Variationally optimized basis orbitals for biological molecules

T. Ozaki
Research Institute for Computational Sciences (RICS), National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

H. Kino
National Institute for Material Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

(Received 21 June 2004; accepted 27 July 2004)

Numerical atomic basis orbitals are variationally optimized for biological molecules such as proteins, polysaccharides, and deoxyribonucleic acid within a density functional theory. Based on a statistical treatment of results of a fully variational optimization of basis orbitals (*full* optimized basis orbitals) for 43 biological model molecules, simple sets of preoptimized basis orbitals classified under the local chemical environment (*simple* preoptimized basis orbitals) are constructed for hydrogen, carbon, nitrogen, oxygen, phosphorous, and sulfur atoms, each of which contains double valence plus polarization basis function. For a wide variety of molecules we show that the simple preoptimized orbitals provide well convergent energy and physical quantities comparable to those calculated by the full optimized orbitals, which demonstrates that the simple preoptimized orbitals possess substantial transferability for biological molecules. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1794591]

I. INTRODUCTION

Biological systems such as protein and deoxyribonucleic acid (DNA) are intrinsically large enough to make conventional first-principle calculations based on density functional theories (DFT) highly difficult, even if a massively parallel computer is used. Therefore, to extend the applicability of the DFT to biological molecules, considerable efforts have been devoted over the last decade.[1–11] A strategy to realize such large-scale DFT calculations is to use well localized basis orbitals[10–23] and to solve a resultant eigenvalue problem with sparse matrices by $O(N)$ methods.[3–8] Along this line, an important issue related to a trade-off between the computational accuracy and efficiency is how localized basis orbitals are constructed in the real space. Although a finite elements method[21] and a wavelet method[22,23] give ways of constructing the localized basis orbitals systematically, the size of the resultant Hamiltonian matrix is considerably large enough to hamper the realization of unified approaches with several $O(N)$ methods which have different convergence properties depending on the band gap of systems.[8,24] A possible choice to overcome this difficulty is the use of localized atomic basis orbitals the number of which is relatively small.[10–20] While it was thought that the lack of a systematic improvement is a serious drawback in this approach, recently, we have demonstrated that the accuracy and efficiency can be systematically controlled for a wide range of materials by adjusting two simple parameters: a cutoff radius and the number of orbitals.[10,11] Therefore, the localized atomic basis orbital is a practical choice for the realization of the large-scale DFT calculations coupled with $O(N)$ methods. Once the localized atomic basis orbitals are employed, it is desirable to reduce the number of basis orbitals as few as possible in terms of computational efficiency, while keeping a high degree of accuracy. In this sense, double valence plus polarization function (DVP) is regarded as a compromise between the computational accuracy and efficiency, since the DVP not only takes into account the environment dependence of orbitals for valence electrons, but also responds to the polarization of valence orbitals. However, it was still an ambiguous issue how to determine the radial shapes of basis orbitals even if we limit basis orbitals within the DVP. Recently, it has been shown that the radial shape of basis orbital can be variationally optimized by a simple orbital optimization method based on the force theorem.[10,11] The orbital optimization method enables us to automatically determine the radial shape so that the total energy can be minimized. Thus, the orbital optimization method gives a valid way of determining the radial shape of basis orbital in a given system. When the orbital optimization method is applied to biological molecules, it should be noted that preoptimized basis orbitals which are constructed for atoms located in the similar environment in advance could possess substantial transferability since the major part of a biological molecule is formed by inert parts which provide electrostatic potential and geometrical constraints for the chemically active parts. If the preoptimized orbitals possess an adequate transferability, it will be possible to assign preoptimized basis orbitals stored in the database to each atom based on a chemical sense similar to the assignation of empirical potentials. In addition, the construction of the database for the preoptimized orbitals would lead to a justification for the well-known fact that the character of each atom can be predicable based on functional groups in a molecule. Thus, our aim is to construct preoptimized basis orbitals for biological molecules so that the first-principles studies can be feasible for large-scale biomolecules with a considerable degree of accuracy. In this paper, we show that the orbital optimization method can be successfully applied to construct basis orbitals within the DVP for biological molecules, and that a small numbers of preopti-

mized basis orbitals can reproduce the total energy and physical quantities comparable to those calculated by the full optimized orbitals for a wide variety of molecules.

This paper is organized as follows: In Sec. II, we show the way of generating preoptimized atomic basis orbitals for biological molecules based on the full orbital optimization for 43 biological model molecules and a statistical treatment of the full optimized orbitals. In Sec. III, we demonstrate that the preoptimized orbitals possess substantial transferability for a wide variety of molecules. In Sec. IV, we conclude with discussing the applicability and the limitations of the basis set.

II. CONSTRUCTION OF OPTIMIZED ORBITALS

Based on the orbital optimization method[10,11] we construct preoptimized basis orbitals for biological molecules by the following two steps: (i) a full orbital optimization where all basis orbitals in a set of model molecules are variationally constructed; (ii) a simplification of the full optimized basis orbitals to construct a set of simple preoptimized basis orbitals by averaging them statistically.

Our construction of preoptimized basis orbitals is based only on the variational principles that the total energy is minimized with respect to the radial shape of basis orbitals. Therefore, it should be noted that no experimental result is employed in making of optimized basis orbitals.

In the first step we optimize all basis orbitals for a set of selected model molecules by the orbital optimization method coupled with the geometry optimization. In this optimization, ten steps of the orbital optimization are performed in every twenty steps of the geometry optimization by a steepest decent (SD) method with a variable prefactor for accelerating the convergence until the maximum magnitude of calculated force becomes below 10^{-4} (Hartree/Bohr), while the maximum step of the geometry optimization is limited up to 200 steps. The radial shape of basis orbitals of each atom in each molecule can vary differently to minimize the total energy in this orbital optimization. Therefore, we have the full set of basis orbitals differently optimized in all the model molecules. The set of optimized orbitals are referred to as the *full* optimized orbitals for the later discussion in this paper. In a series of the optimizations, the basis specifications are given in the abbreviation, which has been previously discussed,[10,11] as follows: H4.5-$s52^*p51^*$, C5.0-$s52^*p52^*d51^*$, N4.5-$s52^*p52^*d51^*$, O4.5-$s52^*p52^*d51^*$, S6.0-$s52^*p52^*d51^*$, P6.5-$s52^*p52^*d51^*$, and Na9.0-$s52^*p52^*d51^*$, where H, C, N, O, S, P, and Na indicate the atomic symbol, and the subsequent value gives a cutoff radius (a.u.) used in the generation of numerical primitive orbitals, thus indicating that the radial shape can vary within the cutoff radius. $s52^*$ means that two optimized s orbitals are constructed from five primitive orbitals, and * implies the restricted optimization that the radial shape of orbitals is independent on the magnetic quantum number m. Thus, this specification of basis orbitals implies a DVP. The primitive numerical orbitals are defined as the ground and excited states of an atom with the confinement potential as described in Refs. 10 and 11. It should be noted that the use of numerical orbitals is crucial for the efficient employment of the preoptimized orbitals, since a linear combination of the numerical orbitals is trivially transformed to a single numerical orbital by a single numerical table unlike analytic orbitals. Due to this benefit, no additional computational cost is required in the construction of Hamiltonian and overlap matrices for optimized orbitals compared to that for the primitive orbitals, while in case of analytic orbitals the computational effort depends on the number of primitive orbitals.[25–28] The cutoff radius of basis orbitals for each element is determined from the convergence of the total energy and bond length in a dimer molecule which is a severe test for the convergence with respect to basis orbitals because of the smallest number of neighboring atoms.[20] We find a trade-off between the computational accuracy and efficiency at the cutoff radii that we used in this study in the convergence properties in dimer molecules. To replace the deep core potential with a tractable shallow one, we use factorized norm conserving pseudopotentials[29,30] with multiple projectors.[31] The cutoff radii of pseudopotentials used in this study are found in Table I of Ref. 11. A relativistic correction is not included in the generation of pseudopotentials except for Pt. For the exchange correlation, a generalized gradient approximation (GGA)[32] is used with the nonlinear partial core correction (NLPCC)[33] except for a hydrogen atom. The real space grid techniques are used with the energy cutoff of 160 (Ryd) in numerical integration[18] and the solution of Poisson's equation using the fast Fourier transformation (FFT). All DFT calculations were performed using our DFT code, OpenMX.[34]

As model molecules of which basis orbitals are fully optimized, we consider 43 biological molecules, that is, 19 tripeptides for modeling of proteins with the amide linkage and N- and C-terminuses, five monomers of nucleosides, three dimers of nucleosides, a dimer of nucleotides for DNA and ribonucleic acid (RNA), seven disaccharides connected with a glycosidic bond for polysaccharides, three molecules for lipid, and five molecules including adenosine monophosphate (AMP), adenosine diphosphate (ADP) and adenosine triphosphate (ATP) for acid and nucleotide as listed in Table I. They contain 803 hydrogen, 494 carbon, 130 nitrogen, 270 oxygen, eight phosphorous, two sulfur, and seven sodium atoms as a whole. For amino acids with basic and acidic side chains tripeptides are constructed so that the total charge neutrality is maintained, while tripeptides terminated by N- and C-terminuses of glycine residue are considered for amino acids with nonpolar and uncharged polar side chains. In these tripeptides all amino and carboxyl groups are ionized. DNA and RNA are biological polymers formed by five kinds of nucleotides connected with phosphodiester linkage in a strand and with hydrogen bonds between two strands. Therefore, as a set of minimum models for DNA and RNA we consider five monomers, adenine (Ad), guanine (Gu), cytosine (Cy), thymine (Th), and uracil (Ur), of nucleosides being the building block, three kinds of dimers of nucleosides connected with hydrogen bonds, and a dimer of nucleotides connected by a phosphodiester linkage with an anhydrous sodium atom, while as the nucleotide dimer only a cytosine dimer are calculated for simplicity. For polysaccha-

TABLE I. Total energy and mean absolute deviations in the geometrical structure of 43 biological model molecules optimized by using three different sets of basis orbitals: the full optimized, simple optimized, and primitive basis DVPs. The total energies calculated by the primitive and simple optimized orbitals are given as the difference (Hartree/atom) for those calculated by the full optimized orbitals. The mean absolute deviation (MAD) between the full optimized and the other orbitals in optimized structures is calculated so that it can be minimized by varying six parameters: the relative position between centers of mass and relative Euler angles. In the calculation of the mean absolute deviation in bond length (MAD_{BL}) between the full optimized and the other orbitals, bond lengths below 2.2 Å are taken into account. Abbreviation for tripeptides follows the single notation of amino acid. $NaCy_2$ means a cytosine dimer connected by a phosphodiester linkage with an anhydrous sodium atom. In lipid molecules, DP and PC mean hydrogen terminated molecules corresponding to the tail and the head of DPPC. For all the saccharide molecules, disaccharide molecules with a glycosidic bond between C1 and C4 carbon atoms are considered. D-GlcpNAc and D-GlcpA means β-N-acetyl-D-glucosamine and β-D-glucuronic acid.

Molecule	No. of atoms	E_{tot} full opt. (Hartree)	ΔE_{tot} primitive (Hartree/atom)	ΔE_{tot} simple opt. (Hartree/atom)	MAD primitive (Å/atom)	MAD simple opt. (Å/atom)	MAD_{BL} primitive (Å/bond)	MAD_{BL} simple opt. (Å/bond)
Peptide								
GGG	24	−137.4707	0.0155	0.0006	0.0611	0.0041	0.0330	0.0034
GAG	27	−144.6684	0.0160	0.0022	0.0734	0.0089	0.0411	0.0043
GVG	33	−158.9803	0.0131	0.0006	0.0667	0.0081	0.0352	0.0048
GLG	36	−166.1907	0.0136	0.0018	0.0595	0.0070	0.0313	0.0043
GIG	36	−166.1367	0.0123	0.0005	0.0631	0.0118	0.0296	0.0055
GPG	31	−157.7752	0.0148	0.0017	0.0645	0.0142	0.0346	0.0087
GFG	37	−182.7426	0.0135	0.0009	0.0500	0.0113	0.0284	0.0063
GMG	34	−169.8930	0.0143	0.0016	0.0649	0.0051	0.0353	0.0038
GWG	41	−205.3539	0.0131	0.0008	0.0693	0.0175	0.0318	0.0071
GCG	28	−155.4893	0.0150	0.0009	0.0545	0.0277	0.0299	0.0126
DKG	44	−222.3680	0.0149	0.0018	0.0581	0.0147	0.0307	0.0072
DRG	46	−242.6496	0.0150	0.0020	0.0584	0.0132	0.0299	0.0063
DHG	40	−229.3943	0.0160	0.0022	0.0644	0.0151	0.0312	0.0063
EKG	47	−229.5418	0.0140	0.0015	0.0523	0.0176	0.0281	0.0065
GNG	31	−177.4721	0.0166	0.0018	0.0653	0.0285	0.0285	0.0106
GQG	34	−184.6284	0.0147	0.0006	0.0597	0.0164	0.0290	0.0056
GSG	28	−160.8116	0.0162	0.0019	0.0772	0.0125	0.0364	0.0076
GTG	31	−167.9436	0.0154	0.0017	0.0779	0.0271	0.0375	0.0078
GYG	38	−198.8777	0.0144	0.0016	0.0315	0.0155	0.0065	0.0048
D(R)NA								
Ad	31	−166.4850	0.0134	0.0025	0.0624	0.0332	0.0327	0.0088
Gu	32	−182.6085	0.0123	0.0001	0.0469	0.0310	0.0204	0.0073
Cy	29	−156.4758	0.0144	0.0021	0.0538	0.0058	0.0301	0.0038
Th	31	−169.0891	0.0147	0.0016	0.0513	0.0175	0.0304	0.0047
Ur	29	−178.0807	0.0164	0.0021	0.0641	0.0276	0.0363	0.0101
Ad-Th	62	−335.5943	0.0136	0.0018	0.0512	0.0330	0.0273	0.0103
Gu-Cy	61	−339.1560	0.0133	0.0015	0.0578	0.0153	0.0312	0.0057
Ad-Ur	61	−360.7095	0.0140	0.0012	0.0502	0.0100	0.0278	0.0057
$NaCy_2$	60	−380.4148	0.0145	0.0021	0.0561	0.0227	0.0286	0.0074
Saccharide								
arabinose	37	−215.8139	0.0150	0.0019	0.0383	0.0067	0.0252	0.0040
D-GlcpNAc	57	−310.1688	0.0132	0.0006	0.0477	0.0291	0.0268	0.0093
D-GlcpA	43	−292.3236	0.0165	0.0009	0.0531	0.0153	0.0298	0.0075
fructose	45	−262.3936	0.0134	0.0007	0.0485	0.0243	0.0302	0.0088
fucose	43	−230.1633	0.0133	0.0012	0.0488	0.0119	0.0309	0.0056
glucose	45	−262.5093	0.0156	0.0027	0.0437	0.0123	0.0282	0.0053
ribose	43	−215.8258	0.0127	0.0020	0.0502	0.0120	0.0315	0.0052
Lipid								
DP	104	−307.1726	0.0088	0.0010	0.0638	0.0181	0.0347	0.0072
PC	28	−127.2425	0.0111	0.0009	0.0516	0.0327	0.0320	0.0132
oleic acid	54	−160.0938	0.0090	0.0012	0.0549	0.0181	0.0294	0.0078
Acid								
citric acid	21	−153.6811	0.0196	0.0011	0.0551	0.0044	0.0262	0.0038
lactic acid	12	−69.9763	0.0163	0.0013	0.0642	0.0170	0.0448	0.0074
Others								
AMP	37	−267.5090	0.0133	0.0021	0.0546	0.0225	0.0271	0.0103
ADP	42	−352.2806	0.0150	0.0005	0.0381	0.0282	0.0223	0.0096
ATP	47	−437.1293	0.0167	0.0011	0.0613	0.0098	0.0269	0.0057
Av.			0.0143	0.0014	0.0567	0.0171	0.0288	0.0069

ride molecule, seven disaccharide molecules with a glycosidic bond, which are formed by typical 5- and 6-carbon sugars and sugar derivatives, are taken into account with the exception of the stereoisomers in our set of model molecules. Although there are isomers for disaccharides by the position of a glycosidic bond, among the isomers we select a disaccharide that bulky lateral chains are far from each other as much as possible so as to avoid the steric hindrance between

them. For lipid, dipalmitoyl phosphatidylcholine (DPPC), which is a dominant component in cell membranes, and oleic acids, which is a typical unsaturated fatty acid, are considered, where for saving computational time DPPC is divided at the bond connecting the chiral carbon atom and the head group into two hydrogen terminated molecules. In this paper the two molecules, corresponding to the tail and the head of DPPC, are referred to as DP and PC, respectively. In addition to these molecules, citric acid, lactic acid, AMP, ADP, and ADP, which are typical biological molecules, are included in our set of model molecules. Before the full optimization, the initial structures are optimized for the tripeptides, the nucleoside monomer and dimers, and the nucleotide dimer by a molecular mechanics (MM) using a software TINKER[35] with the AMBER98 force field,[36] and the disaccharides by a software SWEET2[37] with the MM3 force field.[38] The experimental structures for AMP,[39] ADP,[40] and ATP (Ref. 41) with one, two, and three anhydrous sodium atoms, respectively, and unoptimized structures created by a molecular modeling software MOLDA (Ref. 42) for the other molecules are used as the initial structures.

The total energy differences between the result of the full optimization, i.e., the simultaneous optimization for both the orbitals and geometry, and that of the geometry optimization using the primitive DVP with the same cutoff radii as the full optimized DVP are given in Table I. We see that the total energy is stabilized by about 0.014 (Hartree) per atom in the orbital optimization. Although the full optimized orbitals provide the minimum energy for the set of model molecules within DVP with the given cutoff radii, it is easy to imagine that application of the full optimized orbitals to new molecules suffers from some trouble in the assignment of basis orbitals, because there are very similar but different chemical environments in the model molecules.

In the second step of the construction of preoptimized orbitals, we therefore simplify the full optimized orbitals, i.e., construct a set of a small number of basis orbitals from the full optimized orbitals by analyzing them. For the later discussion we refer a set of optimized basis orbitals generated by a simplification of the full optimized orbitals as the *simple* preoptimized orbitals. Because the same cutoff radii are used in both the primitive and full optimized orbitals, the energy gain is attributed to the change in the radial shape of basis orbitals. In order to quantify the change in the radial shape of the basis orbitals we define a deviation index D by

$$D = d_1 d_2, \tag{1}$$

with

$$d_1 = \sqrt{\int (R_{\mathrm{opt}}^2 - R_{\mathrm{pri}}^2)^2 r^2 dr}, \tag{2}$$

$$d_2 = \int (rR_{\mathrm{opt}}^2 - rR_{\mathrm{pri}}^2) r^2 dr, \tag{3}$$

where R_{opt} and R_{pri} are radial functions of the full optimized and primitive orbitals. d_1 and d_2 give a total amount for the deviation between two orbitals, and the difference between a centroid of the radial function with a weight of r^2, respectively. Therefore, the absolute value of D means the degree of the change in the radial shape, and the negative sign and positive sign correspond to shrinking and expanding of orbital, respectively.

In Fig. 1 the deviation indices of hydrogen, carbon, nitrogen, and oxygen atoms in our model molecules are plotted against the effective charge of atom calculated by an electrostatic potential (ESP) fit method.[43] In this analysis, we use the ESP fit method, since we find that the ESP fit method gives more reasonable effective charges, fitting in well with a chemical sense, than Mulliken population analysis. In general hydrogen and oxygen atoms have positive and negative effective charges, respectively. In addition the dispersion of the deviation indices is relatively small, indicating the weak environment dependency of basis orbitals of these atoms in biological molecules. For the hydrogen atoms the first *s* and the first *p* orbital tend to shrink, while the second *s* orbitals expands by the orbital optimization. For the oxygen atoms both the effective charge and the deviation indices exhibit a highly small dispersion, and the deviation indices show that the orbital optimization causes large shrinking and expanding of the first and second *s* orbitals, respectively. These weak dependencies of basis orbitals on the chemical environment imply that a single set of optimized DVP is sufficient for hydrogen and oxygen atoms. Moreover, the same independence of the effective charge and the deviation index on the environment are found for phosphorous and sulfur orbitals as shown in Table II, while the same plot as Fig. 1 is not shown due to the paucity of sampling atoms. The deviation indices show that for phosphorous and sulfur atoms the orbital optimization largely affects the first *d* orbitals with a small standard deviation. Our analysis based on the deviation index supports that the full optimized basis orbitals are significantly simplified for each of hydrogen, oxygen, phosphorus, and sulfur atoms in biological molecules. Therefore, we decide to construct a single optimized DVP for each of hydrogen, oxygen, sulfur, and phosphorus atoms by a simple arithmetic average of all the full optimized orbitals in our model molecules, and refer to them as *simple* optimized orbitals. The simplification of the full optimized basis orbitals for hydrogen atom is less justified, since the dispersion of the deviation indices for hydrogen atom is relatively large compared to that of oxygen atom. However, it would be better to avoid a detailed classification of hydrogen atoms, since the chemical environment of hydrogen atom can be easily varied by a proton transfer in the hydrogen bonding in biological molecules. A justification for the simplification will be discussed by a comparison between the full and simple optimized orbitals in a practical way later on. In Fig. 2 the radial shape of the simple optimized orbitals and the primitive orbitals are shown. In fact, we find that the radial shape varies as previsioned from the deviation index. It can be pointed out that the changes of the first *s* and *p* optimized orbitals of hydrogen and oxygen atoms obey the charge state. The shrinking and expanding of the first *s* and *p* optimized orbitals of hydrogen and oxygen atoms are consistent with that they are charged up positively and negatively, respectively, while the second orbitals vary oppositely. The change of polarization functions of hydrogen, oxygen, phosphorous, and sulfur atoms is probably attributed to the correction of poor

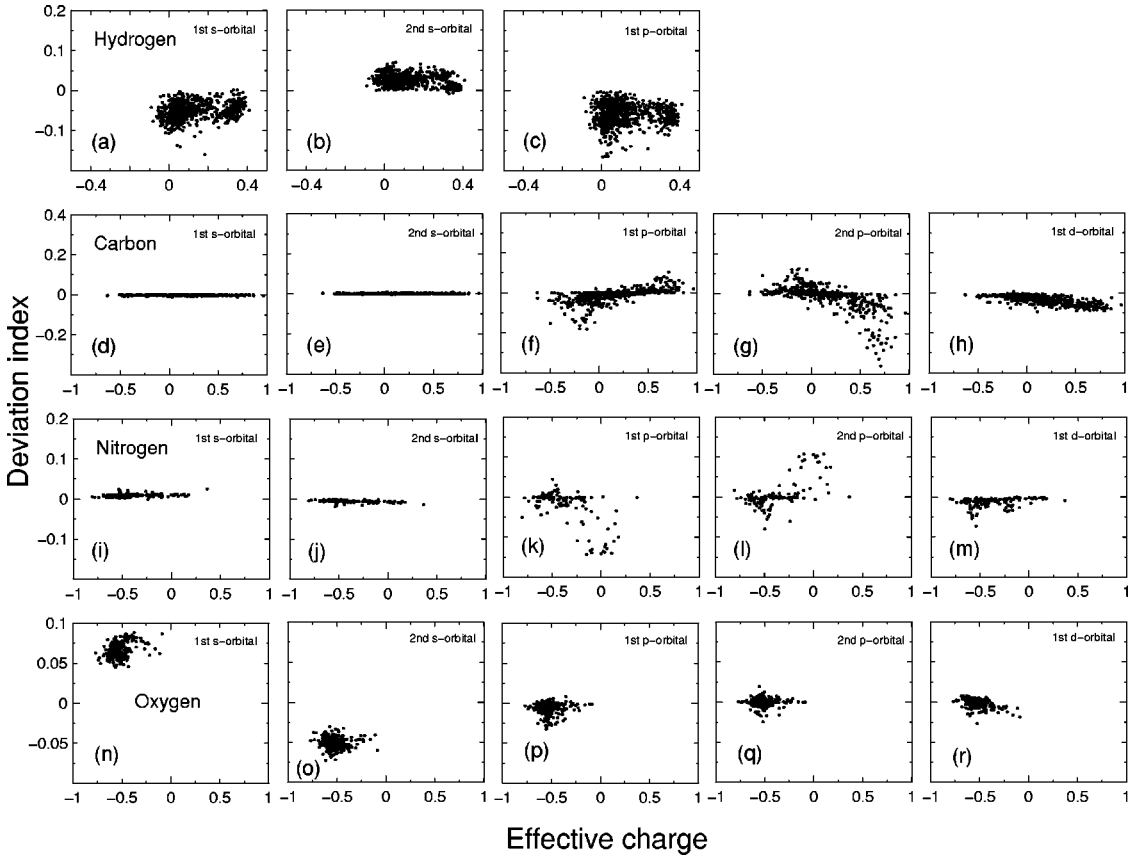

FIG. 1. The deviation index defined by Eq. (1) against the effective charge for (a)–(c) hydrogen, (d)–(h) carbon, (i)–(m) nitrogen, (n)–(r) oxygen atoms in 43 biological model molecules optimized by the full optimized DVP. The effective charge is calculated by an electrostatic potential (ESP) fit method in which sampling points are given by the grids in the real space between two shells of 1.0 and 2.0 times the van der Waals radius. In this grid generation, the volume per grid is 0.0153 (Bohr^3), which yields typically 20 000 points per molecule. In this ESP fit only the conservation of the total charge is considered as a constraint.

primitive *d* orbitals rather than the response to the charge state, since the primitive polarization functions being originally unbound states are calculated under the confinement potential.

On the other hand, both the basis orbitals of the carbon and nitrogen atoms possess widely dispersed and correlated effective charge and the deviation index each other, indicating that basis orbitals are differently optimized with a dependency on the chemical environment. The deviation index, especially for the *p* orbitals, is linearly dependent on the effective charge. However, it would be difficult to classify optimized basis orbitals based on only the effective charge because of the considerable fluctuation in the correlation. Therefore, a detailed analysis is needed for substantial classification of the full optimized basis orbitals. The average effective charge and deviation index of carbon atoms classified by the chemical environment are given in Table S-I of the E-PAPS supplemental material.[44] In this classification, the carbon atoms are distinguished by the neighboring atoms. First, carbon atoms bonding to oxygen atom are classified

TABLE II. Average effective charge and deviation index defined by Eq. (1) for hydrogen, oxygen, phosphorus, sulfur, and sodium in 43 biological model molecules calculated in the full optimized DVP. The effective charge is calculated by a electrostatic potential (ESP) fit method with the same condition as in the caption of Fig. 1. The standard deviation is also given in parentheses.

Species	No.	ESP charge	D_{s1}	D_{s2}	D_{p1}	D_{p2}	D_{d1}
H	803	0.1170	−0.0531	0.0262	−0.0593		
		(0.1202)	(0.0207)	(0.0145)	(0.0264)		
O	270	−0.5233	0.0663	−0.0510	−0.0066	−0.0004	−0.0011
		(0.0981)	(0.0094)	(0.0074)	(0.0075)	(0.0047)	(0.0053)
P	8	0.7344	−0.0003	−0.0007	−0.0077	0.0001	−0.1544
		(0.1586)	(0.0040)	(0.0039)	(0.0053)	(0.0014)	(0.0189)
S	2	−0.2641	0.0005	−0.0000	−0.0027	0.0029	−0.1329
		(0.0368)	(0.0002)	(0.0001)	(0.0001)	(0.0011)	(0.0554)
Na	7	0.7253	−0.0368	0.0238	0.0002	−0.0014	−0.0219
		(0.0311)	(0.0214)	(0.0167)	(0.0004)	(0.0007)	(0.0128)

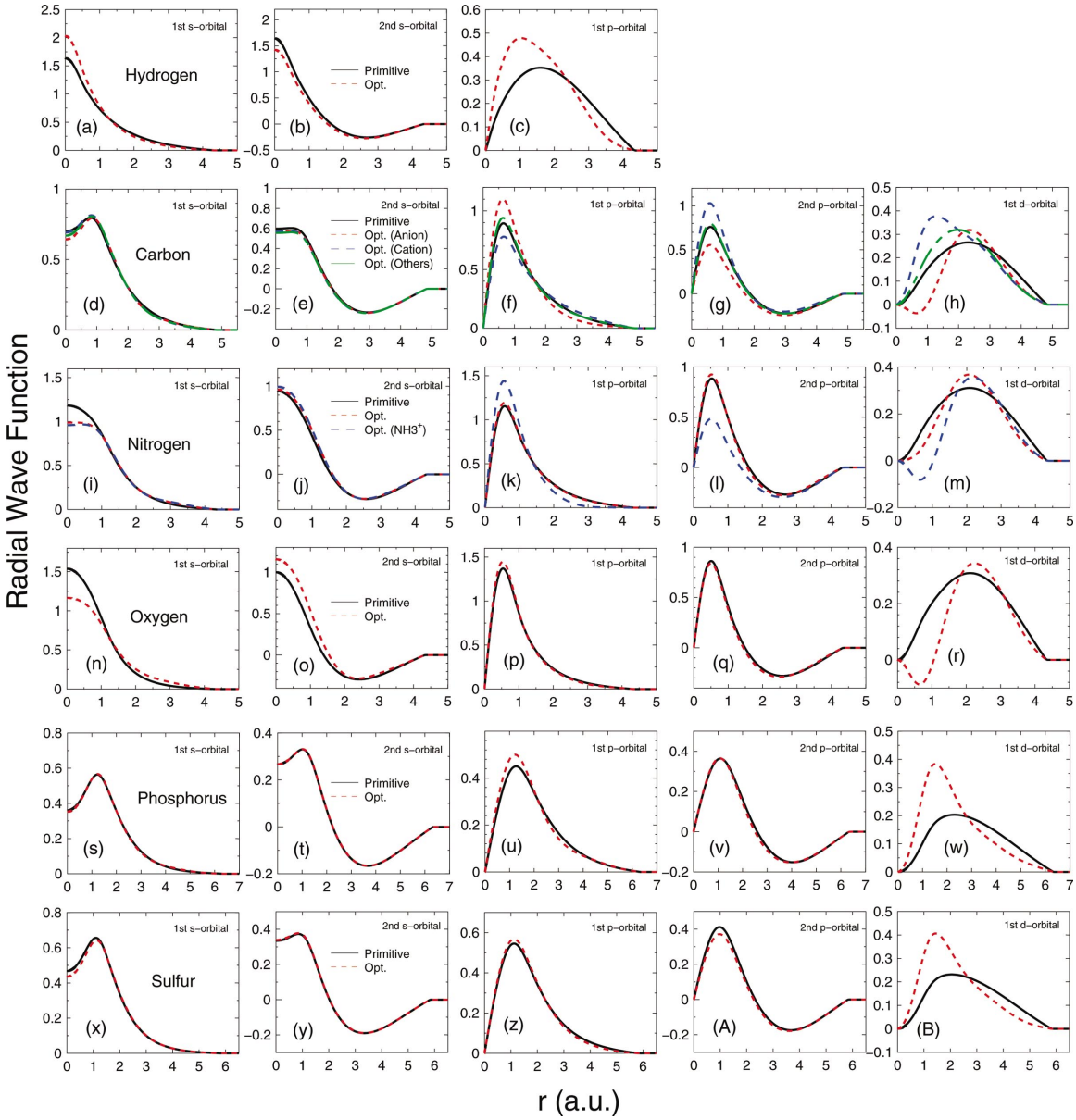

FIG. 2. (Color) Primitive and simple optimized radial wave functions of (a)–(c) hydrogen, (d)–(h) carbon, (i)–(m) nitrogen, (n)–(r) oxygen, (s)–(w) phosphorus, and (x)–(B) sulfur atoms. For hydrogen, oxygen, phosphorus, and sulfur atoms, the simple optimized orbitals are constructed by a simple arithmetic average of all the full optimized orbitals in our model molecules. The simple optimized orbitals of carbon and nitrogen atoms are generated by a simple arithmetic average in three and two species classified in Tables S-I and S-II of the E-PAPS supplemental material (Ref. 44), respectively.

because of the rigidity of oxygen atoms in the chemical environmental variation as discussed above. Then, the other carbon atoms are distinguished by the numbers of neighboring hydrogen and nitrogen atoms, and by the bonding form, sp_2 and sp_3. Although mainly the deviation index of carbon atoms depends on the effective charge, the detailed analysis shows that there are several exceptional cases such as C_{sp_2}-H_0N_3 of an arginine and C_{sp_2}-H_1N_0 of a benzene ring. Thus, we classify them to three species, *anion*, *cation*, and *others*, based on the deviation index rather than the effective charge. The classification of carbon atoms can be obviously depicted by a deviation index map shown in Fig. 3(a). In the deviation index map, averages D_{p1}, D_{p2}, and D_{d1} of each classified carbon atom are plotted for x, y, and z axes with the standard deviation. Because averages D_{s1} and D_{s2} of carbon atoms are relatively negligible compared to D_{p1}, D_{p2}, and D_{d1}, the axes for these deviation indices are excluded in this deviation index map. From the deviation index map, we see that the classification to three species can be easily justified for carbon atoms. The species CON looks to be classified to others, however, the side view given in the inset of Fig. 3(a) implies that it would be better to classify CON to cation. Practically the guess in the classification is supported by the fact that the total energy becomes lower when the species CON is included in cation as shown in Table III. Thus, we decide to classify the species CON to cation. This classification suggests that carbon atoms connected to oxygen atom with a double bond are strongly affected by the oxygen atoms, whereas those connected to the oxygen atom with a single bond such as COH and COR are

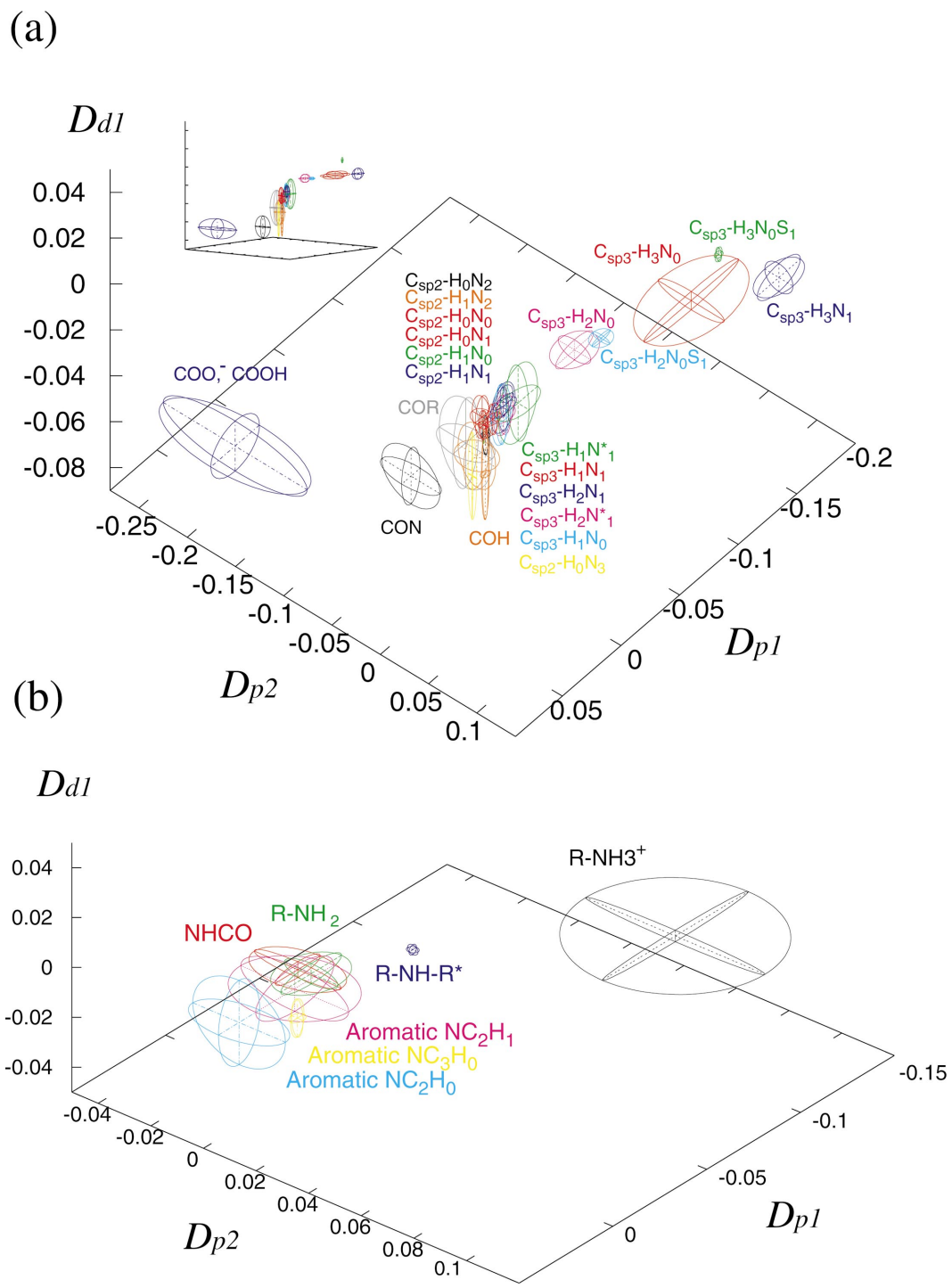

FIG. 3. (Color) Deviation index map of (a) carbon and (b) nitrogen atoms, where averages D_{p1}, D_{p2} and D_{d1} of each classified carbon and nitrogen atoms, listed in Tables S-I and S-II of the E-PAPS supplemental material (Ref. 44), are plotted for x, y and z axes with the standard deviation.

less influenced by the oxygen atom. For each of the three species, simple optimized DVPs are constructed by a simple arithmetic average of all the full optimized orbitals in the species, anion, cation, and others. We see that the radial shape significantly varies in the first and the second simple optimized p orbitals, and the first simple optimized d orbital as shown in Fig. 2. The large change of p orbitals in carbon atom could be related to a large change in the occupancy of electrons involved in p orbitals as discussed in our previous work.[11]

As that of carbon atoms, the average effective charge and deviation index of nitrogen atoms and orbitals classified by the chemical environment, and the deviation index map are shown in Table S-II of the E-PAPS supplemental material[44] and Fig. 3(b), respectively. In this deviation index map of nitrogen atom, axes of D_{s1} and D_{s2} are neglected due to their relatively small values. From Fig. 3(b), we find that the nitrogen atoms can be classified into two species, N and cation. A more detailed classification is not considered, since the classified nitrogen atoms, except for NH_3^+, considerably

TABLE III. Comparison of the total energy (Hartree) of selected molecules in the different classification of the species CON. The same abbreviations as those in Table I are used for molecules. The total energy is calculated by the geometry optimization with the same conditions as given in Sec. II.

Molecule	*Cation* with CON	*Cation* without CON
GGG	−137.4557	−137.4395
GLG	−166.1261	−166.1123
Cy	−156.4127	−156.4112
Th	−169.0390	−169.0250
Ur	−178.0191	−178.0067
b-D-GlcpNAc	−310.1321	−310.1313

overlap each other in Fig. 3(b). Therefore, we construct only two simple optimized DVPs for nitrogen atom by simple arithmetic averaging. As shown in Fig. 2, we see that the first *d* orbital of the species *N* and the *p* and *d* orbitals of the species cation largely vary by the orbital optimization. The large change of *p* orbitals, compared to the small change of *s* orbitals, might be attributed to the same origin as in the carbon atom.

As a consequence of the simplification for the full optimized orbitals, we generate a set of simple optimized DVP for hydrogen, oxygen, phosphorus, and sulfur atoms, and three and two sets of optimized DVPs for carbon and nitrogen atoms, which are available on our website.[34]

III. TRANSFERABILITY

In this section it is shown that the simple optimized orbitals possess a substantial transferability, although these optimized orbitals are generated from the full optimized orbitals by a significant simplification as discussed in Sec. II. Again for the 43 biological model molecules the geometry optimization is performed using the simple optimized orbitals with the same condition as described in Sec. II, and the total energy and geometrical parameters are compared with results by both the primitive and the full optimized orbitals. In Table I we see that the energy difference ΔE_{tot} of the simple optimized orbitals is about ten times smaller than that of primitive orbitals, indicating that the simple optimized orbitals almost span the same occupied subspace as the full optimized orbitals do. In accordance with the energy convergence, the geometrical parameters are also comparable to those of the full optimized orbitals. The average value of the mean absolute deviation (MAD) between optimized structures relative to that of the full optimized orbitals is 0.0567 and 0.0171 Å/atom for the primitive and the simple optimized orbitals, respectively. In addition to the MAD, the mean absolute deviation in bond length (MAD_{BL}) relative to that by the full optimized orbitals is shown in Table I. We see that the MAD_{BL} of the simple optimized orbitals is about four times smaller than that of the primitive orbitals. The average value of MAD_{BL} suggests that the difference in bond length is only 0.007 Å compared to that calculated by the full optimized orbitals. Thus, we find that, in all the model molecules without any exceptional case, the simple optimized orbitals yield convergent geometries more than the primitive orbitals, which are comparable to those of the full optimized orbitals with the difference of about 0.007 Å in bond length.

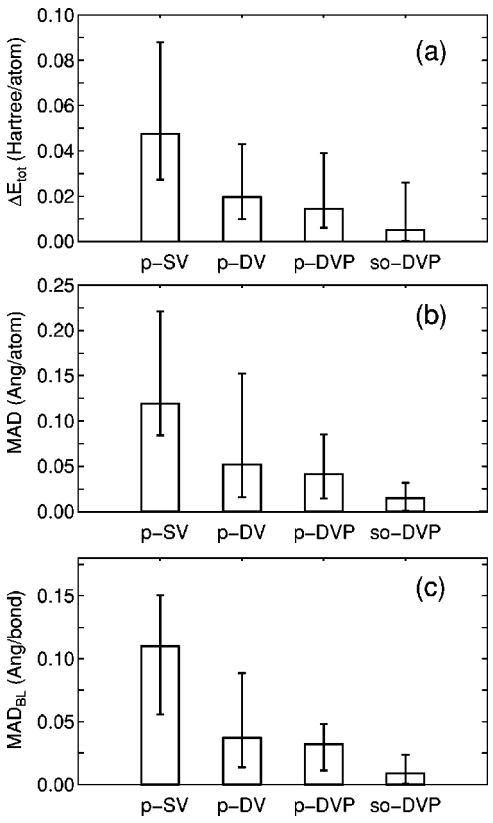

FIG. 4. (a) The average error ΔE_{tot} in the total energy, (b) the average MAD, and (c) the average MAD_{BL}, relative to those of the fo-DVP, of 31 molecules including small typical and biological molecules calculated by five different basis sets: p-SV, p-DV, p-DVP, so-DVP, and fo-DVP. In the abbreviation of basis orbitals, p, so, and fo mean the primitive, the simple optimized, and the full optimized orbitals, respectively, and SV, DV, and DVP represent single valence orbitals, double valence orbitals, and double valence orbitals plus a polarization function, respectively. The error bar gives the maximum and minimum values in those of the 31 molecules. The MAD_{BL} was calculated under the same condition as given in the caption of Table I. For all the list of the 31 molecules, see Table S-III of the E-PAPS supplemental material (Ref. 44).

Moreover, in Table S-III of the E-PAPS supplemental material[44] we show the total energy, geometrical parameters, dipole moment, binding energy, and activation energy for rotational barrier of 31 molecules including small typical and biological molecules calculated by five different basis sets, p-SV, p-DV, p-DVP, so-DVP, and fo-DVP in order to demonstrate the transferability of the simple optimized basis orbitals. The full optimization for the fo-DVP are performed in the same condition as in Sec. II. Without any exceptional case the total energy decreases in order of the p-SV, p-DV, p-DVP, so-DVP, and fo-DVP in all the molecules we calculated. In addition, it should be noted that the total energy of the so-DVP is significantly close to that of the fo-DVP, which clearly demonstrates that the so-DVP basis set has substantial transferability for a wide variety of molecules, even though the so-DVP is generated from the full optimized orbitals by the simple statistical treatment. We also show the average error in the total energy, the average MAD, and the average MAD_{BL} relative to those of the fo-DVP in Fig. 4 to easily verify the transferability of so-DVP, which in fact illustrates the so-DVP is comparable to the fo-DVP. It can be

confirmed in comparison of a methane, ethane, ethylene, acetylene, and benzene molecules that the so-DVP provides well convergent geometrical parameters involved in hydrogen and carbon atoms. These examples show that the so-DVP has an adequate applicability for a wide range of bonding natures, while the bond between carbon atoms varies through single, double, and triple bonds in these molecules. As that of hydrogen and carbon atoms, well convergent geometrical parameters involved in oxygen and nitrogen atoms are also obtained by the so-DVP. In all the cases geometrical parameters calculated by the so-DVP are comparable to those by the fo-DVP. However, it should be mentioned that the so-DVP and the fo-DVP do not give fully convergent results for the double bond involved in oxygen and/or nitrogen atoms, which are confirmed in an oxygen molecule, formaldehyde, formamide, nitrogen dioxide NO_2, and sulfuric acid H_2SO_4. This relatively poor convergence can be attributed to the use of a single polarization function rather than the simplification of full optimized orbitals, because many polarization functions are required to achieve a fully convergent result for representative elements in the right side of the periodic table as shown in our previous study.[11] Thus, we find that it is difficult to obtain a fully convergent result for these elements within the DVP even though the radial shapes are fully optimized. For phosphorus and sulfur atoms, the so-DVP yields substantial improvements of primitive orbitals. The improvement can be seen in PH_3, H_2S, H_2S_2, H_2SO_4 and thioformamide, while the so-DVP does not reach to a fully convergent result for sulfur atom yet because of the same reason as the case of oxygen atom. Although there is no disulfide bond in the 43 model molecules discussed in Sec. II, we see that the disulfide bond in H_2S_2 calculated by the so-DVP is relatively comparable to that by the fo-DVP. As a general trend in the calculated bond length the poor basis sets tend to predict a longer bond length, while the bond length converges to the experimental value with an error of a few percentages as the level of basis set increases. On the other hand, the poor basis sets tend to underestimate the bond angle, while there are exceptional cases.

Due to the importance of hydrogen bonding in biological systems, the transferability for description of the hydrogen bonding is investigated by several model systems: a water dimer $(H_2O)_2$, an acetic acid dimer, guanine-cytosine pair (Gu-Cy), and adenine-thymine pair (Ad-Th) with intermolecular hydrogen bonds, and a maleic acid molecule with an intramolecular hydrogen bond. As an overall feature we find that the so-DVP gives substantial convergent results for geometrical parameters and the binding energy as well as the covalent bonds. However, the bond lengths involved in the hydrogen bond are underestimated by about 0.05 Å even in the use of the fo-DVP, and the binding energy of hydrogen bonding tends to be overestimated compared to the other theories and the experimental values. The deviation might be attributed to both the shorter cutoff radii of basis orbitals and the GGA to the exchange-correlation potential. Since it has been reported that the use of diffuse orbitals is needed to accurately describe the hydrogen bonding,[45] the convergence properties for basis orbitals with a longer tail should be investigated. A study is being done to clarify the relationship between the cutoff radius of basis orbitals and the description of hydrogen bonding, and to find a compromise between the computational accuracy and efficiency. The details will be presented elsewhere. In contrast to the convergence properties of the covalent bond length, we see an opposite dependency in the convergence properties of the bond length involved in the hydrogen bonding. As the level of basis set increases, the bond length involved in the hydrogen bonding becomes longer, while the covalent bond length tends to shorten.

In the actual applications of the simple optimized orbitals to biological systems, the simple optimized orbitals will be often used together with the primitive orbitals, since other representative elements and transition metals can be constituents in biological molecules. In such cases, one may suspect the transferability of the simple optimized orbitals. Therefore, in order to study the capability of the simple optimized orbitals used together with the primitive orbitals, we calculate three molecules, a monofluoromethane molecule CH_3F, a cisplatin molecule, and a carboplatin molecule. The primitive orbitals for fluorine atom (F5.0-s2p2d1) and platinum atom (Pt7.5-s3p3d2f1) are used in the calculations of these molecules by the so-DVP, where the abbreviation of basis orbitals are given in parentheses. As expected, the total energy and the geometrical parameters calculated by the so-DVP are close to those by the fo-DVP. Therefore, these examples clearly illustrate that the transferability of the so-DVP remains even though the so-DVP is used together with the primitive orbitals. Although the bond length between carbon and fluorine atoms in CH_3F is overestimated compared to the other theoretical and experimental values even if the fo-DVP is used, this overestimated bond length comes from the same reason as the case of the oxygen atom, which again suggests that higher quality of basis orbitals more than DVP for representative elements of the sixth and seventh group is required for the full convergence.

IV. CONCLUSIONS

To conclude, we have successfully generated numerical atomic basis orbitals for biological molecules such as proteins, polysaccharides, and deoxyribonucleic acid (DNA) using the orbital optimization method based on the force theorem within the density functional theory. Following the fully variational optimization of basis orbitals for 43 biological model molecules, a few sets of simple preoptimized basis orbitals are constructed by a detailed classification of full optimized orbitals using a deviation index and by a statistical treatment without employing any experimental results. Moreover, the transferability of the simple preoptimized orbitals is demonstrated by calculating the total energy, geometrical parameters, dipole moment, binding energy, and activation energy for rotational barrier of 31 molecules including small typical and biological molecules calculated using five different basis sets: p-SV, p-DV, p-DVP, so-DVP and fo-DVP. We find that the so-DVP gives substantial convergent results for a wide variety of molecules without any exceptional case in all the molecules we have studied. For hydrogen and carbon atoms the so-DVP provides quite satisfactory convergent results comparable to other theoretical

and experimental values. On the other hand, for nitrogen, oxygen, and, sulfur atoms the so-DVP gives less convergent results compared to the other theoretical and experimental values, while the results by the so-DVP are comparable to those by the fo-DVP even in these elements as well as hydrogen and carbon atoms. The less convergent results suggest that more polarization functions are required to achieve a full convergence with respect to basis orbitals for these elements in the fifth, sixth, and seventh group of the periodic table because of strong polarization of those wave functions. From the results of the molecules including hydrogen bonding we find that a further tuning is needed for the cutoff radius of basis orbitals to accurately describe the weak hydrogen bonding, since the so-DVP and the fo-DVP tend to underestimate the bond length involved in the hydrogen bonding by about 0.05 Å compared to other theoretical and experimental values, although a part of the poor description of hydrogen bonding may be attributed to the GGA. However, The results of the fo-DVP suggests that the less convergence for the elements in the fifth, sixth, and seventh group and the underestimation of bond length involved in the hydrogen bonding should not be attributed to the simplification of the full optimized basis orbitals, but a smaller number of basis orbitals and shorter cutoff radii in the DVP. The further improvement of basis orbitals will be discussed elsewhere to accurately describe the elements in the fifth, sixth, and seventh group, and the hydrogen bonding. In spite of these two difficulties, our optimization scheme for basis orbitals is regarded as an attempt to extend applicability of DFT calculations to large-scale systems by making full use of the commonality in the chemical environment of constituent atoms in specific systems such as biological systems. In fact, we demonstrated that the so-DVP yields substantial improvement of the primitive orbitals set, indicating that a basis set special to biological systems are capable of making DFT calculations in realistic large-scale systems feasible with a considerable degree of accuracy. Thus, we conclude that our variational optimization of basis orbitals is a valid way of constructing a specially purposed basis orbital set and that the preoptimized orbitals so-DVP are applicable to a wide variety of biological molecules with a considerable degree of accuracy, while our so-DVP has two difficulties to overcome.

ACKNOWLEDGMENTS

The authors would like to thank S. Tsuzuki for helpful comments on transferability of optimized orbitals. One of us (T.O.) was partly supported by NEDO under the Nanotechnology Materials Program, Research and Development for Applying Advanced Computational Science and Technology of Japan Science and Technology Corporation (ACT-JST), and NAREGI Nanoscience Project, Ministry of Education, Culture, Sports, Science, and Technology, Japan.

[1] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
[2] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992) and references therein.
[3] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).
[4] S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).
[5] G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).
[6] X.-P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993); M. S. Daw, *ibid.* **47**, 10895 (1993).
[7] T. Ozaki and K. Terakura, Phys. Rev. B **64**, 195126 (2001).
[8] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999) and references therein.
[9] K. Kitaura, T. Sawai, T. Asada, T. Nakano, and M. Uebayasi, Chem. Phys. Lett. **312**, 319 (1999).
[10] T. Ozaki, Phys. Rev. B **67**, 155108 (2003).
[11] T. Ozaki and H. Kino, Phys. Rev. B **69**, 195113 (2004).
[12] O. F. Sankey and D. J. Niklewski, Phys. Rev. B **40**, 3979 (1989).
[13] A. A. Demkov, J. Ortega, O. F. Sankey, and M. P. Grumbach, Phys. Rev. B **52**, 1618 (1995).
[14] J. P. Lewis, P. Ordejon, and O. F. Sankey, Phys. Rev. B **55**, 6880 (1997).
[15] D. Sanchez-Portal, P. Ordejon, E. Artacho, and J. M. Soler, Int. J. Quantum Chem. **65**, 453 (1997).
[16] W. Windl, O. F. Sankey, and J. Menendez, Phys. Rev. B **57**, 2431 (1998).
[17] S. D. Kenny, A. P. Horsfield, and H. Fujitani, Phys. Rev. B **62**, 4899 (2000).
[18] J. Junquera, Ó. Paz, D. Sánchez-Portal, and E. Artacho, Phys. Rev. B **64**, 235111 (2001); J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez-Portal, J. Phys.: Condens. Matter **14**, 2745 (2002) and references therein.
[19] C. K. Gan, P. D. Haynes, and M. C. Payne, Phys. Rev. B **63**, 205109 (2001).
[20] J. D. Talman, Phys. Rev. Lett. **84**, 855 (2000).
[21] E. Tsuchida and M. Tsukada, Phys. Rev. B **54**, 7602 (1996); E. Tsuchida and M. Tsukada, J. Phys. Soc. Jpn. **67**, 3844 (1998).
[22] K. Cho, T. A. Arias, J. D. Joannopoulos, and P. K. Lam, Phys. Rev. Lett. **71**, 1808 (1993).
[23] T. A. Arias, Rev. Mod. Phys. **71**, 267 (1999).
[24] D. R. Bowler, M. Aoki, C. M. Goringe, A. P. Horsfield, and D. G. Pettifor, Modell. Simul. Mater. Sci. Eng. **5**, 199 (1997).
[25] S. F. Boys, Proc. R. Soc. London, Ser. A **200**, 542 (1950); **201**, 125 (1950); **258**, 402 (1960).
[26] W. J. Hehre, I. Radom, P. V. R. Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory* (Wiley, New York, 1986).
[27] S. Huzinaga, *Gaussian Basis Sets for Molecular Calculations* (Elsevier, Amsterdam, 1984).
[28] T. H. Dunning, Jr. and P. J. Hay, *in Methods of Electronic Structure Theory*, edited by H. F. Schaefer III (Plenum, New York, 1977), p. 1.
[29] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).
[30] L. Kleinman and D. M. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).
[31] P. E. Blochl, Phys. Rev. B **41**, 5414 (1990).
[32] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[33] S. G. Louie, S. Froyen, and M. L. Cohen, Phys. Rev. B **26**, 1738 (1982).
[34] Our DFT code, OpenMX, the basis orbitals, and pseudo potentials used in this study are available on a web site (http://staff.aist.go.jp/t-ozaki/) in the constitution of the GNU General Public Licence.
[35] P. Ren and J. W. Ponder, J. Comput. Chem. **23**, 1497 (2002).
[36] D. A. Pearlman *et al.*, Comput. Phys. Commun. **91**, 1 (1995); D. A. Case, D. A. Pearlman, J. W. Caldwell *et al.*, AMBER6, University of California, San Francisco, 1999.
[37] A. Bohne, E. Lang, and C. W. von der Lieth, J. Mol. Model. [Electronic Publication] **4**, 33 (1998); A. Bohne, E. Lang, and C. W. von der Lieth, Bioinformatics **15**, 767 (1999).
[38] N. L. Allinger, Y. H. Yuh, and J. H. Lii, J. Am. Chem. Soc. **111**, 8551 (1989); J. H. Lii and N. L. Allinger, *ibid.* **111**, 8566 (1989); **111**, 8576 (1989).
[39] T. Nakatsu, H. Kato, and J. Oda, 12AS (PDB ID), the Protein Data Bank (PDB).
[40] D. R. Tomchick and J. L. Smith, 1AO0 (PDB ID), the Protein Data Bank (PDB).
[41] H. S. Subramanya, A. J. Doherty, S. R. Ashford, and D. B. Wigley, 1A0I (PDB ID), the Protein Data Bank (PDB).
[42] H. Yoshida and H. Matsuura, J. Chem. Software **3**, 147 (1997).
[43] U. C. Singh and P. A. Kollman, J. Comput. Chem. **5**, 129 (1984); L. E. Chirlian and M. M. Francl, *ibid.* **8**, 894 (1987); B. H. Besler, K. M. Merz Jr., and P. A. Kollman, *ibid.* **11**, 431 (1990).
[44] See EPAPS Document No. E-JCPSA6-121-301439 for three tables and references. A direct link to this document may be found in the online article's HTML reference section. The document may also be reached via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html) or from ftp.aip.org in the directory /epaps/. See the EPAPS homepage for more information.
[45] B. J. Lynch, Y. Zhao, and D. G. Truhlar, J. Phys. Chem. A **107**, 1384 (2003).