Accuracy of order-*N* density-functional theory calculations on DNA systems using CONQUEST

2008 J. Phys.: Condens. Matter 20 294201

(http://iopscience.iop.org/0953-8984/20/29/294201)

J. Phys.: Condens. Matter **20** (2008) 294201 (10pp) doi:10.1088/0953-8984/20/29/294201

Accuracy of order-N density-functional theory calculations on DNA systems using CONQUEST

T Otsuka[1], T Miyazaki[1], T Ohno[1], D R Bowler[2,3,4] and M J Gillan[2,3,4]

[1] National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0045, Japan
[2] Materials Simulation Laboratory, UCL, Gower Street, London WC1E 6BT, UK
[3] London Centre for Nanotechnology, UCL, 17-19 Gordon Street, London WC1H 0AH, UK
[4] Department of Physics and Astronomy, UCL, Gower Street, London WC1E 6BT, UK

E-mail: OOTSUKA.Takao@nims.go.jp and MIYAZAKI.Tsuyoshi@nims.go.jp

Received 1 February 2008
Published 24 June 2008
Online at stacks.iop.org/JPhysCM/20/294201

Abstract

In preparation for large-scale modelling of DNA systems using the linear-scaling density-functional theory methods implemented in the CONQUEST code, we investigate the effect of the approximations used in the code for DNA test systems. The results of CONQUEST calculations on single DNA bases and on hydrogen-bonded base pairs are compared with experimental ones and with the results from other codes in order to gauge the errors incurred by the use of pseudo-atomic orbital (PAO) basis sets and to assess the accuracy of different density functionals. We then use calculations on hydrated and unhydrated DNA systems containing up to ∼3400 atoms to test the effect of the spatial cut-off R_L required to achieve linear-scaling operation in CONQUEST. We find that PAO basis sets of double-zeta plus polarization quality give satisfactory results, and that generalized gradient functionals reproduce well the energetics of hydrogen bonding between base pairs. The linear-scaling errors can readily be rendered negligible with moderate values of R_L.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

The linear-scaling density-functional theory (DFT) code CONQUEST [1–3] has a well established ability to treat systems containing many thousands of atoms, and its scaling with respect to number of atoms is known to be good. We therefore expect that it will perform well on large, complex biological systems such as DNA. In preparation for such work, we investigate here in detail the accuracy of the methods used in the code when applied to DNA systems, focusing on the errors incurred by the use of localized orbital methods and by the spatial cut-off required to obtain practical linear-scaling. We shall show that the localized orbital errors are acceptably small, and that errors due to spatial cut-off can readily be made completely negligible for DNA systems.

The great importance of being able to model DNA at the atomic scale has long been recognized, and there is an enormous amount of published modelling work, based almost entirely on empirical interaction models. Atomistic modelling of DNA is an important tool in studying a wide range of questions relevant to biology and medicine. Such questions include the way the conformations of DNA are determined by the base sequence, the aqueous environment, and the nature and concentration of counter-ions, as well as the mechanisms by which gene expression is controlled by sequence-specific binding of the DNA molecule to proteins and drug molecules. More recently, modelling has begun to be important in studying the many applications of DNA in nanotechnology, for example in the formation of nano-wires and the functionalization of inorganic surfaces or carbon nanotubes [4–8]. We see the present preparatory work as being of potential long-term importance for the modelling of these and other problems concerning DNA systems.

Although the modelling of large biomolecular systems has hitherto almost always been based on parametrized interaction models, such as those incorporated in the AMBER code [9], there are strong long-term reasons for trying to develop approaches in which the entire system is directly modelled using quantum mechanical techniques [10–12]. The central problem with parametrized models is that of transferability, in other words the assumption that interaction parameters derived for one environment will remain valid in other environments. The interaction parameters characterizing particular types of interaction (electrostatic, non-bonding, bond-stretching, bond-bending, dihedral, etc) in particular types of system are generally derived from quantum chemistry calculations in the gas phase [13]. However, the models are usually intended for use in a condensed phase, often in aqueous solution. But important molecular characteristics, such as dipole moment, may change markedly on going from vacuum to solution, so that the issue of transferability is highly non-trivial. This issue is, of course, well recognized, and efforts have been made to address it, but it is not clear whether it is capable of being completely overcome. It is important to recall here that the reliability of interaction models is often difficult to check against experiment, and if discrepancies are found, it may not be clear which feature of the model is in error. A further problem becomes apparent when the biomolecule interacts with an inorganic substrate, such as a semiconductor or oxide material, or a carbon nanotube [14–16]. In such cases, parametrized interaction models may not be available, so that a development effort is needed for each new case. All these issues give a strong motivation for developing methods that will allow electronic structure methods to be used to model the entire system, the case of interest in the present work being DNA systems in an aqueous environment treated by DFT.

Modelling techniques based on DFT have been very widely used to study the structure, dynamics and electronic structure of an enormous variety of condensed-matter problems across a wide range of scientific disciplines (see e.g. [17]). The reason why DFT has been the electronic structure method of choice is that it scales with system size more favourably than the Hartree–Fock (HF) and post-HF methods of quantum chemistry, so that it can treat large complex systems containing many atoms. Nowadays, DFT simulations on systems of a few hundred atoms with fully converged plane-wave basis sets are fairly routine. However, biomolecular systems are often far larger than this. When the number of atoms N becomes much larger than a few hundred, the cpu time required by standard DFT algorithms scales as N^3, and this means that calculations on systems of more than $\sim$1000 atoms are extremely challenging. Nevertheless, large-scale DFT calculations on DNA performed with standard algorithms have been reported [11, 12]. QM/MM methods are sometimes applied to biological systems to reduce the computational effort, but it is unclear in many cases whether the QM region is large enough, and whether the effects of the boundary between QM and MM regions are negligible. Another approach based on electronic structure techniques for modelling large biomolecules is the set of fragment molecular orbital (FMO) methods proposed by Kitaura *et al* [18, 19], in which a large system is treated by dividing it into smaller fragments. The FMO approach can be applied both with DFT and with HF and post-HF techniques, and there have been a number of studies of DNA systems using the technique [20–22].

However, the ability of DFT to treat very large systems has made major advances in the last 10 years. These advances have been achieved by the development of linear-scaling or $O(N)$ techniques, in which the memory and cpu time requirements are proportional to N [23]. There are now several independent $O(N)$ DFT codes [24–29], including our own CONQUEST code [1–3, 30], which has the important feature of being able to employ as basis sets either pseudo-atomic orbitals (PAOs) or systematically improvable basis functions akin to finite elements, which if necessary can be used to achieve complete basis set convergence. To our knowledge, the first application of $O(N)$ DFT calculations to study a DNA system was achieved using the SIESTA code [10], though the system studied contained somewhat less than 1000 atoms. The linear scaling and parallel scaling properties of CONQUEST were demonstrated many years ago [2], and the practical ability of the code to perform very large DFT calculations has been shown very recently by structural optimization work on three-dimensional Ge islands on Si(001), which employed systems of over 20 000 atoms [3, 30–32]. It is this recent work that gives confidence that very large DNA systems hydrated with explicit water should now be within range of CONQUEST calculations.

Our large-scale work on Ge islands was preceded by a considerable amount of development and testing, which involved quite extensive comparisons of CONQUEST calculations on much smaller systems with the results of standard DFT calculations. These comparisons were vital in establishing the validity of the basis sets and the adequacy of the $O(N)$ spatial cut-off used in the large-scale calculations. We take exactly the same approach here. Indeed, the majority of the results we shall present here concern tests on basis sets and exchange–correlation functionals for individual DNA bases, and for base pairs. It is only after these tests that we present illustrative results on the effect of the $O(N)$ spatial cut-off for systems of over 1000 atoms. As we shall see, the overall outcome is that the approximations involved in achieving $O(N)$ operation for DNA systems appear to be remarkably benign.

The rest of the paper is organized as follows. In the next section, we summarize briefly the techniques used in CONQUEST. In section 3, we present tests on the equilibrium bond lengths of single DNA bases to check the accuracy of the PAO basis sets. Next, the hydrogen bonds in DNA base pairs in the Watson–Crick configuration have been examined for the accuracy of PAOs and exchange correlation functionals. For these tests on single bases and base pairs, where we are concerned only with the reliability of the PAOs or the exchange correlation functionals, a diagonalization method is employed instead of the $O(N)$ technique. Then we report our investigation of the technical settings needed to achieve accurate $O(N)$ calculations, presenting illustrative CONQUEST calculations on large DNA systems both with and without hydrating water molecules. In the last section, concluding remarks are given.

2. Methods

The details of the calculation methods used in CONQUEST are explained in our previous papers [1–3, 33] and the recent progress of the code is shown in [34]. Here, we summarize the main points that we shall need to refer to in the present work.

In CONQUEST, we use the Kohn–Sham density matrix ρ defined as

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_n f_n \psi_n(\mathbf{r}) \psi_n(\mathbf{r}')^\star. \quad (1)$$

Here, $\psi_n(\mathbf{r})$ is the Kohn–Sham eigenfunction for band index n, and f_n is its occupation number. The DFT total energy (the Kohn–Sham total energy) can be calculated from the density matrix, using pseudopotential techniques. For the exchange–correlation part, we can employ either the local density approximation (LDA) with the standard Ceperley–Alder functional [35] or the generalized gradient approximation (GGA) with the functional proposed by Perdew, Burke and Ernzerhof (PBE) [36]. In practice, we represent the density matrix using localized orbitals, which we refer to as 'support functions':

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i\alpha, j\beta} \phi_{i\alpha}(\mathbf{r}) K_{i\alpha, j\beta} \phi_{j\beta}(\mathbf{r}'). \quad (2)$$

Here, the support functions $\phi_{i\alpha}(\mathbf{r})$ are functions that are non-zero only inside 'support regions' centred on the atoms, where i labels the atom and α runs over the support functions on a given atom; the coefficients $K_{i\alpha, j\beta}$ are the matrix elements of the density matrix in the non-orthogonal 'basis' of support functions. The support functions $\phi_{i\alpha}(\mathbf{r})$ themselves are expressed as linear combinations of localized basis functions associated with each atom i. CONQUEST provides two types of basis functions, one being B-splines on regular grids [37] and the other being numerical pseudo-atomic orbitals (PAOs) similar to those used in other localized orbital codes [24, 38, 39]. If we use B-splines, we can systematically improve the accuracy of the basis sets so as to achieve plane-wave accuracy. On the other hand, the advantage of PAOs is they are rather efficient, in the sense that we can perform reasonably accurate calculations with a small number of basis functions. We are planning to use PAOs for our study on DNA systems, and in this work, we use PAOs and pseudopotentials compatible with the SIESTA code [24]. Although support functions can be varied in general, in the present work we consider only the case where each support function is represented by a single PAO.

In order to calculate the density matrix, CONQUEST can employ either conventional diagonalization or the $O(N)$ method. When using diagonalization, we express $\psi_n(\mathbf{r})$ as a linear combination of the support functions, and their coefficients are obtained by diagonalizing the Hamiltonian matrix. Then, the density matrix can be calculated directly from equation (2). On the other hand, in the density matrix minimization method used in $O(N)$ operation, we optimize ρ to minimize the total energy with the constraints that ρ is weakly idempotent (all occupation numbers f_n lie between 0 and 1) and ρ gives the correct total valence electron number. To realize the $O(N)$ behaviour, we also need to use the locality of the density matrix, $\rho(\mathbf{r}, \mathbf{r}') \to 0$ when $|\mathbf{r} - \mathbf{r}'| \to \infty$. The procedure used for this in CONQUEST is the method proposed by Li, Nunes and Vanderbilt (LNV) [40] (referred to as LNV method or ADM method), in which K is expressed in terms of an auxiliary density matrix (ADM) $L_{i\alpha, j\beta}$ by the matrix relation:

$$K = 3LSL - 2LSLSL, \quad (3)$$

with $S_{i\alpha, j\beta} \equiv \langle \phi_{i\alpha} | \phi_{j\beta} \rangle$ the overlap matrix of support functions. A spatial cut-off R_L is then imposed on the L-matrix: $L_{i\alpha, j\beta} = 0$ for $|\mathbf{R}_i - \mathbf{R}_j| > R_\mathrm{L}$, where $\mathbf{R}_i$ are the atomic positions. The reason why this procedure automatically yields weak idempotency is described in the original papers. As the method is variational, increase of R_L results in decrease of the total energy, and an infinite R_L should give us the exact result. The decaying behaviour of the density matrix depends on the energy gap of the system. Large energy gaps result in rapid damping of the density matrix, and we thus expect the ADM method to be suitable for biological systems, which usually have large gaps.

To generate localized PAOs, we have used the SIESTA code with the scheme explained in [24], in which the PAO cut-off radius is controlled by an energy shift used in solving the radial wave equation for a given pseudopotential. A larger energy shift gives a more localized PAO basis set, which usually results in higher total energy, but enables us to perform more rapid calculations. It is important to prepare accurate and efficient PAOs for practical calculations.

3. Results and discussion

3.1. Single DNA bases

We begin our tests by calculating the optimized atomic positions of single DNA bases using CONQUEST and comparing them with those obtained by other codes. The aim here is to prepare reliable PAOs which can be used in our future calculations on DNA systems. We have checked some of our results by repeating the calculations using the SIESTA code, and have confirmed that the results from the two codes are essentially the same. Since some of the energy terms are calculated in a completely different way in the two codes, the agreement provides a useful consistency check. Further technical details of the calculation methods with the PAO basis sets used by CONQUEST are given in [34].

Before presenting our results on single DNA bases, we mention briefly tests performed on simple molecules, including H_2O and NH_3, using LDA. We prepared various PAOs with different energy shifts for single-ζ (SZ), single ζ with polarization function (SZP), double-ζ (DZ) and double-ζ with polarization function (DZP). From the results, we have found that we need at least DZP basis sets and that the results by DZP are accurate enough. With DZP, the difference of the calculated bond lengths from the experimental ones is about 1%, which is consistent with other DFT calculations [41], while smaller basis sets give errors larger than 4%. In addition, we have found that DZP basis sets correctly reproduce the pyramid structure (C_{3v}) as the ground state of NH_3, while SZ basis sets

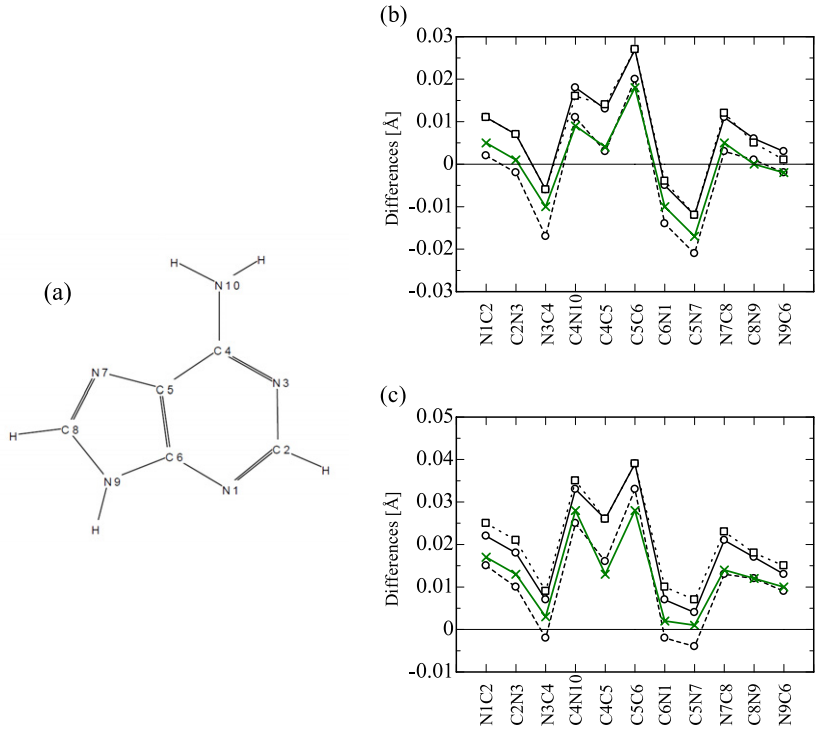

Figure 1. (a) Structure of the adenine molecule, showing indices used to label the atoms. Panels (b) and (c) show errors in calculated bond lengths (Å units) using LDA and GGA-PBE respectively, with PAO basis sets having energy shifts of 50 meV (dotted), 100 meV (solid) and 300 meV (dashed). GAUSSIAN03 results are plotted as green solid line with crosses.

(This figure is in colour only in the electronic version.)

Table 1. Cut-off lengths of the PAOs (Å units), prepared using different energy shifts. In preparing DZP basis sets, the split valence method implemented in the SIESTA code is employed.

	LDA				GGA			
	$l=0$		$l=1$		$l=0$		$l=1$	
	$\zeta=1$	$\zeta=2$	$\zeta=1$	$\zeta=2$	$\zeta=1$	$\zeta=2$	$\zeta=1$	$\zeta=2$
50 meV								
H	3.20	2.17			3.20	2.14		
C	2.71	1.86	3.31	1.98	2.64	1.84	3.31	1.98
N	2.38	1.58	2.91	1.66	2.32	1.56	2.91	1.64
O	2.08	1.36	2.61	1.41	2.08	1.35	2.61	1.41
100 meV								
H	2.97	2.14			2.90	2.09		
C	2.51	1.84	2.99	1.93	2.45	1.82	2.99	1.93
N	2.21	1.56	2.63	1.62	2.15	1.54	2.63	1.62
O	1.98	1.35	2.36	1.40	1.93	1.33	2.42	1.40
300 meV								
H	2.49	2.04			2.43	1.99		
C	2.16	1.79	2.51	1.82	2.16	1.77	2.58	1.84
N	1.90	1.52	2.26	1.56	1.90	1.52	2.26	1.54
O	1.70	1.33	2.03	1.35	1.71	1.31	2.03	1.33

incorrectly give the planar structure (D_{3h}). From these results, we have decided to use DZP basis sets also for the single DNA bases.

For the isolated DNA bases adenine (A), thymine (T), guanine (G) and cytosine (C), we optimize the atomic positions by using LDA or GGA with three different sets of DZP bases prepared by using 50, 100 or 300 meV as the energy shift. (Hereafter, we refer to these PAO basis sets as 50 meV-PAO, 100 meV-PAO, and 300 meV-PAO, respectively.) The cut-off radii of the PAO basis sets are reported in table 1. For the charge density, we use 400 Ryd as the cut-off energy of the FFT grids.

Figure 1 shows the calculated bond lengths of adenine using these three PAO basis sets with LDA and GGA. In the figure, the deviations of the calculated bond lengths from the experimental ones [42] are plotted. The results obtained by the

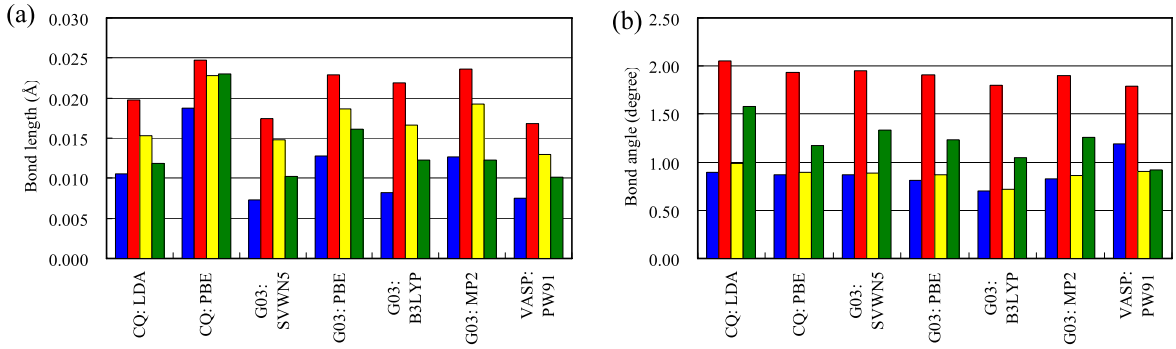

Figure 2. Mean absolute deviation of calculations relative to experiment for (a) bond lengths (Å) and (b) bond angles (degrees) for the DNA bases (left to right) adenine (blue), cytosine (red), guanine (yellow) and thymine (green). Results are shown (left to right) for CONQUEST (LDA and GGA-PBE), GAUSSIAN03 (LDA, GGA-PBE, B3LYP and MP2) and VASP (GGA-PW91). CONQUEST results were obtained with 100 meV-DZP basis sets, GAUSSIAN03 results with cc-pVDZ basis sets. VASP results are from [44].
(This figure is in colour only in the electronic version.)

code GAUSSIAN03 [43] using DZP (cc-pVDZ) are also shown in the figure. Several points emerge from these results. We note first that the error of the bond lengths is about 0.01 Å in LDA. Optimized bond lengths by GGA are larger than LDA ones by about 0.005 Å and the results by GGA are consistent with other DFT calculations [44, 45]. Second, we have found that the calculated bond lengths are almost unchanged even when we go from 50 meV-PAO to 100 meV-PAO. If we use 300 meV-PAOs, the calculated bond lengths are reduced by about 0.01 Å. We conclude that the differences from using different PAO basis sets are small for the bond lengths of the isolated adenine molecule. We note also that the CONQUEST results, especially those with 300 meV-PAO, are close to the GAUSSIAN03 results.

For the four isolated DNA bases, we show the mean absolute differences of the calculated bond lengths and angles from the experimental ones in figures 2(a) and (b), respectively; calculated LDA and GGA results using CONQUEST with 100 meV-PAO basis sets are compared with those obtained by other codes, including GAUSSIAN03 and the plane-wave code VASP [46]. For GAUSSIAN03 calculations, we have used DZP (cc-pVDZ) basis sets, and results by the hybrid functional B3LYP and by the second-order Møller–Plesset (MP2) perturbation theory are also given for comparison. The results show that all the theoretical methods reproduce the experimental structure well. For all the methods, the differences of the calculated and experimental structures in the cytosine case are somewhat larger than for the other bases, but the mean absolute differences of the bond lengths are still smaller than 0.025 Å. Except for this quantitative difference, the results for single DNA bases (C, T, and G) are similar to those we have found in the adenine case. First, the calculated bond lengths by GGA are a little larger than those by LDA. Although the bond lengths by CONQUEST with 100 meV-PAO basis set are a little larger than those by GAUSSIAN03 with cc-pVDZ basis sets, the differences of the results by the two codes are small enough if the same exchange–correlation functional is used. In addition to the comparison between the two localized orbital codes (CONQUEST and GAUSSIAN03), we also compare the calculated structures with those by the plane-wave code VASP here [44]. It should be noted that the functional proposed by Perdew and Wang (PW91) [47] is used in the GGA calculations by VASP, but this functional is almost the same as PBE [36] and the results should be close to each other. We have observed that the calculated bond lengths by VASP are a little smaller than those by the GGA calculations using CONQUEST or GAUSSIAN03 with DZP. We can expect that larger basis sets in CONQUEST or GAUSSIAN03 calculations would result in shorter bond lengths, and this is consistent with the present results. In this sense, we can estimate the errors from the present choice of the DZP basis sets by the differences between the results by CONQUEST and VASP, and we can conclude that the errors are acceptable. Finally, the comparison of the results by different methods in the GAUSSIAN03 calculations shows that the results by PBE, B3LYP and MP2 are almost the same. From these results, we conclude that the results by CONQUEST with the present choice of DZP basis sets are reliable for the calculated bond lengths of the single DNA bases. Note that, for calculated bond angles, there are no clear differences between the results by different codes or different methods.

3.2. DNA base pairs

We now study the interactions between the DNA bases, consisting of hydrogen bonds between the A–T and G–C pairs in the Watson–Crick configuration shown in figure 3. For successful modelling of DNA systems, a good description of the energetics of the hydrogen bonds is essential. Although we have seen that the optimized structures of single DNA bases are robust with respect to the choice of DZP basis sets and exchange–correlation functionals, the same may not be true for the much weaker hydrogen bonds.

We calculate here the optimized structure and the stabilization energy of the base pairs by LDA and GGA using CONQUEST and compare the results with those from other codes and methods. As before, we use the three DZP basis sets, consisting of 50 meV-, 100 meV- and 300 meV-PAOs. By 'stabilization energy' we mean the energy gain on bringing the two molecules together. It is clear that since we are using localized basis sets, we must correct for basis set superposition error (BSSE), and we use the so-called counterpoise method [48] to reduce this error. However, note

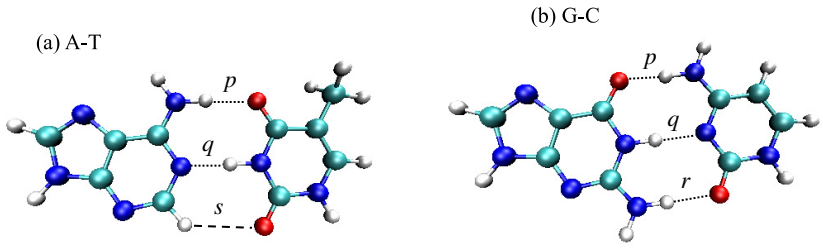

Figure 3. Watson–Crick configuration of the DNA base pairs (a) A–T and (b) G–C (C atoms: light blue; O: red, N: dark blue, H: white). Notation for interatomic distances p, q and s for A–T and p, q and r for G–C characterizing equilibrium geometry of base pairs is shown. (This figure is in colour only in the electronic version.)

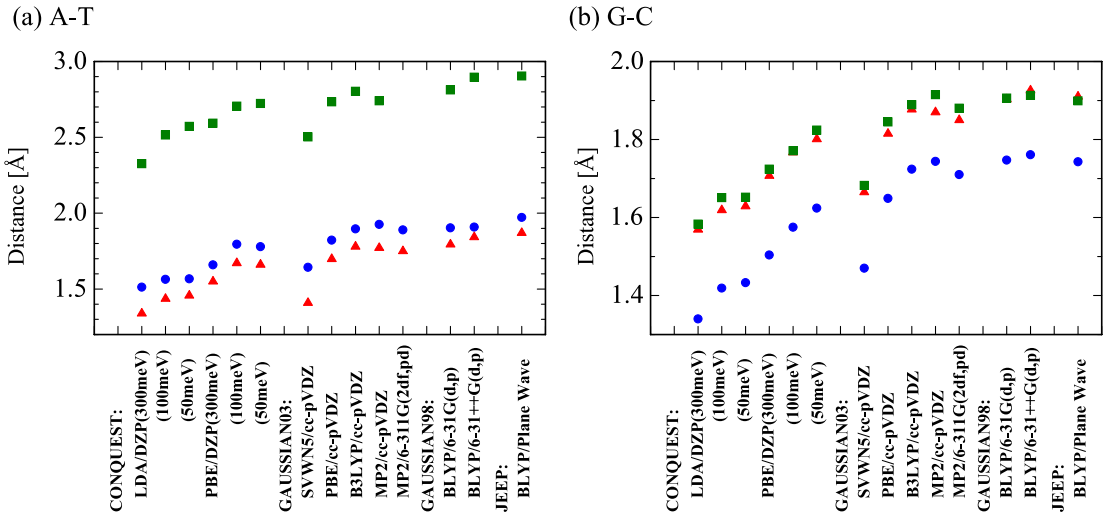

Figure 4. Calculated equilibrium interatomic distances for DNA base pairs (a) A–T and (b) G–C. Distances p, q and s for A–T pair (see figure 3) are shown as blue circles, red triangles and green squares, respectively; distances p, q and r for G–C pair (see figure 3) are shown as blue circles, red triangles and green squares, respectively. CONQUEST results using different DZP basis sets with LDA, GGA-PBE are on the left; GAUSSIAN03 results using LDA(SVWN5), GGA(PBE), MP2 with cc-pVDZ basis sets and the MP2 result from [49] are in the middle; GGA-BLYP results from [50] are on the right.
(This figure is in colour only in the electronic version.)

that it is difficult to calculate forces with this correction and we have not considered any corrections in the optimization of the atomic positions.

Figure 4 reports the calculated structural parameters for the geometries of figure 3, where we use the notation p and q for the hydrogen bond distances in the A–T pair and p, q and r for for those in the G–C pair. The CONQUEST results using the three DZP basis sets with LDA and GGA are plotted in the left of the figures. As in the case of the single DNA bases, we have also calculated the optimized structure of the DNA base pairs using GAUSSIAN03 with LDA(SVWN5), GGA(PBE), B3LYP and MP2. For the MP2 method, the results of hydrogen bond distances from [49] are also shown in the figure. The results by GGA functional BLYP from [50] are also included in figure 4. From the CONQUEST results, we see that the structural parameters calculated with the 300 meV-PAO basis set are significantly different from those with 50 meV or 100 meV-PAOs. The difference is on the order of 0.1 Å, which is about ten times larger than the differences of bond lengths of single DNA bases. We also see that there are large differences between LDA and GGA. LDA hydrogen bond distances are smaller than GGA ones by about 0.2 Å. This effect is also evident in the GAUSSIAN03 results. We note that the CONQUEST results using 50 meV- or 100 meV-PAOs agree well with the results from GAUSSIAN03 using the same exchange–correlation functional. In the figure, the most reliable result is the one obtained by MP2. In figure 4, there are two MP2 results, the left one shows our present result using DZP (cc-pVDZ) basis sets while the right one is from [49] obtained by using 6-311G(2df,pd) basis set. We see that the difference between these two results is small. As pointed out in [51], and as we will see below, the convergence of the stabilization energy by MP2 with respect to the localized basis sets is rather slow. However, the convergence of the optimized structure is much faster if we use DZP or better basis sets. The agreement between the PBE result from CONQUEST and these MP2 results is satisfactory[5,6], while the LDA results show serious discrepancies. These results suggest that GGA-PBE (or B3LYP) is accurate for the description of the structures of A–T and G–C pairs.

[5] We have also compared the present PBE result by Conquest with another MP2 result using TZVPP basis set in [51] for the another measures of the hydrogen bond distances (N(A)–O(T), N(A)–N(T), O(G)–N(C), N(G)–N(C) and N(G)–O(C)). The differences of these structural parameters by the present PBE results are 0.03–0.09 Å, while LDA results show differences of 0.18–0.24 Å.
[6] Reference [55] reports the optimized structure of the DNA base pairs by MP2 is different from a planar structure which is the most stable by HF or DFT methods.

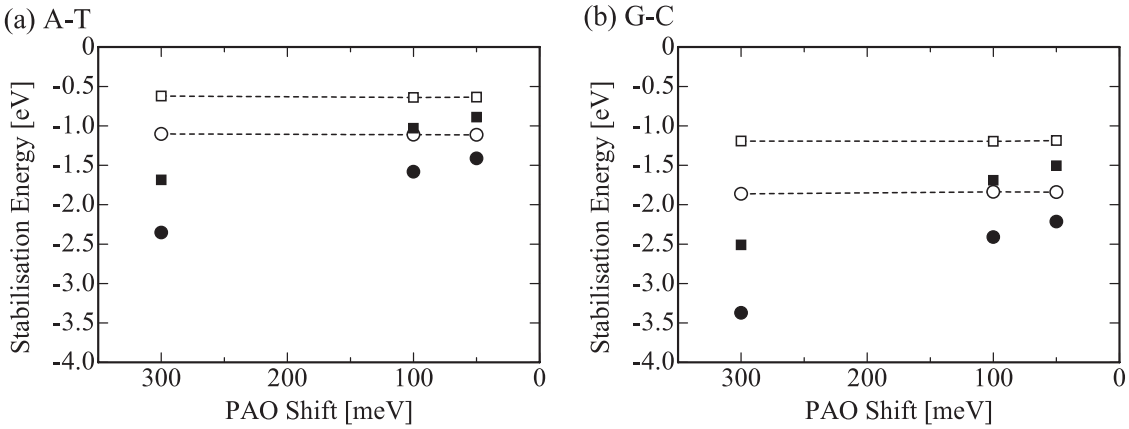

Figure 5. Stabilization energy of (a) A–T and (b) G–C pairs calculated with and without BSSE counterpoise correction using different DZP basis sets specified by their energy shift. Filled circles and squares: uncorrected LDA and GGA-PBE; empty circles and squares: corrected LDA and GGA-PBE. Dashed lines are a guide to the eye.

Table 2. Stabilization energy (eV units) for A–T and G–C pairs by CONQUEST using LDA and PBE. The results by GAUSSIAN03 using cc-pVDZ basis sets with LDA(SVWN5), GGA(PBE), B3LYP and MP2, and results from [51] are also shown. CBS (complete basis sets) limit in the RI-MP2 (approximate resolution of the identity MP2) method means the extrapolated value of the MP2 results with respect to the increase of the basis sets. The CCSD(T) value is evaluated from the energy difference of the RI-MP2 and CCSD(T) results using small basis sets. See [51] for detail.

	CONQUEST		GAUSSIAN03				Ref. [51]	
	LDA	PBE	SVWN5	PBE	B3LYP	MP2	RI-MP2	CCSD(T)
Basis	DZP		cc-pVDZ				CBS limit	
A–T	−1.11	−0.64	−1.05	−0.60	−0.50	−0.48	−0.67	−0.67
G–C	−1.84	−1.20	−1.77	−1.15	−1.05	−0.94	−1.22	−1.25

We now turn to the stabilization energy of the A–T and G–C pairs. Figure 5 shows the stabilization energy with and without BSSE correction. We see that the BSSE corrections are larger for more localized PAO basis sets. However, once BSSE corrections are included, the stabilization energy is almost the same for all the DZP basis sets, the difference of stabilization energy obtained with different DZP basis sets being less than 0.7 kcal mol^{-1} (30 meV). This result suggests that 300 meV-DZP or even more localized DZP basis sets might be accurate enough for the energetics of hydrogen bonds in the Watson–Crick configurations of A–T and G–C pairs. However, we have noted that BSSE corrections were not included in the structure optimization. We therefore think that it remains important to have smaller BSSE to reduce the inconsistency between the present forces and the BSSE-corrected total energy. We believe that the use of 100 meV-DZP basis sets represents a sensible compromise between accuracy, consistency and efficiency.

The BSSE-corrected stabilization energy calculated by LDA or GGA using CONQUEST is shown in table 2, together with GAUSSIAN03 results. For the MP2 calculations on the DNA base pairs, Jurecka *et al* [51] reported that the stabilization energy converges rather slowly with respect to basis set, and they extracted the converged value by extrapolation. They also evaluated the energy correction on going from MP2 to the more accurate CCSD(T) method by calculating the energy difference between these two results using small basis sets. The extrapolated MP2 result and its CCSD(T) corrected value are also included in table 2.

By comparing the CONQUEST and GAUSSIAN03 results, we see again that the two codes agree very well. We also note that the GGA-PBE results are consistent with previous reports [52, 53, 45], some of which use larger basis sets than DZP. This agreement indicates that the present DZP basis sets are reasonably well converged basis sets for PBE calculations. Next, from the comparison between the DFT results and those in the right (RI-MP2 and CCSD(T) corrected from [51]), we can see that the stabilization energy by PBE is close to these values which are most reliable at present. On the other hand, the stabilization energy by LDA is much larger than these values. From table 2 and the results of the optimized structure presented in the above, we can conclude that the local orbital method using CONQUEST with GGA-PBE is fairly accurate for the treatment of the hydrogen bonding of A–T and G–C pairs in DNA systems.

3.3. Order-N calculations on DNA

Having determined what PAO basis sets and exchange–correlation functionals will be needed for future large-scale calculations on DNA systems, we now address the question of $O(N)$ operation, and particularly the cut-off distance R_L to be used in the auxiliary density matrix (ADM) technique.

As a test DNA model, we adopt a DNA decamer hydrated with a large number of water molecules. Specifically, we employ the B-DNA decamer 5′-d(CCATTAATGG)$_2$-3′, which contains 634 DNA atoms. This is hydrated with 932 water

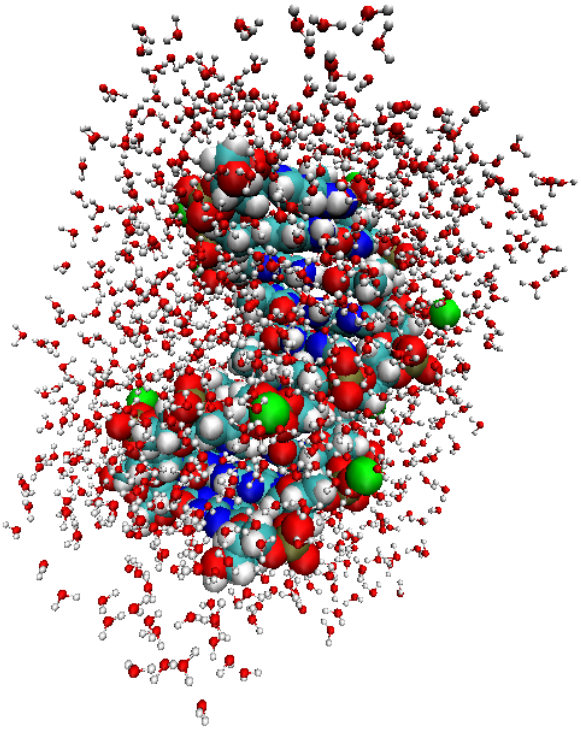

Figure 6. Structure of the 3439-atom hydrated DNA model used in tests of $O(N)$ operation of CONQUEST. The atoms in light blue, red, dark blue, white, gold and green are C, O, N, H, P and Mg atoms, respectively.

(This figure is in colour only in the electronic version.)

molecules, and 9 Mg counter-ions are included for charge neutrality. A view of this 3439-atom system is shown in figure 6. For the general tests of $O(N)$ operation that concern us here, the detailed atomic positions are not particularly important, but we summarize briefly the procedure we have used to generate them. Initial positions were taken from a data set for this DNA decamer in the protein data bank (PDB code: 1WQZ). These positions came from 2.0 Å-resolution x-ray and 3.0 Å-resolution neutron diffraction analysis on the wet crystal at 279 K. The data set includes information about the positions of hydrogen (and deuterium) of the water molecules in the crystal. With these positions, we then added further water molecules using the hydration procedure provided in the AMBER9 package. The system was then equilibrated by using AMBER9 to perform constant-pressure MD, with the PARM99 and TIP3P force fields for the DNA atoms and water molecules respectively. The set of atomic positions resulting from this equilibration was then used for the CONQUEST calculations. The edge lengths of the equilibrated cell are 39.74, 31.03 and 27.09 Å.

Although we have emphasized that the DZP basis set is necessary for the accurate structure determination on DNA systems, we use a SZ basis set for the calculations in this section. We do this because SZ calculations are much more efficient. As we will see in this section, we can employ diagonalization with the SZ basis set for systems having less than 1000 atoms, and we need to do so to check the $O(N)$ results. We expect that the overall electronic structure for a fixed geometry is not very different even if we use SZ basis set. As we pointed out in section 2, the convergence behaviour of the total energy and forces with respect to the cut-off radius of the ADM depends on the electronic structure. In this sense, we expect the results in this paper with the SZ basis set should not be very different from those obtained by DZP basis sets. The calculations with DZP are currently being done, and the results will be reported elsewhere. The SZ basis set used here is generated by using 100 meV as the energy shift. We employ non-self-consistent calculations with the Harris-Foulkes energy functional [54].

First, we have worked on the system which includes only DNA parts without water molecules. Since the system includes only 643 atoms (634 atoms for DNA and 9 Mg atoms) and we work with a SZ basis set at present, we can easily employ conventional diagonalization method for this system. We have also performed $O(N)$ calculations with various cut-offs R_{L} and compared the calculated total energy with the diagonalization results. The dependence of total energy on R_{L} is shown in figure 7 by a red line with circles, with the total energy by diagonalization plotted as a horizontal dotted line. We see that the total energy converges very rapidly,

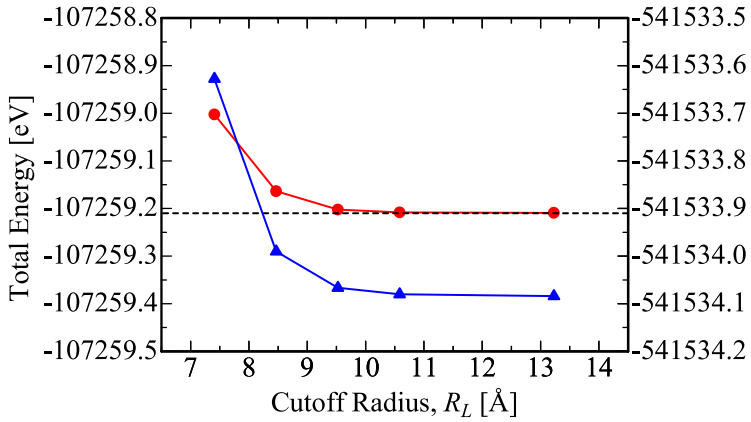

Figure 7. Dependence of total energy on cut-off length R_{L} of the auxiliary density matrix in $O(N)$ calculations on DNA systems. Red line with circles shows the total energy of the DNA system made by removing all water molecules from the system shown in figure 6. Horizontal dashed line shows the total energy of the same system calculated by diagonalization. Blue line with triangles shows the total energy of the system in figure 6.

and the error at $R_L = 8.47$ Å is already only 0.046 eV (1.7 mHartree) in the total system, which corresponds to an error of 7.2×10^{-5} eV/atom. If we increase the cut-off up to 9.53 Å, the error becomes 0.0078 eV (0.28 mHartree) as a total, 1.2×10^{-5} eV/atom. This error is surprisingly small compared with those in semiconducting systems.

This outcome has been confirmed also in the calculations on the whole DNA system including hydrating water molecules shown in figure 6 by a blue line with triangles. Note that the energy scale for the total energy of the system shown in the right side is the same as the one in the left, which is for the system without water molecules. Although the diagonalization method is too expensive for such large systems, we have found that $O(N)$ calculations with the ADM method are feasible and robust. The total energy of the whole system calculated as a function of R_L is plotted in figure 7 and we see that the total energy converges very rapidly. The value at $R_L = 13.23$ Å can be considered as the converged value and the difference from this value is 0.094 eV (3.5 mHartree) at $R_L = 8.47$ Å and 0.017 eV (0.64 mHartree) at $R_L = 9.53$ Å. It should be noteworthy that we have used $R_L = 10.8$ Å in our $O(N)$ calculations on semiconducting surfaces [31, 32]. We believe it is not expensive to use $R_L = 10$ Å as the ADM cut-off distance. With the accuracy of 0.1 mHartree, we will be able to discuss the difference of the total energy induced by a local reaction in a system containing several thousand atoms. This is very encouraging for our future study on DNA systems. We conclude that the present $O(N)$ method using the ADM method is very accurate and promising for DFT studies on DNA systems.

4. Conclusion

In this work, we have investigated in detail the accuracy and the reliable calculation conditions of the theoretical methods used by our linear-scaling DFT code CONQUEST, as a preparation for future DFT studies on DNA systems. First, we have calculated the optimized structures of the four isolated DNA bases. By comparing the results with those by other codes, we have confirmed that the three DZP basis sets, which have different localization lengths, are accurate enough to reproduce the bond lengths of these four bases. The errors in bond lengths by LDA and PBE with these DZP basis sets are about 0.02 Å, which is consistent with other DFT results. Next, the optimized structure and stabilization energy of the DNA base pairs (A–T and G–C) in the Watson–Crick configuration have been examined. The calculated stabilization energy has been compared with those by quantum chemistry methods in [51], which are the most reliable theoretical values at present, to our knowledge. We have found that the stabilization energy by LDA is much larger than these values, while GGA-PBE gives good agreement. We observed a difference in the hydrogen bond distances on the order of 0.1 Å between the results calculated by different DZP basis sets, although the BSSE-corrected stabilization energies are almost the same. We also found that the hydrogen bond distances by LDA are shorter than PBE results by about 0.2 Å and that the optimized structure by PBE is close to that given by the MP2 method. These results show that GGA-PBE gives a rather accurate description of the hydrogen bonds of the A–T and G–C pairs.

Finally, we have employed $O(N)$ calculations on a DNA system containing about 3400 atoms, including hydrating water molecules. We have investigated the accuracy of the ADM method used in the $O(N)$ mode of CONQUEST and have demonstrated that the method is extremely accurate with a moderate value of the ADM cut-off distance R_L. The error in total energy for $R_L = 9.5$ Å is less than 1 mHartree. This result is very encouraging for future studies on DNA systems. We believe that it is now possible to perform reliable and accurate DFT calculations on DNA systems containing several thousand atoms or more, and we hope to report the results of such calculations in the near future.

Acknowledgments

This work is partly supported by Grant-in-Aid for Scientific Research from the MEXT, Japan. DRB is supported by the Royal Society. The calculations in this work were performed by the numerical materials simulator at NIMS, and the facilities of the Cybermedia center at Osaka University and the Earth Simulator Center.

References

[1] Hernández E, Gillan M J and Goringe C M 1996 *Phys. Rev.* B **53** 7147
[2] Bowler D R, Miyazaki T and Gillan M J 2002 *J. Phys.: Condens. Matter* **14** 2781
[3] Bowler D R, Choudhury R, Gillan M J and Miyazaki T 2006 *Phys. Status Solidi* b **243** 989
[4] Ventra M D and Zwolak M 2004 *Encyclopedia of Nanoscience and Nanotechnology* vol 2 (California: American Scientific Publishers) pp 475–93
[5] Endres R G, Cox D L and Singh R R P 2004 *Rev. Mod. Phys.* **76** 195
[6] Kino H, Tateno M, Boero M, Torres J A, Ohno T, Terakura K and Fukuyama H 2004 *J. Phys. Soc. Japan* **73** 2089
[7] Hübsch A, Endres R G, Cox D L and Singh R R P 2005 *Phys. Rev. Lett.* **94** 178102
[8] Mantz Y A, Gervasio F L, Laino T and Parrinello M 2007 *Phys. Rev. Lett.* **99** 058104
[9] Case D A *et al* 2006 *AMBER 9* University of California, San Francisco
[10] de Pablo P J, Moreno-Herrero F, Colchero J, Herrero J G, Herrero P, Baró A M, Ordejón P, Soler J M and Artacho E 2000 *Phys. Rev. Lett.* **85** 4992
[11] Gervasio F L, Carloni P and Parrinello M 2002 *Phys. Rev. Lett.* **89** 108102
[12] Gervasio F L, Laio A and Parrinello M 2005 *Phys. Rev. Lett.* **94** 158103
[13] Duan Y *et al* 2003 *J. Comput. Chem.* **24** 1999
[14] Meeker K and Ellis A B 2000 *J. Phys. Chem.* B **104** 2500
[15] Zheng M, Jagota A, Semke E D, Diner B A, Mclean R S, Lustig S R, Richardson R E and Tassi N G 2003 *Nat. Mater.* **2** 338
[16] Zheng M *et al* 2003 *Science* **302** 1545
[17] Martin R M 2004 *Electronic Structure: Basic Theory and Practical Methods* (Cambridge: Cambridge University Press) The Edinburgh Building, Cambridge CB2 2RU, UK
[18] Kitaura K, Sawai T, Asada T, Nakano T and Uebayasi M 1999 *Chem. Phys. Lett.* **312** 319
[19] Kitaura K, Ikeo E, Asada T, Nakano T and Uebayasi M 1999 *Chem. Phys. Lett.* **313** 701

[20] Sekino H, Sengoku Y, Sugiki S and Kurita N 2003 *Chem. Phys. Lett.* **378** 589
[21] Fukuzawa K, Komeiji Y, Mochizuki Y, Kato A, Nakano T and Tanaka S 2006 *J. Comput. Chem.* **27** 948
[22] Ishikawa T *et al* 2006 *Chem. Phys. Lett.* **427** 159
[23] Goedecker S 1999 *Rev. Mod. Phys.* **71** 1085
[24] Soler J M, Artacho E, Gale J D, García A, Junquera J, Ordejón P and Sánchez-Portal D 2002 *J. Phys.: Condens. Matter* **14** 2745
[25] Ozaki T 2006 *Phys. Rev.* B **74** 245101
[26] Haynes P D, Skylaris C K, Mostofi A A and Payne M C 2006 *Phys. Status Solidi* b **243** 2489
[27] BigDFT project http://www-drfmc.cea.fr/sp2m/L_Sim/BigDFT/index.en.html
[28] Tsuchida E 2007 *J. Phys. Soc. Japan* **76** 034708
[29] Takayama R, Hoshi T, Sogabe T, Zhang S L and Fujiwara T 2006 *Phys. Rev.* B **73** 165108
[30] Gillan M, Bowler D, Torralba A and Miyazaki T 2007 *Comput. Phys. Commun.* **177** 14
[31] Miyazaki T, Bowler D R, Choudhury R and Gillan M J 2007 *Phys. Rev.* B **76** 115327
[32] Miyazaki T, Bowler D R, Ohno T and Gillan M J 2008 in preparation
[33] Goringe C M, Hernández E, Gillan M J and Bush I J 1997 *Comput. Phys. Commun.* **102** 1
[34] Todorovic M, Brazdova V, Torralba A, Choudhury R, Miyazaki T, Gillan M J and Bowler D R 2008 *J. Phys.: Condens. Matter.* submitted
[35] Ceperley D M and Alder B J 1980 *Phys. Rev. Lett.* **45** 566
[36] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[37] Hernández E, Gillan M J and Goringe C M 1997 *Phys. Rev.* B **55** 13485
[38] Horsfield A P and Bratkovsky A M 2000 *J. Phys.: Condens. Matter* **12** R1
[39] Ozaki T and Kino H 2004 *Phys. Rev.* B **69** 195113
[40] Li X P, Nunes R W and Vanderbilt D 1993 *Phys. Rev.* B **47** 10891
[41] Serena P A, Baratoff A and Soler J M 1993 *Phys. Rev.* B **48** 2046
[42] Clowney L, Jain S C, Srinivasan A R, Westbrook J, Olson W K and Berman H M 1996 *J. Am. Chem. Soc.* **118** 509
[43] Frisch M J *et al* 2004 *Gaussian 03, Revision B.05* (Wallingford CT: Gaussian, Inc.)
[44] Preuss M, Schmidt W G, Seino K, Furthmüller J and Bechstedt F 2004 *J. Comput. Chem.* **25** 112
[45] Machado M, Ordejón P, Artacho E, Sánchez-Portal D and Soler J M 1999 *Preprint* physics/9908022
[46] Kresse G and Furthmüller J 1996 *Phys. Rev.* B **54** 11169
[47] Perdew J P 1991 *Electronic Structure of Solids '91* (Berlin: Akademie-Verlag)
[48] Boys S F and Bernardi F 1970 *Mol. Phys.* **19** 553
[49] Kurita N, Danilov V I and Anisimovc V M 2005 *Chem. Phys. Lett.* **404** 164
[50] Fellers R S, Barsky D, Gygi F and Colvin M 1999 *Chem. Phys. Lett.* **312** 548
[51] Jurecka P and Hobza P 2003 *J. Am. Chem. Soc.* **125** 15608
[52] van der Wijst T, Guerra C F, Swart M and Bickelhaupt F M 2006 *Chem. Phys. Lett.* **426** 415
[53] Grimme S 2004 *J. Comput. Chem.* **25** 1463
[54] Miyazaki T, Bowler D R, Choudhury R and Gillan M J 2004 *J. Chem. Phys.* **121** 6186
[55] Danilov V I and Anisimov V M 2005 *J. Biomol. Struct.* **22** 471