Two-center-multipole expansion method: Application to macromolecular systems

Ilia A. Solov'yov,[*],[†] Alexander V. Yakubovich,[*] Andrey V. Solov'yov,[*] and Walter Greiner

Frankfurt Institute for Advanced Studies, Max von Laue Strasse 1, 60438 Frankfurt am Main, Germany

We propose a theoretical method for the calculation of the interaction energy between macromolecular systems at large distances. The method provides a linear scaling of the computing time with the system size and is considered as an alternative to the well-known fast multipole method. Its efficiency, accuracy, and applicability to macromolecular systems is analyzed and discussed in detail.

PACS number(s): 87.15.Kg, 87.15.Aa, 82.20.Kh

I. INTRODUCTION

In recent years, there has been much progress in simulating the structure and dynamics of large molecules at the atomic level, which may include up to thousands and millions of atoms [1–4]. For example, amorphous polymers may have segments each with 10 000 atoms [4] which associate to form partially crystalline lamellae, random coil regions, and interfaces between these regions, each of which may contribute with special mechanical and chemical properties to the system.

With increasing computer powers it became possible to study molecular systems of enormous sizes which were not imaginable just several years ago. For example in Ref. [1] a molecular dynamics simulations of the complete satellite tobacco mosaic virus was performed which includes up to 1 million of atoms. In that paper the stability of the whole virion and of the RNA core alone were demonstrated, and a pronounced instability was shown for the capsid without the RNA.

The study of structure and dynamics of macromolecules often implies the calculation of the potential energy surface for the system. The potential energy surface of a macromolecule carries alot of useful information about the system. For example, from the potential energy landscape it is possible to estimate the characteristic times for the conformational changes [5–7] and for fragmentation [8]. The potential energy surface of a macromolecular system can be used for studying the thermodynamical processes in the system such as phase transitions [9]. In proteins, the potential energy surface is related to one of the most intriguing problems of protein physics: protein folding [9–13]. The rate constants for complex biochemical reactions can also be established from the analysis of the potential energy surface [14,15].

The calculation of the potential energy surface and molecular dynamics simulations often implies the evaluation of pairwise interactions. The direct method for evaluating these potentials is proportional to $\sim N^2$, where N is the number of particles in the system. This places a severe restraint on the treatable size of the system. During the last two decades many different methods have been suggested which provide a linear dependence of the computational cost with respect to N [16–21]. The most widely used algorithm of this kind is the fast multipole method (FMM) [17–24]. The critical size of the system at which this method becomes computationally faster than the exact method is accuracy dependent and is very sensitive to the slope in the N dependence of the computational cost. In Refs. [18,20,25] critical sizes ranging from $N \approx 300$ to 30 000 have been reported. Many discrepancies of the estimates in the critical size arise from differences in the effort of optimizing the algorithm and the underlying code. However, it is also important to optimize the methods themselves with respect to the required accuracy.

The FMM is based on the systematic organization of multipole representations of a local charge distribution, so that each particle interacts with local expansions of the potential. Originally FMM was introduced in Ref. [21] by Greengard and Rokhlin. Later, Greengard's method was implemented in various forms. Schmidt and Lee [20] produced a version based upon the spherical multipoles for both periodic and nonperiodic systems. Zhou and Johnson implemented the FMM for use on parallel computers [26], while Board *et al.* reported both serial and parallel versions of the FMM [25].

Ding *et al.* introduced a version of the FMM that relies upon Cartesian rather than spherical multipoles [18], which they applied to very large scale molecular dynamics calculations. Additionally they modified Greengard's definition of the nearest neighbors to increase the proportion of interactions evaluated via local expansions. Shimada *et al.* also developed a Cartesian-based FMM program [27], primarily to treat periodic systems described by molecular mechanics potentials. In both cases only low order multipoles were employed, since high accuracy was not sought.

In the present paper we suggest a method for calculating the interaction energy between macromolecules. Our method also provides a linear scaling of the computational costs with the size of the system and is based on the multipole expansion of the potential. However, the underlying ideas are quite different from the FMM.

Assuming that atoms from different macromolecules interact via a pairwise Coulomb potential, we expand the potential around the centers of the molecules and build a two-center-multipole expansion using bipolar harmonics algebra. Finally, we obtain a general expression which can be used for calculating the energy and forces between the fragments. This approach is different from the one used in the FMM, where the so-called translational operators were used to expand the potential around a shifted center. Note that the final expression, which we suggest in our theory was not dis-

―――――――
*On leave from the A.F. Ioffe Institute, St. Petersburg, Russia.
†Email address: ilia@fias.uni-frankfurt.de

cussed before within the FMM. Similar expressions were discussed since the early 1950s (see, e.g., Refs. [28–31]). In these papers the two-center-multipole expansion was considered as a new form of Coulomb potential expansion, but the expansion was never applied to the study of macromolecular systems.

We consider the interaction of macromolecules via Coulomb potential since this is the only long-range interaction in macromolecules, which is important for the description of the potential energy surface at large distances. Other interaction terms in macromolecular systems are of the short-range type and become important when macromolecules get close to each other [8]. At large distances these terms can be neglected.

In the present paper we show that the method based on the two-center-multipole expansion can be used for computing the interaction energy between complex macromolecular systems. In Sec. II we present the formalism which lies behind the two-center-multipole expansion method. In Sec. III A we analyze the behavior of the computation cost of this method and establish the critical sizes of the system, when the two-center-multipole expansion method demands less computer time than the exact energy calculation approach. In Sec. III B we compare the results of our calculation with the results obtained within the framework of the FMM. In Sec. IV we discuss the accuracy of the two-center-multipole expansion method.

II. TWO-CENTER-MULTIPOLE EXPANSION METHOD

In this section we present the formalism, which underlies the two-center-multipole expansion method, which will be further referred to as the TCM method. Let us consider two multiatomic systems, which we will denote as A and B. The pairwise Coulomb interaction energy of those systems can be written as follows:

$$U = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{|\mathbf{R}_j^B - \mathbf{R}_i^A|} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{|\mathbf{R}_0 + \mathbf{r}_j^B - \mathbf{r}_i^A|}, \quad (1)$$

where N_A and N_B are the total number of atoms in systems A and B respectively, q_i and q_j are the charges of atoms i and j from the system A and B respectively, $\mathbf{R}_0$ is the vector interconnecting the center of system A with the center of system B, $\mathbf{r}_i^A$ and $\mathbf{r}_j^B$ are the vectors describing the position of charges i and j with respect to the centers of the system A and B, respectively. The centers of both systems can be any suitable points of each of the molecules. It is natural to define them as the centers of mass of the corresponding systems, but in some cases another choice might be more convenient [see, for example, Ref. [8], where we have applied the two-center-multipole (TCM) method for studying fragmentation of alanine dipeptide].

Expression (1) can be expanded into a series of spherical harmonics. The expansion depends on the vectors $\mathbf{R}_0$, $\mathbf{r}_i^A$ and $\mathbf{r}_j^B$. In the present paper we consider the case when

$$|\mathbf{R}_0| > |\mathbf{r}_j^B| + |\mathbf{r}_i^A| \quad (2)$$

holds for all i and j. This particular case is important, because it describes well separated charge distributions, and

can be used for modeling the interaction between complex objects at large distances. In this case the expansion of Eq. (1) reads [32]

$$\frac{q_i q_j}{|\mathbf{R}_0 + \mathbf{r}_j^B - \mathbf{r}_i^A|} = q_i q_j \sum_{L=0}^{\infty} \sum_{M=-L}^{L} \frac{4\pi}{2L+1} \frac{|\mathbf{r}_j^B - \mathbf{r}_i^A|^L}{R_0^{L+1}}$$
$$\times Y_{LM}^*(\Theta_{\mathbf{r}_j^B - \mathbf{r}_i^A}, \Phi_{\mathbf{r}_j^B - \mathbf{r}_i^A}) Y_{LM}(\Theta_{\mathbf{R}_0}, \Phi_{\mathbf{R}_0}). \quad (3)$$

According to Ref. [32] the function

$$|\mathbf{r}_j^B - \mathbf{r}_i^A|^L Y_{LM}^*(\Theta_{\mathbf{r}_j^B - \mathbf{r}_i^A}, \Phi_{\mathbf{r}_j^B - \mathbf{r}_i^A})$$

can be expanded into series of bipolar harmonics

$$|\mathbf{r}_j^B - \mathbf{r}_i^A|^L Y_{LM}^*(\Theta_{\mathbf{r}_j^B - \mathbf{r}_i^A}, \Phi_{\mathbf{r}_j^B - \mathbf{r}_i^A})$$
$$= \sqrt{4\pi(2L+1)!} \sum_{\substack{l_1, l_2 = 0 \\ l_1 + l_2 = L}}^{L} (-1)^{l_2} \frac{(r_i^A)^{l_1}(r_j^B)^{l_2}}{\sqrt{(2l_1+1)!(2l_2+1)!}}$$
$$\times [Y_{l_1}(\Theta_{r_i^A}, \Phi_{r_i^A}) \otimes Y_{l_2}(\Theta_{r_j^B}, \Phi_{r_j^B})]_{LM}^*, \quad (4)$$

where a bipolar harmonic is defined as follows:

$$[Y_{l_1}(\Theta_{r_1}, \Phi_{r_1}) \otimes Y_{l_2}(\Theta_{r_2}, \Phi_{r_2})]_{LM}$$
$$= \sum_{m_1, m_2} C_{l_1 m_1 l_2 m_2}^{LM} Y_{l_1 m_1}(\Theta_{r_1}, \Phi_{r_1}) Y_{l_2 m_2}(\Theta_{r_2}, \Phi_{r_2}). \quad (5)$$

Here $C_{l_1 m_1 l_2 m_2}^{LM}$ are the Clebsch-Gordan coefficients, which can be transformed to the $3j$-symbol notation as follows:

$$C_{l_1 m_1 l_2 m_2}^{LM} = (-1)^{l_1 - l_2 + M} \sqrt{2L+1} \begin{pmatrix} l_1 & l_2 & L \\ m_1 & m_2 & -M \end{pmatrix}. \quad (6)$$

Using Eqs. (4), (6), and (5) we can rewrite expansion (3) as follows:

$$\frac{q_i q_j}{|\mathbf{R}_0 + \mathbf{r}_j^B - \mathbf{r}_i^A|} = q_i q_j \sum_{\substack{l_1, l_2 = 0 \\ l_1 + l_2 = L}}^{\infty} \sum_{\substack{m_1 = -l_1 \\ m_2 = -l_2}}^{l_1, l_2} (-1)^{l_1 + M}$$
$$\times \sqrt{\frac{(4\pi)^3 (2L)!}{(2l_1+1)!(2l_2+1)!}} \begin{pmatrix} l_1 & l_2 & L \\ m_1 & m_2 & -M \end{pmatrix}$$
$$\times \frac{(r_i^A)^{l_1}(r_j^B)^{l_2}}{R_0^{L+1}}$$
$$\times Y_{l_1 m_1}(\Theta_{r_i^A}, \Phi_{r_i^A}) Y_{l_2 m_2}(\Theta_{r_j^B}, \Phi_{r_j^B}) Y_{LM}^*(\Theta_{\mathbf{R}_0}, \Phi_{\mathbf{R}_0}). \quad (7)$$

The multipole moments of systems A and B are defined as follows:

$$Q_{l_1 m_1}^A = \sum_{i=1}^{N_A} q_i (r_i^A)^{l_1} \sqrt{\frac{4\pi}{2l_1+1}} Y_{l_1 m_1}(\Theta_{r_i^A}, \Phi_{r_i^A}),$$

$$Q_{l_2 m_2}^B = \sum_{j=1}^{N_B} q_j (r_j^B)^{l_2} \sqrt{\frac{4\pi}{2l_2+1}} Y_{l_2 m_2}(\Theta_{r_j^B}, \Phi_{r_j^B}). \quad (8)$$

Summing Eq. (7) over i and j, and accounting only for the first L_{max} multipoles in both systems, we obtain

$$U_{mult} = \sum_{\substack{l_1,l_2=0 \\ l_1+l_2=L}}^{L_{max}} \sum_{\substack{m_1=-l_1 \\ m_2=-l_2}}^{l_1,l_2} \frac{(-1)^{l_1+M}}{R_0^{L+1}} \sqrt{\frac{4\pi(2L)!}{(2l_1)!(2l_2)!}}$$
$$\times \begin{pmatrix} l_1 & l_2 & L \\ m_1 & m_2 & -M \end{pmatrix} \times Q_{l_1 m_1}^A Q_{l_2 m_2}^B Y_{LM}^*(\Theta_{R_0},\Phi_{R_0}).$$

(9)

This expression describes the electrostatic energy of the system in terms of a two-center-multipole expansion. Note, that this expansion is only valid when the condition $R_0 > r_i^A + r_j^B$ holds for all i and j, otherwise more sophisticated expansions have to be considered, which is beyond the scope of the present paper.

Summation in Eq. (9) is performed over $l_1,l_2 \in [0 \cdots L_{max}]$; $m_1 \in [-l_1 \cdots l_1]$; $m_2 \in [-l_2 \cdots l_2]$, and the condition $M = m_1 + m_2$ holds. L_{max} is the principal multipole number, which determines the number of multipoles in the expansion.

III. COMPUTATIONAL EFFICIENCY

A. Comparison with direct Coulomb interaction method

In this section we discuss the computational efficiency of the TCM method. For this purpose we have analyzed the time required for computing the Coulomb interaction energy between two systems of charges and the time required for the energy calculation within the framework of the TCM method for different system sizes, and for different values of the principal multipole number.

For the study of the computational efficiency of the TCM method we have considered the interaction between two systems (we denote them as A and B) of randomly distributed charges, for which the condition (2) holds. The charges in both systems were randomly distributed within the spheres of radii $R_A = 1.0N_A^{1/3}$ and $R_B = 1.0N_B^{1/3}$, respectively, and the distance between the centers of mass of the two systems was chosen as $R_0 = 3/2(R_A + R_B)$.

The computational time needed for the energy calculations is proportional to the number of operations required. Thus, the time needed for the Coulomb energy calculation (CE calculation) can be estimated as

$$\tau_{Coul} = \alpha_{Coul} N_A N_B \sim N^2,$$

(10)

where α_{Coul} is a constant depending on the computer processor power and on the efficiency of the code, $N_A \sim N_B \sim N$. From Eq. (10) it follows that the computational cost of the CE calculation grows proportionally to the second power of the system size.

For large systems the TCM method becomes more efficient because it provides a linear scaling with the system size. The time needed for the energy calculation reads
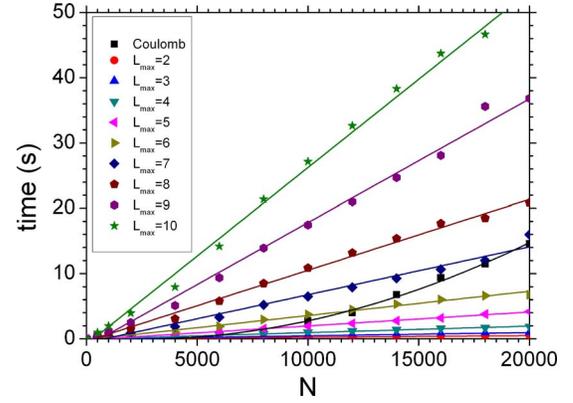

FIG. 1. (Color online) Time needed for energy calculation as a function of the system size.

$$\tau_{mult}(L_{max}) = \beta N + \sum_{l_1=0}^{L_{max}} \sum_{l_2=0}^{L_{max}} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} (N_A \tau_{l_1,m_1} + N_B \tau_{l_2,m_2})$$
$$\approx \alpha_{mult} L_{max}(1+L_{max})^3 N,$$

(11)

where the first term βN corresponds to the computer time needed for allocating arrays in memory and tabulating the computationally expensive functions such as $\cos(\Phi)$ and $\exp(im\Phi)$. $\tau_{l,m}$ is the time needed for evaluation of the spherical harmonic at given l and m, and α_{mult} is a numerical coefficient, which depends on the processor power and on the efficiency of the code. In general it is different from α_{Coul}.

In Fig. 1 we present the dependencies of the computer time needed for the CE calculation (squares) and for the computation of energy within the TCM method for different values of the principal multipole number as a function of system size. This data was obtained on a 1.8 GHz 64-bit AMD Opteron-244 computer.

From Fig. 1 it is clear that the time needed for the CE calculation has a prominent parabolic trend that is consistent with the analytical expression (10). The fitting expression which describes this dependence is given in the first row of Table I. At large N the N^2 term becomes dominant and the other two terms can be neglected. Thus, $\alpha_{Coul} \approx 4.46 \times 10^{-8}$ s.

The fitting expressions which describe the time needed for the energy computations within the TCM method at different values of the principal multipole number are given in Table I, rows 2–10. These expressions were obtained by fitting the data shown in Fig. 1. Note the linear dependence on N. The numerical coefficient in all expressions correspond to the factor $\alpha_{mult} L_{max}(1+L_{max})^3$ in Eq. (11). The fitting expressions in Table I were obtained by fitting of data obtained for systems with large number of particles (see Fig. 1). Therefore these expressions are applicable when $N \gg 1$.

From equations presented in Table I it is possible to determine the critical system sizes at which the TCM method becomes less computer time demanding than the CE calculation. The critical system sizes calculated for different principal multipole numbers are shown in the third column of Table I. These sizes correspond to the intersection points of

TABLE I. Fitting expressions for the computational time needed for the CE calculation and for the energy computation within the TCM method at different values of the principal multipole number L_{max} (second column). System sizes, for which the Coulomb energy calculation becomes more computer time demanding at a given value of L_{max} are shown in the third column.

L_{max}	$\tau(N)$ (s)	N_{max}
Coulomb	$0.11736 - 0.0002N + 4.6768 \times 10^{-8} N^2$	
2	$-0.01986 + 3.0 \times 10^{-5} N$	4223
3	$-0.03159 + 5.0 \times 10^{-5} N$	4662
4	$-0.04714 + 1.0 \times 10^{-4} N$	5809
5	$-0.16054 + 2.1 \times 10^{-4} N$	8026
6	$-0.14710 + 3.7 \times 10^{-4} N$	11 704
7	$-0.59675 + 7.4 \times 10^{-4} N$	19 308
8	$-0.35383 + 10.9 \times 10^{-4} N$	27 212
9	$-1.15856 + 1.9 \times 10^{-3} N$	44 286
10	$-0.83688 + 2.71 \times 10^{-3} N$	61 892

the parabola describing the time needed for the CE calculation with the straight lines describing the computational time needed for the TCM method. In Fig. 1 one can see six intersection points for $L_{max} = 2-7$.

From Eq. (11), it follows that computation time of the energy within the framework of the TCM method grows as the power of 4 with increasing L_{max}. To stress this fact, in Fig. 2 we present the dependencies of the computation time obtained within the TCM method at different system sizes as a function of principal multipole number. All curves shown in Fig. 2 can be perfectly fitted by the analytical expression (11). In the inset to Fig. 2, we plot the dependence of the fitting coefficient α_{mult} as a function of the system size. From this plot it is seen that α_{mult} varies only slightly for all system sizes considered, being equal to $(1.982 \pm 0.015) \times 10^{-7}$ s.

Thus, the expression for the time needed for the energy calculation within the framework of the TCM method reads as
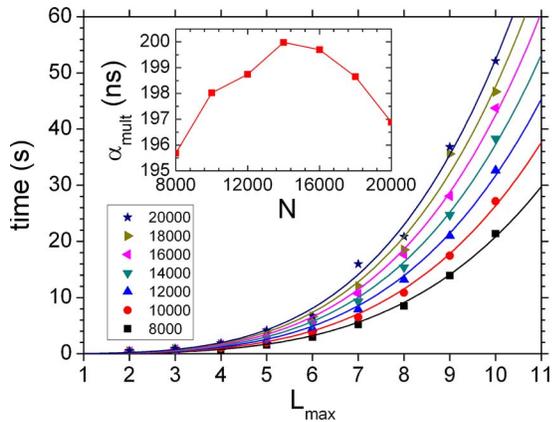

FIG. 2. (Color online) Time needed for the calculation of energy of the systems of different sizes computed within the framework of the TCM method as a function of the principal multipole number L_{max}. In the inset we plot α_{mult} as a function of the system size.

$$\tau_{mult}(L_{max}) \approx 1.98 \times 10^{-7} L_{max}(1 + L_{max})^3 N. \quad (12)$$

Note, that $\alpha_{mult} = 1.98 \times 10^{-7}$ s is larger than $\alpha_{Coul} \approx 4.46 \times 10^{-8}$ s, since in one turn of the TCM method more algebraic operations have to be done, than in one turn of the CE calculation.

From the analysis performed in this section it is clear that the TCM method can give a significant gain in the computation time. However, at larger principal multipole numbers ($L_{max} = 8, 9, 10$) this method can compete with the CE calculation only at system sizes greater than 27 000–61 000 atoms. The accounting for higher multipoles is necessary if the distance between two interacting systems becomes comparable to the size of the systems. In the next section we discuss in detail the accuracy of the TCM method and identify situations in which higher multipoles should be accounted for.

B. Comparison with the fast multipole method

The fast multipole method (FMM) [21–23] is a well known method for calculating the electrostatic energy in a multiparticle system, which provides a linear scaling of the computing time with the system size. In order to stress the computational efficiency of the TCM method in this section we compare the time required for the energy calculation within the framework of the FMM and using the TCM method.

To perform such a comparison we used an adaptive FMM library, which has been implemented for the Coulomb potential in three dimensions [24,33]. We have generated two random charge distributions of different size and calculated the interaction energies between them as well as the required computation time using the FMM and the TCM methods. As in the previous section the charges in both systems were randomly distributed within the spheres of radii $R_A = 1.0 N_A^{1/3}$ and $R_B = 1.0 N_B^{1/3}$, respectively, and the distance between the center of mass of the two systems was chosen as $R_0 = 3/2(R_A + R_B)$.

In Fig. 3 we present the comparison of the computer time needed for the FMM calculation (squares) and for the computation of energy within the TCM method (triangles) as a function of system size. These data were obtained on an Intel(R) Xeon(TM) CPU 2.40 GHz computer. In the upper and lower insets of Fig. 3 we show the relative error of the FMM and of the TCM methods as a function of the system size respectively, which is defined as follows:

$$\eta_{method} = \frac{|U_{Coul} - U_{method}|}{|U_{Coul}|} 100 \% . \quad (13)$$

Here "method" indicates the FMM or the TCM methods. For comparing the efficiency of the two methods we have considered different charge distributions within the size range of 100 to 10 000 particles. Each point in Fig. 3 corresponds to a particular charge distribution. For each system size ten different charge distributions were used. The time of the FMM calculation depends on the charge distribution, as is clearly seen in Fig. 3. Note that for a given system size the calculation time of the FMM can change by more than a factor of 5,
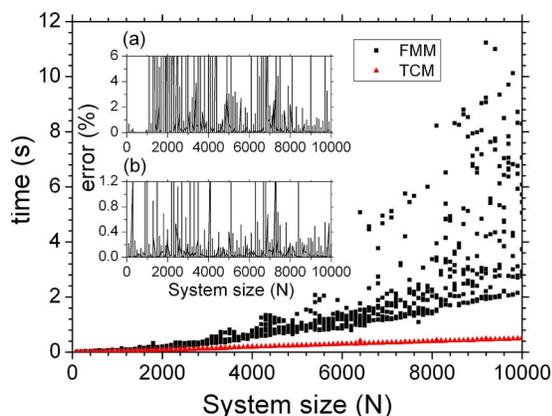
FIG. 3. (Color online) Time needed for the calculation of the interaction energy between two systems as a function the total number of particles calculated within the framework of the TCM method (triangles) and within the framework of the FMM (squares). Plots (a) and (b) in the inset show the relative error of the FMM and of the TCM methods respectively as a function of the system size, computed using Eq. (13).

depending on the charge distribution (see points for N =10 000 in Fig. 3).

For all system sizes FMM requires some minimal computer time for calculating the energy of the system, which increases with the growth of system size (see Fig. 3). The comparison of the minimal FMM computation time with the computation time required for the TCM method shows that the TCM method appears to be significantly faster than the FMM. For N=10 000 FMM requires at least 2.15 s to compute the energy, while the TCM method requires 0.53 s, being approximately 4 times faster.

The results of the TCM method calculation shown in Fig. 3, were obtained for L_{max}=2. The analysis of relative errors presented in the inset to Fig. 3 shows that with this principal multipole number it is possible to calculate the energy between two systems with an error of less than 1% for almost arbitrary charge distribution. Note that for the same charge distributions the error of the FMM is much more, being about 5% in almost all of the considered systems. This allows us to conclude that the TCM method is more efficient and more accurate than the classical FMM.

It is important to mention that in the traditional implementation, FMM calculates the total electrostatic energy of the system while TCM method was developed for studying interaction energy between system fragments. It is possible to modify the FMM to study only interaction energies between different parts of the system. However, the computation cost of the modified FMM is expected to be higher than of the TCM method. This happens because, within the framework of the modified FMM method, the field created by one fragment of the system should be expanded in the multipole series and the interactions of the resulting multipole moments with the charges from another fragment should be calculated. Thus the computation cost of this method will be proportional to $N_A N_B$, where N_A and N_B are the number of particles in two fragments, while the TCM method is proportional to $N_A + N_B$. The computation cost of the modified version of the FMM depends quadratically on the size of the system, because in this method the interacting fragments should be considered as two independent cells, while traditional FMM uses a hierarchical subdivision of the whole system into cells to achieve linear scaling.

So far we have considered only the interaction between two multiparticle systems *in vacuo*, and demonstrated the efficiency of the TCM method in this case, although the TCM method can also be applied to the larger number of interacting systems. The study of structure and dynamics of biomolecular systems consisting of several components (i.e., an ensemble of proteins, DNA, macromolecules in solution) is a separate topic, which is beyond the scope of this paper and deserves a separate investigation.

IV. ACCURACY OF THE TCM METHOD: POTENTIAL ENERGY SURFACE FOR PORCINE PANCREATIC TRYPSIN/SOYBEAN TRYPSIN INHIBITOR COMPLEX

We have calculated the interaction energy between two proteins within the framework of the TCM method and compared it with the exact Coulomb energy value. On the basis of this comparison we have concluded about the accuracy of the TCM method.

In the present paper we have studied the interaction energy between the porcine pancreatic trypsin and the soybean trypsin inhibitor proteins [Protein Data Bank (PDB) [36] entry 1AVW [37]]. Trypsins are digestive enzymes produced in the pancreas in the form of inactive trypsinogens. They are then secreted into the small intestine, where they are activated by another enzyme into trypsins. The resulting trypsins themselves activate more trypsinogens (autocatalysis). Members of the trypsin family cleave proteins at the carboxyl side (or "C-terminus") of the amino acids lysine and arginine. Porcine pancreatic trypsin is a archetypal example. Its natural noncovalent inhibitor (porcine pancreatic trypsin inhibitor) inhibits the enzyme's activity in the pancreas, protecting it from self-digestion.

Trypsin is also inhibited noncovalently by the soybean trypsin inhibitor from the soya bean plant, although this inhibitor is unrelated to the porcine pancreatic trypsin inhibitor family of inhibitors. Although the biological function of the soybean trypsin inhibitor is mostly unknown it is assumed to help defend the plant from insect attack by interfering with the insect digestive system.

The structure of both proteins is shown in Fig. 4. The coordinate frame used for our computations is marked in the figure. This coordinate frame is consistent with the standard coordinate frame used in the PDB.

We use this particular example as a model system in order to demonstrate the possible use of the TCM method. Therefore environmental effects are omitted and we consider only the protein-protein interaction *in vacuo*. The porcine pancreatic trypsin and the soybean trypsin inhibitor include 223 and 177 amino acids, respectively. Both proteins include 5847 atoms. Thus for such system size the TCM method is faster than the CE calculation if $L_{max} \lesssim 4$ (see Table I).

We have calculated the interaction energy between the porcine pancreatic trypsin and soybean trypsin inhibitor as a
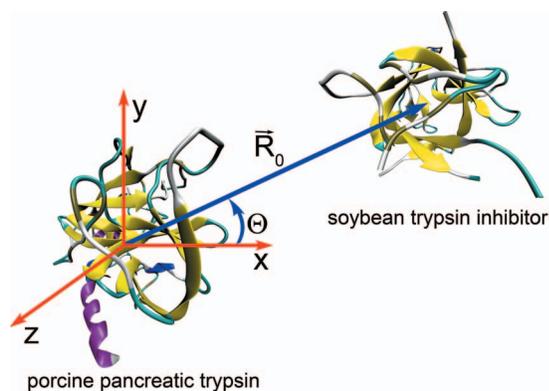
FIG. 4. (Color online) Structure of the porcine pancreatic trypsin and soybean trypsin inhibitor with the coordinate frame used for the energy computation. The figure has been rendered with help of the visual molecular dynamics (VMD) visualization package [38].

function of distance between the centers of masses of the fragments $\vec{R}_0$ and the angle Θ, which is determined as the angle between the x axis and the vector $\vec{R}_0$ (see Fig. 4). We have assumed that the porcine pancreatic trypsin is fixed in space at the center of the coordinate frame and have restricted $\vec{R}_0$ to the xy plane. Of course, the two degrees of freedom considered are not sufficient for a complete description of the mutual interaction between the two systems. For this purpose at least six degrees of freedom are needed. However, for our example of the energy calculation of the porcine-pancreatic-trypsin–soybean-trypsin-inhibitor complex within the framework of the TCM method the two coordinates $\vec{R}_0$ and Θ are sufficient.

The interaction energy of the porcine pancreatic trypsin with the soybean trypsin inhibitor as a function of R_0 and Θ calculated within the framework of the TCM method is shown in Fig. 5. The Coulomb interaction energy between the two proteins is shown in plot (a). In Ref. [8] it has been shown that the interaction energy between two well separated biological fragments arises mainly due to the Coulomb forces. In the present paper we consider $R_0 \in [58, 100]$ Å and $\Theta \in [0, 360]°$, at which condition (2) holds and both proteins can be considered as separated. This means that the potential energy surface shown in Fig. 5(b) describes the interaction energy between the porcine pancreatic trypsin and the soybean trypsin inhibitor on the level of accuracy of at least 90%.

Figure 5(a) shows that one can select several characteristic regions on the potential energy surface marked with numbers 1–4. The corresponding configurations (states) of the system are shown in Fig. 6. The potential energy surface is determined by the Coulomb interactions between atoms, thus at large distances it raises and asymptotically approaches zero. State 1 has the maximum energy within the considered part of the potential energy surface because this state corresponds to the largest contact separation distance between porcine pancreatic trypsin and the soybean trypsin inhibitor being equal to 54.8 Å.

At smaller distances the potential energy decreases due to the attractive forces acting between the two proteins. State 2
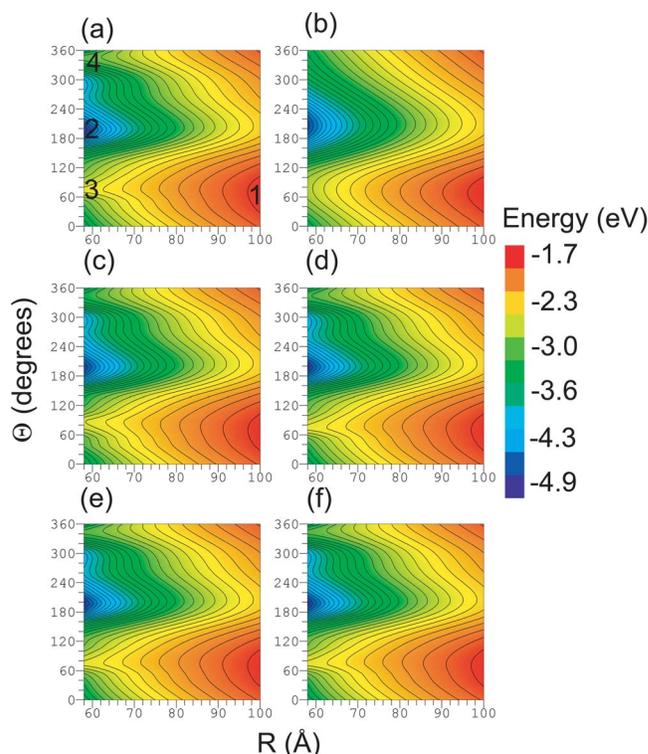

FIG. 5. (Color online) The interaction energies of the porcine pancreatic trypsin with the soybean trypsin inhibitor calculated as the function of R_0 and Θ (see Fig. 4) within the framework of the TCM method at different values of the principal multipole number L_{max}: $L_{max}=2$ (b); $L_{max}=4$ (c); $L_{max}=6$ (d); $L_{max}=8$ (e); $L_{max}=10$ (f). The result of the CE calculation is shown in plot (a).

corresponds to the minimum on the potential energy surface. It arises because a positively charged polar arginine (R125) from the porcine pancreatic trypsin approaches the negatively charged site of the soybean trypsin inhibitor, which includes negatively charged polar amino acids glutamic acid
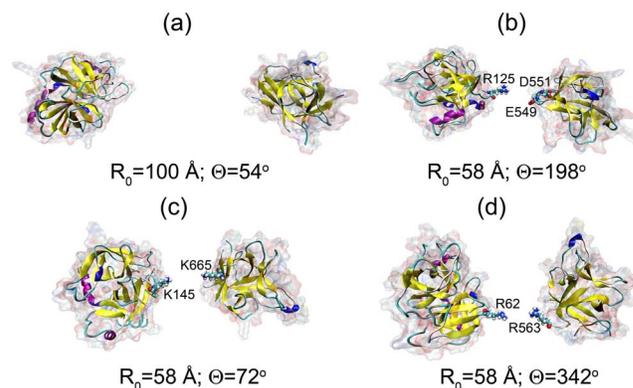

FIG. 6. (Color online) Relative orientations of the porcine pancreatic trypsin and the soybean trypsin inhibitor, corresponding to the selected points on the potential energy surface presented in Fig. 5: state 1 (a), state 2 (b), state 3 (c), state 4 (d). Below each image we give the corresponding values of R_0 and Θ. Some important amino acids are marked according to their PDB identification. The figure was prepared with the help of the VMD visualization package [38].
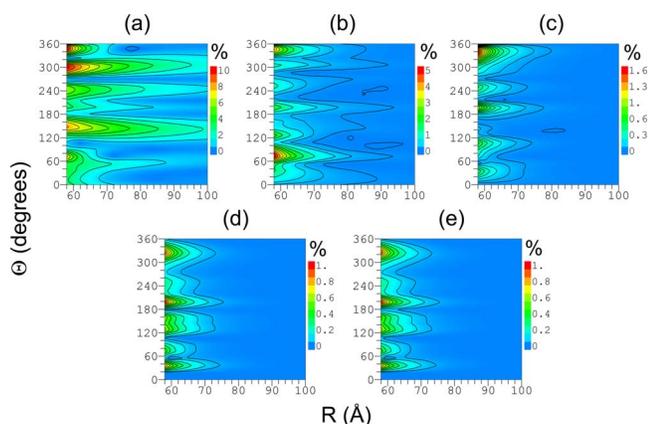
FIG. 7. (Color online) Relative error of the interaction energies of the porcine pancreatic trypsin with the soybean trypsin inhibitor calculated as the function of R_0 and Θ within the framework of the TCM method at different values of the principal multipole number L_{max}: $L_{max}=2$ (a), $L_{max}=4$ (b), $L_{max}=6$ (c), $L_{max}=8$ (d), $L_{max}=10$ (e).

(E549) and aspartic acid (D551) (see state 2 in Fig. 6). The strong attraction between the amino acids leads to the formation of a potential well on the potential energy surface. This observation is essential for dynamics of the attachment process of two proteins, because it establishes the most probable angle at which the proteins stick in the xy plane of the considered coordinate frame ($\Theta=192°$).

States 3 and 4 correspond to the saddle points on the potential energy surface and have energies higher than state 2. They are formed because at these configurations two positively charged polar amino acids from the two proteins become closer in space providing a source of a local repulsive force. In state 3 these amino acids are lysines (K145 and K665) (see state 3 in Fig. 6), and in state 4 these are arginines (R62 and R563) (see state 4 in Fig. 6).

In Fig. 5(b) we show the potential energy surface obtained within the framework of the TCM method with $L_{max}=2$, i.e., with accounting for up to the quadrupole-quadrupole interaction term in the multipole expansion (9). From the figure it is seen that the TCM method describes correctly the major features of the potential energy landscape (i.e., the position of the minimum and maximum as well as their relative energies). However, the minor details of the landscape, such as the saddle points 3 and 4 [see Fig. 5(a)] are missed.

The relative error of the TCM method can be defined as follows:

$$\eta(L_{max})(R_0,\Theta) = \frac{|U_{Coul}(R_0,\Theta) - U_{mult}^{L_{max}}(R_0,\Theta)|}{|U_{Coul}(R_0,\Theta)|} \times 100\%,$$

$$(14)$$

where $U_{Coul}(R_0,\Theta)$ and $U_{mult}^{L_{max}}(R_0,\Theta)$ are the Coulomb energy and the energy calculated at given values of R_0 and Θ within the TCM method, respectively. In Fig. 7(a) we present the relative error calculated according to Eq. (14) for $L_{max}=2$. From this plot it is clear that significant deviation from the exact result arise at $\Theta \sim 50-60°$, $140-150°$, $245°$, $300-$

$310°$, and $350°$. The discrepancy at $\Theta \sim 50-60°$, $\Theta \sim 300-310°$, and $\Theta \sim 350°$ arises because the saddle points 3 and 4, cannot be described within the framework of TCM method with $L_{max}=2$. The discrepancy at $\Theta \sim 140-150°$ and $\Theta \sim 245°$ is due to the error in the calculation of the slopes of minimum 2 at $R_0=58$ Å and $\Theta=198°$.

It is worth noting that the relative error of the TCM method with $L_{max}=2$ is less than 10%. With increasing distance between the proteins, the relative error decreases, and becomes less than 5% at $R_0 \geqslant 72$ Å and less than 3% at $R_0 \geqslant 86$ Å [see Fig. 7(a)]. This means that already at $L_{max}=2$ the TCM method reproduces with a reasonable accuracy the essential features of the potential energy landscape. This observation is very important, because TCM method with $L_{max}=2$ requires less computer time than the CE calculation already at $N=4223$ (see Table I). Thus, the TCM method can be used for the identification of major minima and maxima on the potential energy surface of macromolecules and modeling dynamics of complex molecular systems.

Accounting for higher multipoles in the multipole expansion (9) leads to a more accurate calculation of the potential energy surface. Figures 5(c) and 5(d) show the potential energy surfaces obtained within the framework of the TCM method with $L_{max}=4$ and 6, respectively. From these plots it is seen that all minor details of the Coulomb potential energy surface, such as the saddle points 3 and 4 are reproduced correctly. Figure 7(b) shows that the TCM method with $L_{max}=4$ gives the maximal relative error of about 5% at $R_0=58$ Å and $\Theta=75°$, in the vicinity of the saddle point 3. The relative errors in the vicinity of the saddle point 4 and minimum 2 are equal to 4 and 1 %, respectively. The error becomes less than 1% for all values of angle Θ at $R_0 \geqslant 70$ Å. For $L_{max}=6$, the largest relative error is equal to 1.5% at $R_0=58$ Å and $\Theta=340°$ (saddle point 4), becoming less than 1% at $R_0 \geqslant 61$ Å [see Fig. 7(c)].

By accounting for the higher multipoles in the multipole expansion (9) one can increase the accuracy of the method. Thus, with $L_{max}=8$ and 10 it is possible to calculate the potential energy surface with the error less then 1% [see Figs. 5(e) and 5(f) and Figs. 7(d) and 7(e)]. Although the time needed for computing the potential energy surfaces with $L_{max}=8$ and 10 is larger than the time needed for computing the Coulomb energy directly (see Table I), we present these surfaces in order to stress the convergence of the TCM method.

V. CONCLUSION

In the present paper we have proposed a method for the calculation of the Coulomb interaction energy between pairs of macromolecular objects. The suggested method provides a linear scaling of the computational costs with the size of the system and is based on the two-center-multipole expansion of the potential. Analyzing the dependence of the required computer time on the system size, we have established the critical sizes at which our method becomes more efficient than the direct calculation of the Coulomb energy.

The comparison of efficiency of the TCM method with the efficiency of FMM allows us to conclude that the TCM

method has proved to be faster and more accurate than the classical FMM.

The method based on the two-center-multipole expansion can be used for the efficient computation of the interaction energy between complex macromolecular systems. To determine that we have considered the interaction between two proteins: porcine pancreatic trypsin and the soybean trypsin inhibitor. The accuracy of the method has been discussed in detail. It has been shown that accounting of only four multipoles in both proteins gives an error in the interaction energy less than 5%.

The TCM method is especially useful for studying dynamics of rigid molecules, but it can also be adopted for studying dynamics of flexible molecules. In this work we have developed a method for the efficient calculation of the interaction energy between pairs of large multiparticle systems, e.g., macromolecules, being *in vacuo*. The investigation of biomolecular systems consisting of several components (i.e., complexes of proteins, DNA, macromolecules in solution) and the extension of the TCM method for these cases deserves a separate investigation. If a system of interest consists of several interacting molecules being placed in a solution, one can use the TCM method to describe the interaction between the molecules and then to take account of the solution as implicit solvent. This can be achieved using for example the formalism of the Poisson-Boltzmann [34,35], similar to how it was implemented for the description of the antigen-antibody binding and unbinding process [14,15]. The other possibility is to split the whole system into boxes and account for the solvent explicitly by calculating the interactions between the boxes and the molecules of interest. This can be achieved by using the TCM method or a combination of the FMM and the TCM methods. In this case the FMM can be used for the calculation of the resulting effective multipole moment of the solvent, while the TCM method is much better suitable for the description of the macromolecules energetics and dynamics. Note that all of the suggested methodologies provide linear scaling of the computation time on the system size.

The results of this work can be utilized for the description of complex molecular systems such as viruses, DNA, protein complexes, etc., and their dynamics. Many dynamical features and phenomena of these systems are caused by the electrostatic interaction between their various fragments and thus the use of the two-center-multipole expansion method should give a significant gain in their computation costs.

ACKNOWLEDGMENTS

[1] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten, Structure (London) **14**, 437 (2006).

[2] A. Y. Shih, A. Arkhipov, P. L. Freddolino, and K. Schulten, J. Phys. Chem. B **110**, 3674 (2006).

[3] D. Lu, A. Aksimentiev, A. Y. Shih, E. Cruz-Chu, P. L. Freddolino, A. Arkhipov, and K. Schulten, Phys. Biol. **3**, S40 (2006).

[4] H. Meyer and J. Baschnagel, Eur. Phys. J. E **12**, 147 (2003).

[5] A. V. Yakubovich, I. A. Solov'yov, A. V. Solov'yov, and W. Greiner, Eur. Phys. J. D **39**, 23 (2006).

[6] I. A. Solov'yov, A. V. Yakubovich, A. V. Solov'yov, and W. Greiner, Phys. Rev. E **73**, 021916 (2006).

[7] I. A. Solov'yov, A. V. Yakubovich, A. V. Solov'yov, and W. Greiner, Zh. Eksp. Teor. Fiz. **129**, 356 (2006) [J. Exp. Theor. Phys. **102**, 314 (2006)].

[8] I. A. Solov'yov, A. V. Yakubovich, A. V. Solov'yov, and W. Greiner, Zh. Eksp. Teor. Fiz. **130**, 534 (2006) [J. Exp. Theor. Phys. **103**, 463 (2006)].

[9] A. V. Yakubovich, I. A. Solov'yov, A. V. Solov'yov, and W. Greiner, Eur. Phys. J. D **40**, 363 (2006).

[10] C. Chen, Y. Xiao, and L. Zhang, Biophys. J. **88**, 3276 (2005).

[11] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).

[12] A. Liwo, M. Khalili, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **102**, 2362 (2005).

[13] F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, Biophys. J. **83**, 3525 (2002).

[14] E. S. Henriques and A. V. Solov'yov (unpublished).

[15] E. S. Henriques and A. V. Solov'yov, Eur. Phys. J. D. (to be published).

[16] W. F. van Gunsteren and H. J. C. Berendsen, Angew. Chem., Int. Ed. Engl. **29**, 992 (1990).

[17] H. G. Petersen, D. Soelvason, and J. W. Perram, J. Chem. Phys. **101**, 8870 (1994).

[18] H.-Q. Ding, N. Karasawa, and W. A. Goddard III, J. Chem. Phys. **97**, 4309 (1992).

[19] C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Lett. **230**, 8 (1994).

[20] K. E. Schmidt and M. A. Lee, J. Stat. Phys. **63**, 1223 (1991).

[21] L. Grengard and V. Rokhlin, J. Comput. Phys. **60**, 187 (1985).

[22] J. Board and K. Schulten, Comput. Sci. Eng. **2**, 56 (2000).

[23] M. R. Pincus and H. A. Scheraga, J. Phys. Chem. **81**, 1579 (1977).

[24] P. Gibbon and G. Sutmann, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms, Lecture Notes*, edited by J. Grotendorst, D. Marx, and A. Muramatsu (John von Neumann Institute for Computing, Jülich, 2002), Vol. 10, p. 467.

[25] J. A. Board, J. W. Causey, J. F. Leathrum, Jr., A. Windemuth, and K. Schulten, Chem. Phys. Lett. **198**, 89 (1992).

[26] F. Zhao and S. L. Johnson, SIAM (Soc. Ind. Appl. Math.) J.

Sci. Stat. Comput. **12**, 1420 (1991).

[27] J. Shimada, H. Kaneko, and T. Takada, J. Comput. Chem. **15**, 29 (1994).

[28] R. J. Buehler and J. O. Hirschfelder, Phys. Rev. **83**, 628 (1951).

[29] C. G. Joslin and C. G. Gray, J. Phys. A **17**, 1313 (1984).

[30] G. P. Gupta and K. C. Mathur, Phys. Rev. A **23**, 2347 (1981).

[31] W. P. Wolf and R. J. Birgeneau, Phys. Rev. **166**, 376 (1968).

[32] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).

[33] The FMM library for long range Coulomb interactions was provided by the complex atomistic modeling and simulations group, Forschungszentrum Jülich, http://www.fz-juelich.de/zam/fcs/

[34] F. Fogolari, P. Zuccato, G. Esposito, and P. Viglino, Biophys. J. **76**, 1 (1999).

[35] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. Mac-Cammon, Proc. Natl. Acad. Sci. U.S.A. **98**, 10037 (2001).

[36] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, Nucleic Acids Res. **28**, 235 (2000).

[37] H. K. Song and S. W. Suh, J. Mol. Biol. **275**, 347 (1998).

[38] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).