; electrostatic solvation energy, ΔG^{elec}_{COSMO}, in kcal/mol; and empirical dispersion energy, E_{dis}, in kcal/mol. Mean values of the total energies as obtained with the fragment-based (FRAG) and conventional (CONV) calculations and their differences ($\Delta_{FRAG-CONV}$, in kcal/mol) are also indicated.

Table 2. Average Values (kcal/mol of peptide) for the Different Energy Components of the Interaction Energy among the **POG10** Peptide Chains[a]

$\Delta\bar{E}^{CP\text{-}uncorrected}_{PBE/SVP}$	BSSE	$\Delta\bar{G}^{elec}_{COSMO}$	$\Delta\bar{E}_{disp}$	$\Delta\bar{E}_{int}$[b]
−105.6 (1.1)	85.7 (0.2)	37.1 (1.0)	−82.5 (0.1)	−65.4 (0.2)

[a] Standard errors are given in parentheses. [b] $\Delta\bar{E}_{int} = \Delta\bar{E}^{CP\text{-}uncorrected}_{PBE/SVP} + BSSE + \Delta\bar{G}^{elec}_{COSMO} + \Delta\bar{E}_{disp}$.

energy (~0.002 au) is comparable to that reported in previous fragment energy calculations,[10,14,20] these results suggest that inclusion of solvent effects in the fragment QM calculations should improve the accuracy of the fragment-based approaches given that the intramolecular long-range interactions could be dampened out by the electrostatic screening exerted by the surrounding solvent continuum.

Due to the linear structure of collagen, we expect that the performance of the fragment-energy calculations for larger collagen models would be equally satisfactory and that other molecular properties of collagen molecules (e.g., gradients) could be also computed within a reasonable accuracy. Finally, we note that, in terms of CPU time, a single-point energy calculation on the [Ace-(Pro-Hyp-Gly)$_4$-Nme]$_3$ system using the fragment approach took about 9 h on one x86−64 processor. The same energy value obtained with conventional QM calculations required about 80 h of CPU time.

Fragment Calculations on the POG10 Triple Helix. The results of our fragment energy calculations on the full **POG10** system (1089 atoms) are summarized in Table 2, which contains the average values of the various energetic components contributing to the interchain interaction energy. The calculations were done on 100 snapshots extracted from our previous MD simulation.[29] The total interaction energy amounts to −65.4 kcal/mol of peptide, which gives an average value of −6.5 kcal/mol for every −(Pro-Hyp-Gly)− triplet of residues. As expected, all the energy components considered in the calculations (gas-phase electronic energy, empirical dispersion energy, and electrostatic solvation

energy) contribute significantly to the interaction energy. Of particular interest can be the large weight of the BSSE as estimated by the CP calculations, 85.7 kcal/mol. Clearly, the omission of the BSSE corrections would have resulted in an unphysical overestimation of the interaction energy. On the other hand, the inability of the PBE DFT functional to recover most of the intermolecular dispersion energy justifies the addition of the empirical dispersion energy. In fact, the combination of DFT QM methods and empirical dispersion energy has been used in previous computational studies that apply DFT to study weak nonpolar interactions.[33−35] Although the three peptide chains intertwined into the triple helix establish many hydrogen-bond interactions that can be described reasonably by the PBE calculations, we see in Table 2 that the dispersion energy is the largest stabilizing contribution to the interchain interaction energy of the **POG10** triple helix. Hence, it turns out that the close packing of the peptide chains plays a crucial role in the overall stabilization of the triple helix.

Perhaps the bottom line from the calculations summarized in Table 2 is that the QM fragment energy approach may constitute a promising alternative for studying the intermolecular interactions in large biomolecules. For the collagen model peptide studied in this work, the error introduced by the fragmentation technique can be rather small (<1 kcal/mol) as suggested by the preliminary test calculations. However, we do note again that when using a DFT level of theory in the fragment calculations for large biomolecules, correction of the BSSE and inclusion of dispersion energy are a must in order to obtain meaningful results for interaction energies.

Intramolecular BSSE. As shown in Table 2, the CP correction to the interchain interaction energy is quite large, +85.7 kcal/mol at the PVE/SVP level, due to the large size of the **POG10** system and the relatively small size of the double-ζ SVP basis set. In principle, the use of larger basis sets should reduce significantly the magnitude of the BSSE but at the cost of increasing the CPU time. Nevertheless, it is most likely that assessing and correcting the BSSE will also be required when carrying out fragment energy calculations on biomolecules using medium-sized basis sets (cc-pVDZ, TZVP, ...). Moreover, it is becoming increasingly clear that the relative energies of different conformations of large and flexible biomolecules are quite sensitive to the size of the basis set and that part of this dependence arises from the *intramolecular* BSSE.[36] Although this (presumably small) effect has been commonly ignored so far, there is now some solid computational evidence in the recent literature indicating that the intramolecular BSSE can severely impair the accuracy of the energetic QM predictions for polypeptide systems.[36−38]

Given that we are interested in computing the relative stability of the triple-helix conformation with respect to the compact form of the isolated chains (see below), we decided to estimate the magnitude of the intramolecular BSSE in our QM calculations. For this purpose, the CP method of Boys and Bernardi could be applied by taking atomic fragments, but this alternative would result in a large number of extra QM calculations as well as in problems in the assignation of charge, multiplicity, and electronic state of the atomic fragments.[39] Hence, we followed a more pragmatic approach that consists of the definition of proper molecular fragments within the large system and adding H-link atoms to saturate the exposed chemical bonds. Subsequently, the BSSE in the interaction among the resulting fragments is computed using the standard CP procedure. A similar approach has been employed previously by other authors.[36] For example, Valdés et al. estimated the intramolecular BSSE in [*n*]-helicene molecules consisting of all-ortho-annulated benzene rings by computing the CP-corrected interaction energies of benzene pairs, in which the Cartesian coordinates of the C atoms are identical to those in the helicene.[36]

After some computational experimentation, we decided to employ the following fragmentation protocol for estimating the intramolecular BSSE of the **POG10** peptide chains. (1) For each **POG10** structure, a pair list of nonbonded (beyond 1−4) interactions involving heavy atoms is built using a distance criteria (X···Y < 4.0 Å). (2) Each peptide chain is broken into four smaller fragments by removing three glycine residues. These glycine residues are automatically selected in order to *maximize* the number of nonbonded interactions among the resulting fragments (see Figure 2a and 2b). H-Link atoms are added to saturate the exposed bonds. (3) The standard CP method is used to compute the value of the BSSE corresponding to the interactions among the four fragments (*intra*-$BSSE_1$; see Figure 2b). (4) The BSSE due to the interactions between the formerly removed glycine residues and the nearby groups is estimated by building a molecular cluster in which the three glycine residues are surrounded by the closer residues. Then the CP procedure is applied again to estimate the BSSE arising from the simultaneous interactions between the three glycines and the rest of the groups (*intra*-$BSSE_2$; see Figure 2c). (5) The total intramolecular BSSE of the peptide chain is approximated by adding together the two BSSE values computed in 3 and 4.

The QM calculations for estimating the intramolecular BSSE were done on 100 MD snapshots of the free **POG10** chain.[29] Thus, we found that, at the PBE/SVP level, the average value of the intramolecular BSSE for the isolated **POG10** chain in its folded state amounts to 92.7 kcal/mol of peptide, which is even greater than the BSSE related to the interchain interactions in the triple-helix state (85.7 kcal/mol). For the sake of consistency, the same protocol was applied on each of the three chains in the triple-helix conformation. In this case, the peptide chains are quite extended and their intramolecular BSSE is predicted to be only 3.1 kcal/mol on average. All these CP-corrected QM calculations can be combined to estimate the energetic penalty for the folded **POG10** chain to adopt its extended conformation in the triple helix, the average value being +30.8 kcal/mol (in terms of $E_{PBE/SVP} + BSSE_{intra} + E_{disp} + \Delta G^{elec}_{COSMO}$). Neglecting the intramolecular BSSE in the folded state of **POG10** would lead to a very large unrealistic value (∼120 kcal/mol) for the relative energy between the folded and the extended forms of the peptide chain.

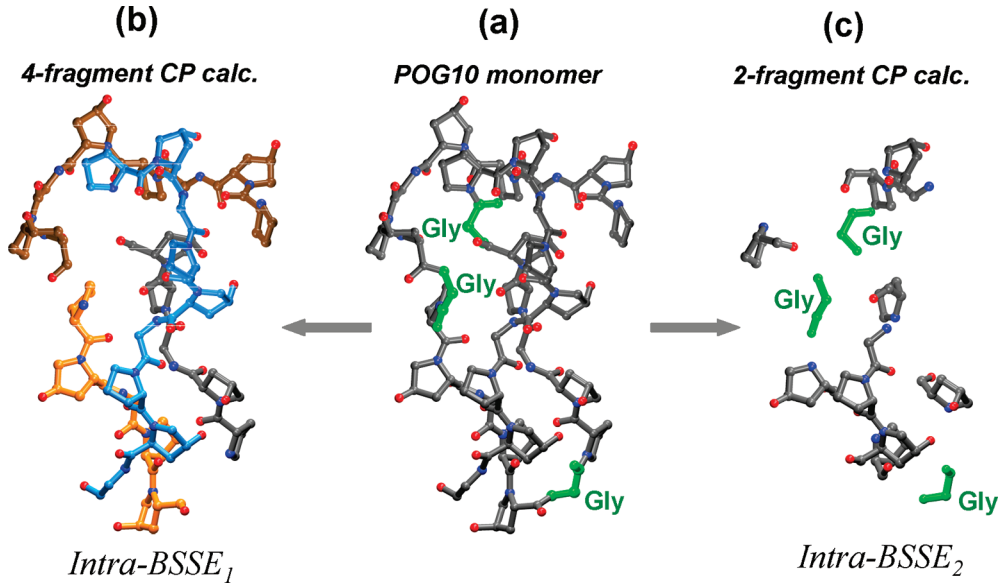


Figure 2. Ball-and-stick models of a **POG10** chain in its monomer state showing the fragmentation procedure followed to correct the *intramolecular* BSSE through CP calculations. (a) On the basis of a nonbonded interaction pair list, three glycine residues (in green) are selected in order to maximize the number of nonbonded interactions among the peptide fragments that result upon removal of the glycine residues. (b) BSSE arising from the interactions among the four peptide chains (C atoms are shown in different colors) is estimated using the CP procedure. (c) A molecular cluster is constructed from the coordinates of the glycine residues selected in a and those of the nearby peptide residues that interact directly with the marked glycines. The BSSE associated to the interaction between the glycines and the nearby groups is again estimated by means of CP calculations.

Table 3. Average Values and Standard Errors (in kcal/mol of peptide) of the Free-Energy Components for the Transition from the Monomeric to the Triple-Helix State at 300 K

	mean value	standard error		mean value	standard error
$\Delta\bar{E}_{PBE/SVP}$	53.7	*9.7*	$\Delta\bar{H}^{norm}_{MM\text{-}GBSA}$	0.7	*0.1*
$\Delta\bar{E}^{CP\text{-}corrected}_{PBE/SVP}$	49.8	*9.6*	$-T\Delta\bar{S}^{norm}_{MM\text{-}GBSA}$	8.8	*0.9*
$\Delta\Delta\bar{G}^{elec}_{COSMO}$	−73.2	*9.3*	$-T\Delta S_{conf}$	0.4	-
$\Delta\bar{E}^{solute}_{disp}$	−10.7	*0.4*	$\Delta\bar{G}^{CP\text{-}corrected\ a}$	−11.7	*1.8*
$\Delta\bar{E}^{solute-solvent}_{disp}$	11.0	*0.8*	$\Delta\bar{G}^{a}$	−7.8	*2.1*
$\Delta\bar{G}_{cav}$	−1.2	*0.1*			

[a] Assuming a standard state of 0.001 M.

Free Energy for the Transition from Monomer to Triple Helix. As shown above, the fragment QM calculations complemented with the empirical dispersion formula can give insight into the nature of the interactions holding the peptide chains in the triple-helix conformation. However, the actual stability of the triple helix is determined by the free-energy change for dissociation to give the free peptide monomers. In our previous work,[29] we found that the isolated **POG10** peptide in aqueous solution adopts a stable folded conformation, and therefore, by combining the fragment QM data on the triple helix with the results of QM calculations on a representative set of **POG10** monomers, one could estimate the corresponding free-energy change for the peptide aggregation process leading to the **POG10** triple helix, provided that the selected QM method gives a compensated description of the conformational and intermolecular interaction energies. By taking advantage of our previous computational experience, we combined the QM energies with further molecular-mechanical data in order to ensure a balanced description of other free-energy components (solute−solvent vdW interactions, thermal contributions to free energy, etc.). More specifically, we used the following expression in order to compute the average free energy of the **POG10** system both in its triple-helix and monomer states

$$\bar{G} = \bar{E}^{CP-corrected}_{PBE/SVP} + \bar{E}^{solute}_{disp} + \bar{E}^{solute-solvent}_{disp} + H^{norm}_{MM-GBSA} - T\bar{S}^{norm}_{MM-GBSA} + \Delta\bar{G}^{elec}_{COSMO} \quad (17)$$

where the gas-phase $\bar{E}^{CP\text{-}corrected}_{PBE/SVP}$ energy, which includes the intermolecular and intramolecular BSSE corrections, and the electrostatic solvation energy ($\Delta\bar{G}^{elec}_{COSMO}$) are computed by means of fragment-based (triple helix) and standard (monomer) QM calculations; the $\bar{E}^{solute}_{disp} + \bar{E}^{solute-solvent}_{disp}$ dispersion energy terms are computed with the same empirical formula, and normal mode molecular mechanical calculations are used to estimate the thermal contributions to free energy. The change in the average values of these energetic components for the monomer → triple-helix transition are collected in Table 3, which also includes the corresponding small differences in the cavitation free energy and the conformational entropy that were computed following the procedures described in our previous work.[29]

We see in Table 3 that the QM energy terms (gas-phase and solvation energy) as well as the empirical dispersion

energies change significantly on going from the monomer to the triple helix. In agreement with our previous molecular mechanical and Poisson−Boltzmann (MM-PB) calculations, the QM-based approach predicts also that the driving force for the formation of the triple helix is mainly provided by the electrostatic solvation energy. The total ΔG value obtained with the CP-corrected QM energies amounts to −11.7 kcal/mol, with a statistical uncertainty of 1.8 kcal/mol (standard error). This value is in moderate agreement with the most accurate experimental estimate at 300 K, −6.4 kcal/mol, which has been derived from differential scanning calorimetry.[29,40] The purely MM-PB calculations together with a broader sampling give a ΔG value of −6.2 (1.2) kcal/mol.[29] The larger difference between the QM-based calculations and experiment is most likely due to several factors like the small error in the fragment-based QM calculations, the remaining inaccuracy in the correction of the intramolecular BSSE, slight unbalances in the combination of QM and MM data in eq 17, as well as by some limitations of the PBE DFT functional to reproduce the electrostatic and H-bond interactions. All these potential sources of error, which are not present in the MM-PB calculations, could be mitigated by gaining more computational experience and improving the details of the mixed QM−MM computational protocol. On the other hand, it turns out that the ΔG value obtained with the CP-uncorrected QM energies (−7.8 kcal/mol) is closer to the experimental estimate. Nevertheless, this result is somewhat fortuitous given that, in the particular case of the **POG10** system, the sum of intra- and intermolecular interactions remains approximately constant upon the monomer → triple-helix transition.

Summary and Conclusions

In this work we reviewed several computational methods developed during the last years for computing the energy of large molecules using only fragment energies. Although some of the previous methods have been introduced independently to each other, a comparative analysis reveals their common roots, which, in our opinion, can be traced back to the general formalism of the MBE method. For biomolecules constructed with repetitive building blocks (residues, secondary structural elements,...), it is proposed that a simple thermochemical approach is probably the best option for formulating a *standard* fragment energy method. The validity of the fragment QM energy strategy has been tested intensively considering a challenging problem for simulation methodologies, that is, the prediction of the interchain interaction energy and the free energy for dissociation of a prototypical collagen model. The comparison of our fragment-based energies with experimental data and former theoretical results shows that the actual applicability of the fragment QM methods in biomolecular simulations will rely heavily on the proper combination of QM and MM calculations as well as in the conformational sampling performed by MM methods. Moreover, the correction of the inter- and intramolecular BSSE will be critically important for obtaining realistic energies of either interaction or conformational changes.

Since the MM-PB method predicts a more accurate value than the fragment-based QM calculations for the ΔG change in the monomer → triple-helix transition of the **POG10** system, one may raise the question of whether the fragment QM approaches are really needed. Clearly, the fragment QM calculations would have a broader applicability since they can be used to investigate all kinds of interactions and chemical transformations involving biomolecules. For example, most of the current force fields have been developed without specifically considering the interactions of biomolecules with metal ions, clusters, or surfaces, and therefore, the application of fragment-QM methodologies to study *biomaterials* could provide reliable energetic data, which in turn could be useful for the development and validation of new MM parameters. In addition, we point out that the QM charge densities obtained in the fragment calculations contain much valuable information that can be used for estimating other QM properties (e.g., electrostatic potential) and deriving QM descriptors (e.g., for determining ligand affinity). Similarly, the fragment QM calculations could also be used to outline electron pathways connecting the electron donor and acceptor sites in redox metalloproteins[41] and the energy gaps between electronic states. Therefore, with the continuous improving in the efficiency of QM methodologies, the decreasing cost of computer hardware, as well as a necessary standardization of the fragment energy approach by means of intensive computational experimentation, the full QM description of large biomolecules could be done regularly in the near future.

Computational Methods

DFT Calculations. Density functional theory methods have become the most popular QM methodology for the study of biomolecules because they include electron correlation effects at a relatively cheap computational cost. In principle, the Perdew−Burke−Ernzerhoff (PBE)[42] and Tao−Perdew−Staroverov−Scuseria (TPSS)[43] functionals are particularly attractive for performing fragment energy calculations, since they are nonempirical GGA functionals that give results with an acceptable quality in any type of chemical systems including macromolecules and condensed phases. In this work, we used the PBE functional combined with a double-ζ plus polarization basis set (SVP). [44] The reliability of the PBE/SVP level of theory was assessed by carrying out some validation calculations on a small triple-helix system (see below).

All DFT calculations were performed using the TURBOMOLE suite of programs,[45] in the framework of the multipole accelerated resolution-of-the-identity approximation (MARI-J) using the appropriate auxiliary basis set.[46,47] To estimate the effect of the solvent environment on the DFT energies, we used the conductor-like screening model (COSMO) included in TURBOMOLE in which the solvent dielectric continuum is approximated by a scaled conductor.[48] The optimized atomic COSMO radii (r_H = 1.3 Å, r_C = 2.0 Å, r_N = 1.83 Å, and r_O = 1.72) were used to generate the solvent-accessible molecular cavity.[49] Note that in the thermochemical fragment energy calculations reported in this work long-range electrostatic effects are truncated in the different fragment calculations and that, therefore, a molecular cavity is constructed around each fragment system

Table 4. Average Values and Standard Deviations for the Interaction Energy (kcal/mol of THP) among the Three Peptide Chains for 25 Snapshots of the [Ace(Pro-Hyp-Gly)-Nme]$_3$ Trimer

level of theory	ΔE_{int}	level of theory	ΔE_{int}
PBE/SVP[a]	-10.7 ± 1.4	PBE/SVP[b]	-10.7 ± 6.2
PBE/TZVP[a]	-8.6 ± 1.3	PBE/TZVP[b]	-8.2 ± 5.8
PBE/TZVPP[a]	-9.0 ± 1.3		
TPSS/SVP[a]	-9.0 ± 1.1		

[a] Geometries were extracted from the **POG10** MD simulations and relaxed via MM energy minimization. [b] Geometries were extracted from the **POG10** MD simulations.

($Y_i - B_i^R$, $Y_{i-1} - A_i - B_i^R$, ...). This is fully consistent with the estimation of the full system energy from a combination of reaction energies (eqs R3 and R4).

Since the GGA density functionals are unable to describe dispersive interactions, the DFT energy terms were augmented with an dispersion energy contribution, E_{disp}, which was computed using an empirical formula that has been introduced by Elstner et al.[34] in order to extend their approximate DFT method for the description of dispersive interactions. The E_{disp} expression consists basically of a $-C_6/R^6$ term, which is appropriately damped for short R distances. We used the same parameters for C, N, O, and H and combination rules as those described by Elstner et al.[47]

Molecular Geometries and Molecular Mechanical Calculations. Molecular geometries of the **POG10** system were taken from our previous study on the relative stability of collagen model peptides.[29] The triple-helix and monomer states of **POG10** were subject to 20 and 50 ns molecular dynamics (MD) simulations, respectively, at constant P (1 atm) and T (300 K) in explicit solvent using the AMBER package.[50] From these MD simulations, a set of 100 snapshots was extracted for each state and the internal geometry of the solute molecules was relaxed throughout energy minimization prior to the QM and MM energy calculations. The snapshots were postprocessed through the removal of all solvent molecules.

Thermal contributions to the enthalpy and entropy of solute molecules were estimated by means of MM normal mode calculations using the NAB package[51] and following the prescriptions described elsewhere.[29] The nonpolar solvation energy was computed by combining the explicit solvent representation with an estimation of the relative change in the cavitation free energy of the solute.[52] In our previous work, the conformational entropy of the solute was computed via an expansion of the so-called mutual information functions.[6]

Validation Calculations of the PBE/SVP Level of Theory. Table 4 summarizes the results of some preliminary validation calculations in which we computed the interchain interaction energy in a small THP model ([Ace-(Gly-Pro-Hyp)-Nme]$_3$; 123 atoms). In these calculations, we used the PBE and TPSS functionals combined with different basis sets ranging from the double-ζ SVP to the triple-ζ plus double polarization TZVPP. All DFT energies include the effect of aqueous solvent (COSMO model) and are combined with the empirical estimate of the dispersion energy. We also corrected the BSSE affecting the intermolecular interaction energy by means of the counterpoise method. Coordinates of the small THP models were taken from 25 truncated snapshots of our previous MD simulations of the **POG10** system after having relaxed the internal geometry of the solute molecules via energy minimizations using the AMBER force field.

We see in Table 4 that the average PBE energies obtained with various basis sets are quite similar, the differences being around 1−2 kcal/mol. The TPSS functional gives similar interaction energies to those provided by PBE. By repeating some calculations without relaxing the internal geometry of the small THP models, we found that the average interaction energies are hardly affected, but standard deviations are much higher (~6 kcal/mol). Overall, we conclude that the PBE/SVP energy calculations on the MM-relaxed geometries may constitute a reasonable compromise between quality and computational cost.

Acknowledgment. This research was supported by the following grants: FICyT (Asturias, Spain) IB05-076 and MEC (Spain) CTQ2007-63266. E.S. and N.D. thank MEC for their FPU and Ramon y Cajal contracts, respectively. We are grateful to Dr. H. Valdés for her careful reading of the manuscript and suggestions.

Supporting Information Available: Equivalence between second-order MBE and KEM; Tables S1 and S2 containing the relative and absolute energies of all the MD structures considered in the test calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Yang, W.; Lee, T.-S. *J. Chem. Phys.* **1995**, *103*, 5674.

(2) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1997**, *107*, 879.

(3) Connolly, J. W. D.; Williams, A. R. *Phys. Rev. B* **1983**, *27*, 5169.

(4) Carlsson, A. E. Beyond pair potentials in elemental transition metals and semiconductors. In *Solid State Physics*; Ehrenreich, H., Turnbull, D., Eds.; Academic Press: Boston, 1990; Vol. 43, p 1.

(5) Drautz, R.; Fähnle, M.; Sanchez, J. M. *J. Phys.: Condens. Matter* **2004**, *16*, 3843.

(6) Matsuda, H. *Phys. Rev. E* **2000**, *62*, 3096.

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

(8) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2006**, *106*, 447.

(9) Huang, L.; Massa, L.; Karle, J. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 12690.

(10) Huang, L.; Massa, L.; Karle, J. *J. Chem. Theory Comput.* **2007**, *3*, 1337.

(11) Huang, L.; Massa, L.; Karle, J. *Proc. Nat. Acad. Sci. U.S.A.* **2008**, *105*, 1849.

(12) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.

(13) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *20*, 6832.

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

(15) Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. *Chem. Phys. Lett.* **2001**, *336*, 163.

(16) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.

(17) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342.

(18) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 683.

(19) Fedorov, D. G.; K, K. *J. Phys. Chem. A* **2007**, *111*, 6904.

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

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

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

(23) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.

(24) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.

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

(26) Vreven, T.; Frisch, M. J.; Kudin, N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701.

(27) Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760.

(28) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889.

(29) Suarez, E.; Diaz, N.; Suarez, D. *J. Phys. Chem. B* **2008**, *112*, 15248.

(30) Brodsky, B.; Persikov, A. V. *Adv. Protein Chem.* **2005**, *70*, 301.

(31) Bella, J.; Brodsky, B.; Berman, H. M. *Structure* **1995**, *3*, 893.

(32) Bella, J.; Eaton, M.; Brodsky, B.; Berman, H. M. *Science* **1994**, *266*, 75.

(33) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

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

(35) Jureka, P.; Cerný, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

(36) Valdés, H.; Klusák, V.; Pitoák, M.; Exner, O.; Starý, I.; Hobza, P.; L., R. *J. Comput. Chem.* **2008**, *29*, 861.

(37) Shields, A. E.; van Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272.

(38) Palermo, N. Y.; Csontos, J.; Owen, M. C.; Murphy, R. F.; Lovas, S. *J. Comput. Chem.* **2007**, *28*, 1208.

(39) Asturiol, D.; Duran, M.; Salvador, P. *J. Chem. Phys.* **2008**, *128*, 144108.

(40) Nishi, Y.; Uchiyama, S.; Doi, M.; Nishiuchi, Y.; Nakazawa, T.; Ohkubo, T.; Kobayashi, Y. *Biochemistry* **2005**, *44*, 6034.

(41) Guallar, V. *J. Phys. Chem. B* **2008**, *112*, 13460.

(42) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.

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

(44) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.

(45) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(46) Sierka, M.; Hogekamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136.

(47) Eichkorn, K.; Treutler, O.; Ohm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652.

(48) Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenzc, J. C. W.; Eckert, F. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187.

(49) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. *J. Phys. Chem. A* **1998**, *102*, 5074.

(50) Case, D. A.; Darden, T. A.; Cheatham, I., T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.

(51) Macke, T.; Case, D. A. Modeling unusual nucleic acid structures. In *Molecular Modeling of Nucleic Acids*; Leontes, N. B., SantaLucia, J. J., Eds.; American Chemical Society: Washington, DC, 1998; pp 379.

(52) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238.

CT8005002