A novel computaional scheme for accurate and efficient evaluation of π-π and π-σ stacking

Yohsuke Hagiwara and Masaru Tateno*

Center for Computational Sciences, University of Tsukuba, Tennodai 1-1-1, Tsukuba Science City, Japan

* To whom the correspondence should be addressed: tateno@ccs.tsukuba.ac.jp

Keywords: protein, DNA, electronic structure, MP2, CCSD(T), basis set limit
PACS classification: 87.15.ag
(87.15. Biomolecules: structure and physical properties, 87.15.ag Quantum calculations)
Running title: π–π/π–σ stacking

Abstract. Stacking involving aromatic rings has significant contribution to structural stability of biological macromolecules. However, conventional calculations such as density functional theory (DFT) and molecular mechanics (MM) fail to estimate such stabilization energies, most of which are fundamentally derived from van der Waals interactions. For the accurate description, higher level *ab initio* calculations, such as CCSD(T), should be employed; however, their computational costs are huge. MM calculations provide better estimation of the interactions of the aromatic rings than the DFT, but not sufficient. In this report, we propose a novel scheme to calculate the interaction energy at the accuracy compatible to the CCSD(T) with the computational costs comparable to the MM calculations. In our scheme, the electron density of the aromatic rings is represented by Gaussian-type functions, and the parameters involved in the functions are determined by an optimization sheme to reproduce the CCSD(T) results. Here, we employ model structures involving tryptophan and tyrosine rings, and successfully obtain the optimal parameter set. By using this type of the representation of stacking proposed, the computational time to calculate the interaction energy is dramatically reduced by 10^{-10} fold, compared with the CCSD(T). *(189 words: limit 200 words)*

1. Introduction

Stacking involving aromatic rings, aliphatic chains, or cations/anions has been widely observed in three-dimensional structures of biological macromolecular systems such as nucleic acids and proteins, and is also known to contribute to thermodynamic stabilization [1-4]. Many theoretical studies have been performed to unravel the dominant energy terms that contribute to such large stabilizations; in these studies, the van der Waals (vdW) energy was shown to be the primary origin of the stabilization energy for stacking between aromatic rings [4, 5]. This indicated that for accurate description of precisely evaluate the electron correlation effects, more sophisticated *ab initio* calculations are required to precisely evaluate the electron correlation effects, such as quantum Monte Carlo (QMC) calculations and the coupled cluster method with singles, doubles, and perturbative triples (CCSD(T)) [6-8]. Thus, it has been shown that conventional *ab initio* calculations, such as Hartree–Fock (HF) and density functional theory (DFT) calculations, are unable to estimate the stabilization energy gained by stacking aromatic rings [4, 5, 9].

Accordingly, to account for the vdW energy using the DFT calculations, various approaches have been proposed; modification of conventional functionals, adding correction terms to the standard DFT energy, and using perturbation methods in the DFT (DFT-SAPT) where the dispersion energy is included as second-order energy [10]. Those approaches have shown a good consistency with CCSD(T) calculations. However, a serious problem included in those calculations is the computational costs, which are too huge to make it impractical to use such methodologies in long-time MD simulations; for instance, more than ~10-ns calcualtions are required for protein folding simulations. Thus, a widely-used way to perform such long MD simulations is to employ molecular mechanics (MM) calculations, where the vdW interaction is described using empirical functions, e.g., Lennard–Jones (LJ) potential. However, it has been reported that the estimates obtained by MM calculations are inadequate in those cases [11-16]. On the contrary, the importance of stacking is well known to structural and functional features of various biological macromolecules including protein folding; actually, for instance, all of the folding simulations for the smallest peptide, i.e. chignolin, which involves stacking between Tyr and Trp, cannot reproduce the conformations of the stacked amino acid residues observed in the experimental structures, even though this stacking has been indicated to be crucial for its structural stabilization of this small peptide [17-19].

The aim of the present study is to provide an accurate description of the stabilization energy from stacking by vdW interactions, which can be efficiently calculated currently using a reasonable level of computational resources. Our goal is to reproduce not only the accuracy required at the CCSD(T) level with a large basis set, but also an efficiency compatible with MM calculations. The scheme developed should then be applicable even to molecular dynamics (MD) calculations, such as protein folding simulations.

2. Methodology

2.1. Energy function

For that purpose, we first express the electron density derived from an atom *i* using a Gaussian-type function:

$$\rho(\mathbf{r}) = \sum_{i}^{N_{atoms}} q_i a_{i,x} \exp\left(-a_{i,x}{}^2 (x - x_\mathrm{i})^2\right) a_{i,y} \exp\left(-a_{i,y}{}^2 (y - y_\mathrm{i})^2\right) a_{i,z} \exp\left(-a_{i,z}{}^2 (z - z_\mathrm{i})^2\right) \quad (1)$$

Here, **r** and q_i show the position vectors and the number of electrons of atom *i*, respectively. x_i, y_i, and z_i are the coordinates of atom *i*, and $a_{i,\lambda}$ (λ = *x*, *y*, *z*) are parameters that regulate distributions of electrons specific to the *x*, *y*, and *z* directions. Anisotropic qualities of the electron density can be taken into consideration when different values are assigned to each $a_{i,\lambda}$. The Gaussian-type function is not essential for the shape of the functions; other functions, such as the Slater-type, can be used to describe the total electron density, although the involvement of the anisotropic effects is complicated, when the Slater-type function is used.

Next, an effective functional is used to describe the vdW energy. One can, in general, utilize any desired potential functional; in this report, we employ the Andersson–Langreth–Lundqvist (ALL) functional, which was developed to correct the vdW energy using an electron density derived from *ab initio* calculations, for example, DFT [20]. The ALL functional describes the vdW interaction between two molecules, denoted by molecules 1 and 2:

$$E_{vdw}^{ALL} = -\frac{6}{(4\pi)^{3/2}} \int_{V_1}\int_{V_2} d^3\mathbf{r}_1 d^3\mathbf{r}_2 \frac{\rho^{1/2}(\mathbf{r}_1)\rho^{1/2}(\mathbf{r}_2)}{\rho^{1/2}(\mathbf{r}_1)+\rho^{1/2}(\mathbf{r}_2)} \frac{1}{r_{12}{}^6} \quad (2)$$

Here, ρ(**r**) is the total electron density at a position vector **r**; $\mathbf{r}_1$ and $\mathbf{r}_2$ are position vectors corresponding to molecules 1 and 2, respectively. The distribution range of the vectors, i.e. those of the electrons, is determined by V_1 and V_2, which show the volumes of molecules 1 and 2, respectively. For numerical calculations of equation (2), the formulation is written using a grid-based description; the grids are distributed around atoms involved in the molecules, and accordingly, this equation is described as a summation of interactions between two atoms, A and B, each of which belongs to either of the regions related to V_1 and V_2, respectively.

$$E_{vdw}^{ALL} = \sum_{A\in V_1}\sum_{B\in V_2} E_{AB} \quad (3)$$

$$E_{AB} = -\frac{6}{(4\pi)^{3/2}} \sum_{\mathbf{r}_1\in A}\sum_{\mathbf{r}_2\in B} \frac{\rho^{1/2}(\mathbf{r}_1)\rho^{1/2}(\mathbf{r}_2)}{\rho^{1/2}(\mathbf{r}_1)+\rho^{1/2}(\mathbf{r}_2)} \frac{1}{r_{12}{}^6} \quad (4)$$

In MM calculations, vdW interactions are described using the LJ potential;

$$E_{vdw}^{Amber} = \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right) \qquad (5)$$

R_{ij} shows the distance between atoms *i* and *j*; A_{ij} and B_{ij} denote the parameters that regulate the degrees of repulsive and attractive energies, respectively. It should be noted here that the interaction energy, described using equation (2), corresponds to the dispersion energy in the vdW interaction, and therefore, the attractive term in the LJ potential can be replaced with equation (2). For the description of bonding and electrostatic energies, we employ the harmonic functions used in the Amber 9.0 program package without modification [21]. Thus, the energy function that describes the total energy of the system involving the stacking between aromatic rings is written as follows.

$$E = E_{bonded} + \sum_{i<j}^{atoms} \frac{q_i q_j}{\varepsilon R_{ij}} + \sum_{i<j}^{atoms} \frac{A_{ij}}{R_{ij}^{12}} - E_{vdw}^{ALL} \qquad (6)$$

Since the ALL functional depends on the electron density $\rho(\mathbf{r})$, parameter values in the functional should be determined through a fitting procedure, so that the total energy curves obtained by higher level *ab initio* calculations prior to this fitting are reproduced by the total energy function involving the ALL functional, i.e. equation (6). This optimization of parameters in the total electron density can be performed as follows. First, the potential energy profiles are calculated by CCSD(T) calculations for the two models in this study, i.e. T-type and parallel-type models. To prepare the model structures in the parallel type conformation, we placed two aromatic rings such that the line through the center of masses of the two rings is perpendicular to each ring, and then, to yield several conformations for obtaining potential curves, one ring was shifted in the perpendicular direction. For the T-type conformation, we placed two rings such that the line through the center of masses is perpendicular to one ring, and is parallel to the other ring. Then, one ring was shifted along the line to yield several conformations.

The accuracy of CCSD(T) calculations is known to be dependent on the basis set, and thus, to estimate accurate energy values from the stacking, a larger basis set is required. However, the computational cost increases markedly as the basis set becomes larger. In this study, CCSD(T) energy is estimated at a basis set limit that mimics energy calculated using a complete basis set, using a procedure described in the section *2.2* [22]. Second, we fit values obtained from the total energy function defined by equation (6), which includes the ALL-functional-based vdW energy to reproduce the potential energy profile obtained by CCSD(T) calculations at the basis set limit. We employed simulated annealing (SA) protocol for the fitting, where $a_{i,\lambda}$, which regulates the distributions of the electron density in equation (1), is used as a variable. In addition, coefficients for the repulsive terms in the LJ potentials used in the Amber 9.0 program are also exploited as another variable in this study. At each step of the SA protocol, a set of those parameters is obtained, and then, the total energy values obtained by equation (6) are calculated for the two modeled structures. Then, deviations between the current outputs using equation (6) and energy values obtained at the CCSD(T) level are evaluated using the Metropolis criteria in the SA protocol.

To summarize, our scheme to obtain an effective potential by computation to describe vdW energy consists of the following four steps: (i) calculation of the total electron density for each grid with use of a parameter set; (ii) calculation of the vdW energy using an effective potential (e.g., the ALL functional) with use of the total electron density; (iii) calculation of the total energy for the stacked aromatic rings with equation (3); and (iv) evaluation of the total energy values obtained in the step (iii) and those from higher level *ab initio* calculations performed beforehand. It should be noted here that the forms of the total electron density, as well as the effective potential to describe the vdW energy, can be substituted by other functionals.

2.2. Estimation of CCSD(T) energy using a complete basis set

CCSD(T) calculations are known to be dependent on the basis sets used; in order to accurately estimate stacking energy, larger basis sets are required. However, computational costs increase significantly with the size of basis sets used. In this study, CCSD(T) energy is estimated at a basis set limit ($E_{\mathrm{CCSD(T)(limit)}}$), which mimics energy calculated using a complete basis set, by exploiting a procedure proposed by Tsuzuki *et al* [22]. According to this scheme, $E_{\mathrm{CCSD(T)(limit)}}$ is calculated on the basis of the following equation:

$$E_{\mathrm{CCSD(T)(limit)}} = E_{\mathrm{MP2(limit)}} + \Delta\mathrm{CCSD(T)(limit)} \qquad (7)$$

where ΔCCSD(T)(limit) denotes the CCSD(T) correction term, i.e., $\Delta\mathrm{CCSD(T)} = E_{\mathrm{CCSD(T)}} - E_{\mathrm{MP2}}$ at the basis set limit. E_{MP2} and $E_{\mathrm{CCSD(T)}}$ denote stacking energies obtained at the MP2 and CCSD(T) levels, respectively. Here, stacking energy at the MP2 level with use of the basis set limit ($E_{\mathrm{MP2(limit)}}$) is estimated by the following equation:

$$E_{\mathrm{MP2(limit)}} = E_{\mathrm{HF(limit)}} + E_{\mathrm{corr(MP2)(limit)}} \qquad (8)$$

$E_{\mathrm{HF(limit)}}$ and $E_{\mathrm{corrMP2(limit)}}$ denote stacking energies at the Hartree–Fock (HF) level (E_{HF}) and MP2 level correlation energies ($E_{\mathrm{corr(MP2)}} = E_{\mathrm{MP2}} – E_{\mathrm{HF}}$), respectively, at the basis set limit. In practice, $E_{\mathrm{MP2(limit)}}$ is obtained by extrapolation of the correlation energy with use of Helgaker's method as follows [23]. First, $E_{\mathrm{coor(MP2)}}$ are calculated by using aug-cc-pVDZ and aug-cc-pvTZ basis sets, and then, these two values are fitted by exploiting a formula, $a+bX^{-3}$ (where X is 2 for aug-cc-pVDZ and 3 for aug-cc-pVTZ). Next, $E_{\mathrm{corr(MP2)(limit)}}$ is estimated by extrapolation of the obtained function, and finally, $E_{\mathrm{MP2(limit)}}$ is obtained by equation 8. The value of ΔCCSD(T)(limit) is estimated by the following equation:

$$\Delta CCSD(T)_{(limit)} \approx \Delta CCSD(T)_{(M)} + \Delta_{(M)}\Delta CCSD(T) \quad (9)$$

Here, ΔCCSD(T)(M) shows ΔCCSD(T) obtained using a medium-size basis set, and Δ(M)ΔCCSD(T) shows a correction term for ΔCCSD(T) obtained using the medium-size basis set, since ΔCCSD(T) is dependent on the size of the basis sets used in calculations. This term is estimated by the following equation:

$$\Delta_{(M)}\Delta CCSD(T) \approx F_{\Delta CCSD(T)}[E_{corr(MP2)(limit)} - E_{corr(MP2)(M)}] \quad (10)$$

where $E_{corr(MP2)(M)}$ denotes $E_{corr(MP2)}$ obtained using the medium-size basis set. $\Delta(M)E_{corr(MP2)}$ is a correction term for $E_{corr(MP2)}$ due to the dependency of the basis set size used in MP2 calculations. $F\Delta_{CCSD(T)}$ is a scaling factor applied to estimate Δ(M)ΔCCSD(T).

Interaction energies of benzene, thiophene and naphthalene dimers, calculated using various basis sets (6-31G*, 6-311G*, 6-311G*, cc-pVDZ and a modified cc-pVTZ basis set), have shown that ΔCCSD(T) is about 20% to 29% of the absolute value of $E_{corr(MP2)}$ [5, 23–27]. These results suggest that Δ(M)ΔCCSD(T) is approximately 25 ± 5% of the absolute value of $\Delta(M)E_{corr(MP2)}$. Therefore, $F_{\Delta CCSD(T)}$ can be set to −0.25. In this manner, CCSD(T) energies for the model systems were calculated to obtain their potentials. All calculations were performed using the Gaussian 03 package [28].

3. Results and discussion

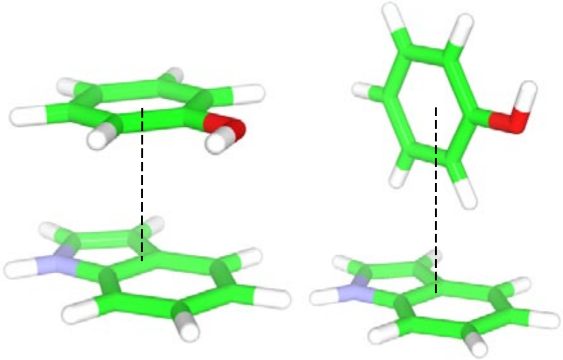

Figure 1.
Modeled structures of the Tyr–Trp system. Left and Right panels show parallel and T-shaped conformations, respectively.

As a test case, we applied the scheme thus developed to stacking between tyrosine (Tyr) and tryptophan (Trp). Modeled structures used have "parallel" conformation, where each base stacks parallel to each other, and "T-shaped" conformation, where each ring stacks perpendicularly to the next. (Fig. 1); for each conformation, we calculated energy potentials with respect to the distance between the centers of mass of two interacting aromatic rings, by MM, HF, DFT, MP2, CCSD(T) calculations at the basis set limit, and a potential function obtained using our scheme.

The energy profiles of the interaction energies with respect to distances between two centers of mass of each aromatic ring are shown in Fig. 2a. According to the profile obtained by CCSD(T) calculations at the basis set limit, energy-minimum geometries for the T- and parallel types are at distances of 4.8 and 3.6 Å, with corresponding interaction energies −3.98 and −3.53 kcal/mol, respectively. By contrast, DFT and HF calculations are unable to estimate the stabilization energy, whereas MP2 calculations significantly overestimate the interaction energy. Compared to those calculations, MM calculations provides potential curves closer to the CCSD(T) calculations. However, interaction energies in optimal-geometries are underestimated, in particular, in the parallel conformations.

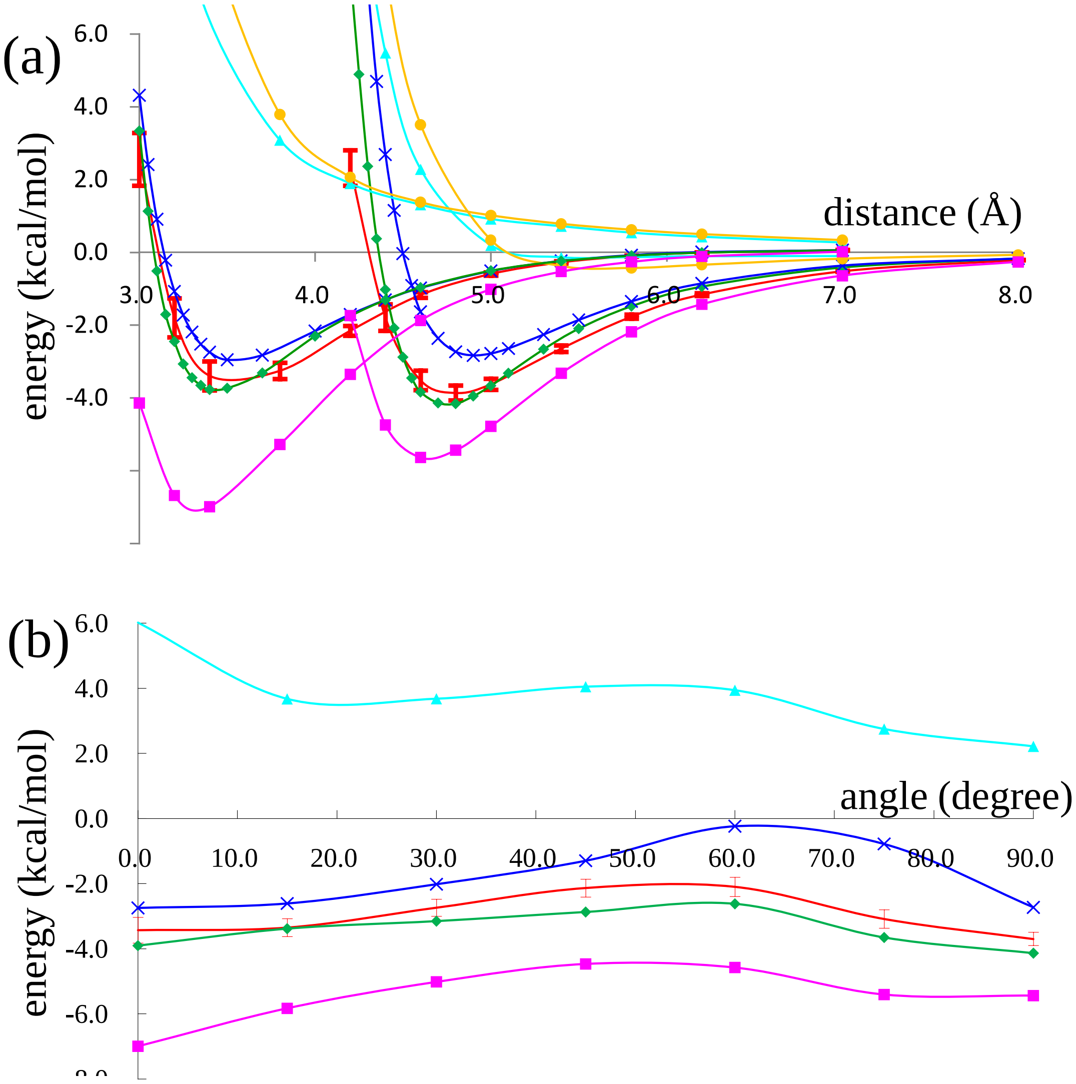

Figure 2.

(a) The potential energy curve with respect to the distance between centers of mass of two aromatic rings (Å). The potential curves obtained by HF (filled circle:●), DFT (triangle: ▲), MP2 (square: ■), MM (cruciform: +), and our scheme (diamond shape: ♦). The potential curve obtained via CCSD(T) at basis set limit is shown with error bars which derives from empirical term described in Methodology section. For each method, two potential curves are shown; the potential curve which have the minimum in larger distance is calculated with respect to T-shaped type conformations, while the other which have the minimum in shorter distance is calculated with use of parallel-type conformations. (b) The potential energy curve with respect to the angle between two aromatic rings (in degrees); the angles are 0°, 15°, 30°, 45°, 60°, 75°, 90°, and the distances between the two center of masses are 3.6 Å, 3.8 Å, 4.0 Å, 4.2 Å, 4.4 Å, 4.6 Å, 4.8 Å, respectively. In those structures, the line through the center of masses of the two rings is perpendicular to one ring. The representation manner is the same as (a).

In contrast, the energy profiles obtained using our scheme are in good agreement with those obtained by CCSD(T) calculations at the basis set limit (Fig. 2a). Thus, it is indicated that the density-based representation proposed in the present study can be fitted to the energy profiles obtained using CCSD(T) calculations at the basis set limit. Further, in order to validate the parameter set obtained by our method, we employed a set of different model structures composed of Tyr and Trp, where the angle between the two aromatic rings is rotated from 0.0° to 90.0° (the angle of 0.0° corresponds to the parallel-type conformation and that of 90.0° to the T-shaped conformation). As a validation test, we generated energy profiles by DFT, MP2, MM, and CCSD(T) calculations at the basis set limit, and the total energy function was obtained. The resulting energy profiles obtained by DFT, MP2, and MM calculations show trends similar to those obtained from the calculations discussed above, which are not consistent with energy profiles obtained by CCSD(T) calculations at the basis set limit (Fig. 2b). By contrast, the energy profiles obtained using our scheme approach very closely to those obtained by CCSD(T) calculations, even though our parameter set is not optimized for conditions used in the calculations. This indicates that completely optimized equation (1) describes all the freedom of the total electron density to represent the interaction of Tyr and Trp, thus confirming the robustness of the developed energy function.

It should be noted here that the computation time using the energy function obtained in this study was dramatically reduced to 1.73 s/step to calculate one point total energy, whereas 1.66×10^{10} s/step is estimated to be required for CCSD(T) calculations using the aug-cc-pVTZ basis set, when using the SGI Altix 3700 system with an Intel Itanium 2 processor (1.6 GHz). Even when our calculation scheme is compared with the method proposed by Tsuzuki *et al* to mimic the CCSD(T) calculation at the basis set limit obtained by MP2 calculations with a huge basis set, the CPU time to calculate our energy function obtained in this study is also significantly reduced—by 10^6-fold.

In this way, using the energy function obtained, we can anticipate performing much more precise simulations, even when the configurations of Tyr and Trp are significantly changed during the calculations, such as molecular dynamics (MD) simulations of biological macromolecules where stacking is involved. In those calculations, our scheme demonstrates its advantages, enabling to significantly reduce the computational time for MD simulations: The CCSD(T) calculations at a basis set limit are performed only to obtain the accurate energy potential as reference for the optimization of the parameters in the electron density function. The optimized parameter values are fixed in MD simulations because of the robustness as mentioned earlier. This leads to significant speed up of MD simulations, since by using our calculation scheme, one can omit the self-consistent field procedure, which is required at every integration step of MD simulations when *ab initio* calculations are used.

4. Acknowledgments

This work is partly supported by grants-in-aid from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) under contract Nos. 19019003, 19340108 and 20051003. Computations were performed using the computer facilities under the “Interdisciplinary Computational Science Program” at the Center for Computational Sciences, University of Tsukuba and the Computer Center for Agriculture, Forestry, and Fisheries Research, MAFF, Japan.

5. References

[1] Meyer E A, Castellano R K and Diederich F 2003 *Angew. Chem. Int. Ed.* **42** 1210.

[2] Hunter C A and Lawson K R, Perkins J and Urch C J 2001 *J. Chem. Soc., Perkin. Trans.* **2**, 651.

[3] Tewari K and Dubey R 2008 *Curr. Med. Chem.* **14** 2911.

[4] Cerny J and Hobza P 2007 *Phys. Chem. Chem. Phys.* **9** 5291

[5] Tsuzuki S, Honda K, Uchimaru T, Mikami M and Tanabe K 2002 *J. Am. Chem. Soc.* **124** 104

[6] Anderson J B 1976 *J. Chem. Phys.* **65** 4121

[7] Head-Gordon M, Pople J A and Frisch M J 1988 *Chem. Phys. Lett.* **153** 503

[8] Pople J A, Head-Gordon M and Raghavachari K J 1987 *J. Chem. Phys.* **87** 5968

[9] Meijer E J and Sprik M 1996 *J. Chem. Phys.* **105** 8684

[10] Sponer J, Riley K E and Hobza P 2008 *Phys. Chem. Chem. Phys.* **10** 2595

[11] Sponer J and Spackova N 2007 *Methods* **43** 278

[12] Fadrna E, Spackova N, Stefl R, Koca J, Cheatham III T E and Sponer J 2004 *Biophys. J.* **87**, 227

[13] Beck D A C, White G W N and Daggett V 2007 *J. Struct. Biol.* **157** 514

[14] Valdes H, Pluhackova K, Pitonak M, Rezac J and Hobza P 2008 *Phys. Chem. Chem. Phys.* **10** 2747

[15] Vondrasek J, Bendova L, Klusak V and Hobza P 2005 *J. Am. Chem. Soc.* **127** 2615

[16] Tateno M and Hagiwara Y 2009 *J. Phys. Condens. Matter.* **21** 064243.

[17] Satoh D, Shimizu S, Nakamura S and TeradaT 2006 *FEBS Lett.* **580** 3422

[18] Seibert M M, Patriksson A, Hess B and van der Spoel D. 2005 *J. Mol. Biol.* **354** 173

[19] Honda S, Yamasaki K, Sawada Y and Mori H 2004 *Structure* **23** 1507

[20] Andersson Y, Langreth D C and Lundqvist B I 1996 *Phys. Rev. Lett.* **76** 102

[21] Case D A *et al* 2006 AMBER 9 University of California, San Francisco

[22] Tsuzuki S, Uchimaru T and Mikami M 2006 *J. Phys. Chem. A.* **110** 2027

[23] Hobza P, Selzle H L and Schlag E W 1996 *J. Phys. Chem.* **100** 18790

[24] Sinnokrot M O, Valeev E V and Sherrill C D 2002 *J. Am. Chem. Soc.* **124** 10887

[25] Tsuzuki S, Honda K, Uchimaru T and Mikami M 2004 *J. Chem. Phys.* **120** 647

[26] Tsuzuki S, Honda K, Uchimaru T and Mikami M 2005 *J. Chem. Phys.* **122** 144323

[27] Tsuzuki S, Honda K and Azumi R 2002 *J. Am. Chem. Soc.* **124** 12200

[28] Frisch M J *et al* 2004 *Gaussian03*. Gaussian, Inc. Wallingford CT