Electronic-structure-based molecular-dynamics method for large biological systems: Application to the 10 basepair poly(dG)·poly(dC) DNA double helix

James P. Lewis*
Department of Physics and Astronomy, Box 871504, Arizona State University, Tempe, Arizona 85287-1504

Pablo Ordejón
Departmento de Física, Falcultad de Ciencas, Universidad de Oviedo, 33007 Oviedo, Spain

Otto F. Sankey
Department of Physics and Astronomy, Box 871504, Arizona State University, Tempe, Arizona 85287-1504

(Received 23 August 1996)

Combining several recently developed theoretical techniques, we have developed an electronic-structure-based method for performing molecular-dynamical simulations of large biological systems. The essence of the method can be summarized in three points: (i) There are two energy scales in the Hamiltonian and each is treated differently—the strong intramolecular interactions are treated within approximate density-functional theory, whereas the weak intermolecular interactions (e.g., hydrogen bonds) are described within a simple theory that accounts for Coulomb, exchange, and hopping interactions between the weakly interacting fragments. (ii) A localized basis of atomic states is used, yielding sparse Hamiltonian and overlap matrices. (iii) The total energies and forces from the sparse Hamiltonian and overlap matrices are solved using a linear scaling technique to avoid the N^3 scaling problem of standard electronic structure methods. As an initial benchmark and test case of the method, we performed calculations of a deoxyribonucleic acid (DNA) double-helix poly(dG)·poly(dC) segment containing ten basepairs, with a total of 644 atoms. By a dynamical simulation, we obtained the minimum-energy geometry and the electronic structure of this DNA dehydrated segment, as well as the full dynamical matrix corresponding to the relaxed structure. The vibrational data and energy band gap obtained compare qualitatively well with previous experimental data and other theoretical results. [S0163-1829(97)04512-8]

I. INTRODUCTION

In the early years of molecular quantum mechanics, the prospect of understanding the microscopic properties of a biological system from first-principles quantum mechanics was perhaps overestimated. Decades later we find that *ab initio* approaches have made an *insignificant* contribution in the field of large biological systems. Due to ‘‘unfulfilled promises’’ most biologists (including biochemists, biophysicists, etc.) are inherently skeptical about deriving practical results from first principles. As such, empirical atomic or intermolecular potential methods and semiempirical quantum-mechanical methods became the method of choice for simulating large biological systems.

Empirical methods have become quite popular among biologists, which is evident by the widespread use of programs such as AMBER (Ref. 1) and CHARMM,[2] which have proven successful for reducing the computational time required in simulations with a large number of atoms. They have also accurately demonstrated their ability to energy minimize structural conformations to those which are in general agreement with experimental results and biological theory. This is not unexpected since the parameters of empirical as well as semiempirical methods are fit to a large database resulting from many experiments.

Although empirical methods are sufficient to answer many structural and thermodynamic questions, biologists see the need to include more than just two-body interactions explicitly.[3] Many effects, such as the electronic polarizability, are inadequately described by only two-body interactions, but including many-body potentials complicates the empirical model significantly because of the increasing number of parameters. Many-body effects are best described by quantum-mechanical methods, since they are inherently many body; in addition, they require no fitting of parameters.

In dealing with biological systems, it is desirable to study more than just the structural and thermodynamical questions. Certain microscopic properties of a system such as electron transport, charge densities, electronic structure, and other optical and electronic properties are naturally described using an electronic structure method. An example of the importance of studying microscopic properties in biological systems is found in the chlorophylls. In these systems, light energy is absorbed, promoting electrons to higher orbitals, and enhancing the potential for transfer of these electrons to suitable acceptors. Studying the electron-transfer properties of these molecules can only be accomplished via an understanding of the microscopic (quantum) properties of such systems. In another example, it has been suggested that there is a correlation between the electronic structure of deoxyribonucleic acid (DNA) and proteins and their biological functions.[4] Quantum-mechanical methods are then necessary to investigate the microscopic properties desired in each of these two examples.

One must appreciate the enormous hurdles that must be overcome in applying electronic structure methods to biological systems. At the atomic level, there are four major types of interactions in biological systems—covalent, ionic, hydrogen bonding, and van der Waals. In order to simulate such systems from first principles successfully, a well-described electronic structure method must be developed to account correctly for each type of interaction. In electronic structure theory, the largest concentration of research efforts has been in the development of methods to describe correctly the strong intramolecular covalent and ionic interactions. Many methods have been created as a result of the materials theory research thrust from the 1970s to the present (for a review of such methods see Refs. 5, 6, and 7). In particular, the method of Sankey and Niklewski,[8] which is used as the basis for modeling intramolecular interactions in this paper, has been shown to work well for a plethora of systems. In dealing with the more subtle and difficult weak interactions, attempts have also been made to derive simple schemes for hydrogen-bonded systems. The weak intermolecular interactions in this paper are modeled using a simplified electronic structure method that was previously developed and applied to the interactions between water molecules[9] and between DNA triplets.[10]

A huge roadblock to simulating large biological molecules from first principles has been the enormous computational intensity required. The electronic structure community is keenly aware of these difficulties, and many attempts have been made to develop techniques which surmount some of these difficulties by using physically motivated approximations. Using a localized basis set reduces the computer requirements from the supercomputer level down to the workstation level for simulations involving no more than 200 atoms. However, diagonalization of the Hamiltonian, which is required to obtain the electronic eigenvalue (band structure) energy, is an $O(N^3)$ computational algorithm, where N is the number of electron orbitals. The simulation of more than a few hundred atoms using first-principles techniques seems improbable with this type of scaling. An important redeeming factor of the localized basis set is that the Hamiltonian is sparse. Recently, several techniques have been developed to take advantage of this sparseness and to solve for the band structure energy using linear scaling algorithms (commonly referred to as order N).[11–23]

An electronic structure, based on the Sankey-Niklewski method with improvements to include a hydrogen-bonding model and a linear system-size scaling algorithm, has been developed for the purpose of doing molecular-dynamical simulations of large biological systems. As an introductory benchmark and test case of this method, we consider a segment of DNA containing ten basepairs of poly(dG)$\cdot$poly(dC). In general, the DNA molecule is formed of two strands coiled about one another, yielding the familiar double-stranded helix. The outer edges of each strand are formed by a sugar-phosphate backbone which follows the helical path. Attached to each sugar along the backbone is a base unit (adenine, guanine, thymine, or cytosine) which is roughly perpendicular to the axis of the helix. The binding forces within each strand come from strong covalent intramolecular interactions. The two strands are hydrogen bonded across the base on one strand to a base of the other, forming the purine-to-pyrimidine complementary basepairs $A\cdot T$ and $G\cdot C$. Hence, within the DNA molecule, as in most biological systems, there are (at least) two energy scales to consider: strong intramolecular interactions, such as covalent bonding, and weak intermolecular interactions, such as hydrogen bonding. This factor, plus the size of the system, makes the DNA molecule an ideal candidate for performing a benchmark test of the techniques which will be presented in this paper.

The organization of this paper is as follows. After discussing in Sec. II the electronic structure based method for calculating large biological systems, results are given for the 10 basepair poly(dG)$\cdot$poly(dC) DNA segments in Sec. III. Finally, in Sec. IV a summary of our results are given and prospective work is discussed.

II. AN ELECTRONIC-STRUCTURE-BASED MOLECULAR-DYNAMICS METHOD FOR LARGE BIOLOGICAL SYSTEMS

A. Electronic Hamiltonian

Many biological systems, such as the DNA double helix, consist of two or more weakly interacting fragments where within each fragment strong intramolecular interactions are present. To describe both the weak and strong interaction energy scales simultaneously with sufficient accuracy, we conveniently characterize the Hamiltonian in terms of a sum of strong intramolecular and weak intermolecular components. This allows the use of two different models of calculation, where each model is appropriately formulated for each type of interaction. The technique of splitting the total energy into intramolecular and intermolecular interactions was used previously by others such as Harris[6] and Gordon and Kim.[24] For two weakly interacting fragments, we write the total single-particle local orbital Hamiltonian $\mathcal{H}$ as

$$\begin{aligned}\mathcal{H} &= \mathcal{H}^{\text{strong}} + \mathcal{H}^{\text{weak}}\\ &= \begin{pmatrix} H_{11}^{\text{strong}} & 0 \\ 0 & H_{22}^{\text{strong}} \end{pmatrix} + \begin{pmatrix} \delta H_{11}^{\text{weak}} & h_{12}^{\text{weak}} \\ h_{21}^{\text{weak}} & \delta H_{22}^{\text{weak}} \end{pmatrix}. \end{aligned} \tag{1}$$

The submatrices H_{11}^{strong} and H_{22}^{strong} represent the intramolecular interactions within fragments 1 and 2, respectively, $\delta H_{11}^{\text{weak}}$ and $\delta H_{22}^{\text{weak}}$ are shifts of the intramolecular interactions due to the intermolecular electrostatic, exchange and overlap interactions, and h_{21}^{weak} and h_{12}^{weak} are the intermolecular ''hopping'' matrix elements between the two fragments.

The intramolecular interactions are determined from a first-principles local-orbital method developed by Sankey and Niklewski.[8] This is done using the Harris functional[6] within the local-density approximation (LDA), and using the pseudopotential approximation.[25] The electronic eigenstates are expanded as a linear combination of pseudoatomic orbitals within a localized sp^3 basis for carbon, nitrogen, oxygen, and phosphorous, and an s-basis for hydrogen. This method has been applied to many covalent systems, and has proven to be computationally fast and quantitatively accurate (see (Refs. 26 and 27 and references therein).

To calculate weak intermolecular interactions such as in hydrogen-bonded systems, we use a method that was previously developed and discussed in Ortega, Lewis, and Sankey,[9] which evolved from earlier work.[28] This method

TABLE I. Decay constants used in the Slater-type orbitals. Only the valence orbitals are used, so $1s$ for H, and $2s$ and $2p$ for C, N, and O.

Atom	ζ_s	ζ_p
H	1.27	
C	2.45	1.75
N	2.76	1.95
O	3.23	2.25

was shown to work well for water and for isolated DNA basepairs and triplets.[10] In the calculation of all the intermolecular interactions, the atomic orbitals of the basis set are assumed to be of Slater type, since this is the expected asymptotic shape in the intramolecular regions. These orbitals are of the form

$$\psi_\alpha^0=\psi_{nlm}=N_{nlm}r^{n-1}Y_l^m(\hat{r})e^{-\zeta r}, \tag{2}$$

where n, l, and m are the appropriate quantum numbers, the functions $Y_l^m(\hat{r})$ are the spherical harmonics, and N_{nlm} is the normalization constant. The decay constants ζ for the p orbitals were obtained from Hehre, Stewart, and Pople,[29], which lists optimum exponents for several molecules. Note the approximation that the decay constant falls off as the square of the orbital energy, $\zeta\sim\sqrt{E}$. Thus $\zeta_p/\zeta_s=\sqrt{E_p}/\sqrt{E_s}$ was used to determine the decay constants for the s orbitals (see Ref. 30 for energy level values). The values determined for the decay contants are shown in Table I. Although the calculation of the intermolecular contributions to the Hamilronian matrix elements is described in detail in Ref. 9, we now briefly sketch the main ideas in this section.

The intramolecular shift $\delta H_{11}^{\rm weak}$, and similarly $\delta H_{22}^{\rm weak}$, is determined from a sum of electrostatic, exchange, and overlap interactions,[9]

$$\delta H_{11}^{\rm weak}=(\delta H_{11})^{\rm electrostatic}+(\delta H_{11})^{\rm exchange}+(\delta H_{11})^{\rm overlap}, \tag{3}$$

between the two weakly interacting fragments. Within our approach, these terms act only on the diagonals of δH. The electrostatic and exchange interactions $(\delta H_{11})^{\rm electrostatic}$ and $(\delta H_{11})^{\rm exchange}$ are calculated based on many-body interactions which are formulated via second quantization, providing a simple picture with the correct physical insight.[9] The exchange interaction is based on the Hartree-Fock formalism to avoid the exchange-correlation problems that occur in the LDA (for example, see Refs. 31–34).

Within the intramolecular shifts, the overlap interaction term comes from a correction factor which arises from the fact that the orbitals are asumed to be orthogonal. For reasons of convenience, the transformation from the original, nonorthogonal basis to the orthogonalized set is done in two steps. First, the orbitals are made orthogonal to each other only *within* each fragment. This step is described in the Sec. II B. The intermediate basis so obtained (which is orthogonal within each fragment, and nonorthogonal between fragments) is further orthogonalized so that the orbitals located on fragment 1 are made orthogonal to those on fragment 2 by performing a Löwdin transformation[35] of the Hamiltonian matrix,

$$\mathcal{H}=\mathcal{O}_i^{-1/2}\mathcal{H}_i\mathcal{O}_i^{-1/2} \tag{4}$$

where $\mathcal{H}_i$ and $\mathcal{O}_i$ are the Hamiltonian matrix and overlap matrices in the intermediate basis, respectively. Note that $\mathcal{O}_i$ contains only nonzero elements between orbitals from different fragments, except for unity along the diagonals. These overlaps beween weakly interacting fragments are small (typically less than 0.1); therefore, the overlap matrix operator can be represented by the Taylor-series expansion

$$\mathcal{O}_i^{-1/2}=(I+S)^{-1/2}=I-\tfrac{1}{2}S+\tfrac{3}{8}S^2 \tag{5}$$

up to second order in the overlap. After multiplying out Eq. (4), diagonal corrections to $\mathcal{H}_i$ yield the overlap corrections to the intramolecular shifts, represented by the term $(\delta H_{11})^{\rm overlap}$.

Within the approximations made for the expansion of the overlap up to second order, the hopping matrix elements, $h_{12}^{\rm weak}$ and $h_{21}^{\rm weak}$ may be modeled using the Bardeen tunneling current.[36,37] Therefore, the hopping matrix element between orbitals on fragment 1 and orbitals on fragment 2, $h_{12}^{\rm weak}$ (and similarly for $h_{21}^{\rm weak}$), is written as

$$h_{12}^{\rm weak}=\gamma T_{12}^{\rm Bardeen}, \tag{6}$$

where $T_{12}^{\rm Bardeen}$ is the Bardeen tunneling current given by

$$T_{12}^{\rm Bardeen}=-\hbar^2/2m\int_{\sigma_{12}}(\psi_1\vec{\nabla}\psi_2-\psi_2\vec{\nabla}\psi_1)\cdot\vec{dS}. \tag{7}$$

The wave function ψ_1 is a localized orbital of an atom located on fragment 1, and the wave function ψ_2 is a localized orbital of an atom located on fragment 2. The wave functions ψ_1 and ψ_2 are generally nonorthogonal to each other. The correction factor γ (typically 1.4 for an atomic state bound near a Rydberg in energy) takes into account the approximations used in deriving Eq. (7) (removal of three-center interactions and any overlaping between the atomic potentials on the fragments). In additon to hydrogen-bonding interactions, the total intermolecular contribution to the total energy must include the van der Waals interactions. The energies from the r^{-6} van der Waals attractions (sometimes referred to as dispersion energies) may, in the extreme case, account for almost 50% of the total binding energy between DNA bases.[38] For larger molecules, such as the DNA bases, van der Waals interactions are necessary because of the complexity of the charge distribution and fluctuations from which such interactions are derived. Dispersion energies are due to correlation effects, and only recently have begun to be tackled within the framework of density-functional theory.[39]

A recent review article discusses empirical van der Waals energies based on the Slater-Kirkwood approximation.[40] This approximation is based on a weighted average of the dispersion coefficients C_6 due to each individual atoms polarizability, and its effective number of electrons. Using this model, the van der Waals interactions are added into our calculations through the equation

TABLE II. Atomic polarizabilities and C_6 parameters used in the Slater-Kirkwood approximation for the van der Waals interactions. [40]

Atom	α_i	C_{6ii}
H	2.60	2.8
C	6.38	19.1
N	6.90	22.8
O	5.42	16.8
P	24.32	190.8

$$\delta U_{mn}^{vdW}=\frac{C_{6mn}}{r_{mn}^{6}}, \tag{8}$$

where

$$C_{6mn}=\frac{2\alpha_m\alpha_n C_{6mm}C_{6nn}}{\alpha_m^2 C_{6nn}+\alpha_n^2 C_{6mm}}. \tag{9}$$

The subscripts m and n signify atoms on molecules i and j, respectively. The α's are the atomic polarizabilities assuming additivity. The atomic C_6 coefficients are obtained from the effective electron number and the atomic polarizability. Table II shows the values of α and C_6 that were used for H, C, N, O, and P.[40]

B. Linear scaling solution

For large systems, determining the electronic eigenvalue (band structure) energy via matrix diagonalization must be avoided due to the N^3 scaling. A linear in N scaling technique is the desirable method of choice for determining the band structure energy. In the last few years, a number of methods with linear scaling have been proposed, most of which are applicable provided that the Hamiltonian and overlap matrices of the system are sparse in a localized orbitals basis.[11–23] This is certainly the case for both $\mathcal{H}^{\rm strong}$ and $\mathcal{H}^{\rm weak}$ and for the overlap matrices in our formulation for large systems. Most of the linear scaling methods proposed so far relay on the localization properties of the magnitudes of interest like the density matrix or the electronic wave functions. For instance, for nonmetallic systems, the *occupied* electron orbitals can be constructed to be exponentially localized Wannier-like states.[41] Beyond some (small) cutoff range R_c, the overlap and Hamiltonian interactions between these Wannier-like occupied orbitals can be neglected. Since a given electron orbital overlaps significantly with only a finite number of other electron orbitals, independent of the system size, it is inherently possible to use an order-N method. In the method of Ordejón *et al.*,[15] which we use here, the following energy functional is formulated

$$\tilde{E}=2\,\mathrm{Tr}[(\mathbb{1}+(\mathbb{1}-S))H], \tag{10}$$

where H and S are the (sparse) Hamiltonian and overlap matrices for the system under consideration. The trace is taken in the occupied subspace, which has dimensions of $N_e/2$, where N_e is the number of electrons. The ground state of the system is obtained by minimizing the energy functional $\tilde{E}$ with respect to all possible *localized* states, which are expanded in terms of the atomic orbital basis (only those entering the localization sphere of radius R_c are included in this expansion). The advantage of the functional of Eq. (10) is the fact that no orthogonality constraints need to be imposed during the minimization, since the form of the functional drives the wave functions toward orthogonalization. In other words, the minimum is achieved for orthogonal functions, and for the exact ground-state band energy.[15]

The energy functional is minimized iteratively using the method of conjugate gradients.[42] In this method, a succession of line minimizations is performed, where the minimization directions are given by the ''forcelike'' gradient of the energy functional, corrected to make the successive directions orthogonal to each of the former iterations. The ''forces'' of the energy functional are just the derivatives of the energy functional with respect to the coefficients of the expansion of the occupied orbitals in terms of the basis functions. This procedure is repeated until the value of the energy functional is minimized and unchanged within some tolerance. All the computations involved (gradients and energy functional) scale linearly with the number of electrons, as long as the occupied states are localized within a radius R_c. This technique avoids the $O(N^3)$ complexity involved in the orthogonalization process present in standard iterative minimization procedures, as well as in matrix diagonalization. We refer the reader to Refs. 15 and 21 for the details of the order-N method used in this work.

In the procedure to minimize the energy functional, a judicious initial guess for the wave functions is needed. In the first molecular-dynamics time step of the simulation, we make an initial guess that takes into account the chemistry of the system. We build Wannier-like functions which are centered in bonds and lone pairs, and which are initially taken as the bonding combination of the hybrids forming the bond, or the pure lone-pair orbitals, respectively. For subsequent simulation time steps, the initial guess of wave functions are taken from the solution of preceding simulation time step. The initial wave functions in the first time step are far from orthonormality and from the Born-Oppenheimer surface, so a relatively large number of minimization iterations is needed, compared with subsequent time steps. Once the energy functional is minimized, a value for the band-structure energy is obtained along with the orthonormal set of wave functions which are then used in the calculation of the atomic forces, charge densities, etc. For the energy functional given by Eq. (10), the band-structure force can be readily evaluated, using a variation of the Hellmann-Feynman theorem.[15,21,43,44]

We recall from Sec. II A that a Löwdin orthogonalization within each fragment was to be performed. Since strong intramolecular interactions are computed in the original, nonorthogonal basis, the intramolecular Hamiltonian has to be transformed to the intermediate basis according to the Löwdin transformation:

$$\mathcal{H}_i^{\rm strong}=\mathcal{O}^{-1/2}\mathcal{H}^{\rm strong}\mathcal{O}^{-1/2}, \tag{11}$$

where $\mathcal{O}$ is the intramolacular overlap matrix,

$$\mathcal{O}=\begin{pmatrix} O_{11} & 0 \\ 0 & O_{22} \end{pmatrix}. \tag{12}$$

Similarly, in the hydrogen-bonding model, the electrostatic and exchange contributions require knowing the Löwdin charges, which are the occupation numbers of the orbitals of each atom. These charges, which are computed from the wave-function solution of the isolated noninteracting fragments $\Psi_i=\Sigma_\alpha c_{i\alpha}\psi_\alpha$, are given by

$$q_\alpha=2\sum_{i\ \mathrm{occ}}\left|\sum_{\alpha'} c_{\alpha'}\mathcal{O}^{1/2}_{\alpha'\alpha}\right|^2. \tag{13}$$

We see that both $\mathcal{O}^{-1/2}$ and $\mathcal{O}^{1/2}$ are needed to perform the Löwdin orthogonalization and to calculate the Löwdin charges, respectively. Normally, these matrices would be obtained by diagonalization of the intramolecular overlap matrix $\mathcal{O}$; however, this is an $O(N^3)$ operation, and a more efficient approach must be developed. In the hydrogen-bonding model, as discussed previously, the term $\mathcal{O}_i^{-1/2}$ was written as $(I+S)^{-1/2}$, and computing $\mathcal{O}_i^{-1/2}$ was accomplished through the use of a Taylor-series expansion up to $O(S^2)$, since terms in the intermolecular overlap matrix are small (similarly for $\mathcal{O}_i^{1/2}$). However, terms in the intramolecular overlap matrix are not small, and this procedure cannot be used in this case. Keeping within the spirit of order-*N* techniques, we found it accurate and efficient to calculate $\mathcal{O}^{1/2}$ and $\mathcal{O}^{-1/2}$ by expanding it into a Chebyshev series,

$$f(x)=\left[\sum_{k=0}^{N-1} c_k T_k(x)\right]-\tfrac{1}{2}c_0. \tag{14}$$

Chebyshev polynomials are defined over the range $[-1,1]$; therefore, the function $\mathcal{O}^{1/2}$ or $\mathcal{O}^{-1/2}$, where the eigenvalues of $\mathcal{O}$ are roughly in the range $[a,b]\epsilon[0.10,3.00]$, is scaled appropriately by

$$y=\frac{x-\frac{1}{2}(b+a)}{\frac{1}{2}(b-a)}. \tag{15}$$

The use of only approximately ten terms yields a very good convergence of the Chebyshev series to $\mathcal{O}^{1/2}$, but 20 terms are needed for a very good convergence of the Chebyshev series to $\mathcal{O}^{-1/2}$, because this function is not as stable for smaller overlap values. It must be noted that, in principle, successive multiplications of the overlap matrix in the series expansion will lead to less and less sparse matrices, which would eventually lead to a supralinear scaling of the matrix multiplications. This is avoided by using ‘‘absorbing boundary conditions’’ on each matrix multiplication after the first product. (Absorbing boundary conditions means we set to zero in the product matrix all elements which are zero in both factor matrices). The elements which are neglected decay exponentially with the distance from the nearest nonzero element of the original overlap matrix $\mathcal{O}$, and represent a reasonable approximation in the spirit of an order-*N* method.

It must be pointed out that the calculated total charge $N_{\mathrm{calc}}=\Sigma_\alpha n_\alpha$ is not precisely equal to the total number of electrons N_e due to errors. There are two error sources: (i) the energy functional Eq. (10) gives a number of electrons smaller than the exact number because of the localized functions used;[21] and (ii) the calculation of the Löwdin populations uses an approximation of $\mathcal{O}^{1/2}$. Since the n_α populations are used to determine long-range Coulomb interactions, small errors in the total charge may give rise to nonnegligable errors in the total energy. In order to correct this, the orbital populations are renormalized in such a way that the total charge is equal to the exact number of electrons, i.e.,

$$n_\alpha^{\mathrm{corrected}}=\frac{N_e}{N_{\mathrm{calc}}}n_\alpha. \tag{16}$$

This corrects for the errors in the Löwdin charges in a mean-field manner, and in fact stabilizes the solution in such a way that even fewer terms in the Chebyshev series of $\mathcal{O}^{1/2}$ and $\mathcal{O}^{-1/2}$ can be taken to yield good convergence in the total energy.

As a final technical detail, we mention that, since the orthonormal wave functions of the isolated noninteracting fragments are needed for calculating the Löwdin charges, the total band-structure energy is necessarily obtained in a two-step procedure. First, the energy functional of Eq. (10) is minimized to determine the band structure energy for only the strong intramolecular interactions (i.e., $\mathcal{H}^{\mathrm{strong}}$, which, for example could be one strand of a DNA double helix). This yields the solution for the wave functions of the isolated noninteracting fragments. The Löwdin charges are calculated from these orthonormal wave functions. Second, from these Löwdin charges we calculate the intermolecular electrostatic and exchange contributions, and the energy functional of Eq. (10) is again minimized to determine the band-structure energy for the total interaction picture (i.e., $\mathcal{H}^{\mathrm{strong}}+\mathcal{H}^{\mathrm{weak}}$).

In conclusion, we presented an electronic structure based method with the purpose of performing simulations of large biological molecules. The method combines three different techniques, providing a means to model the strong intramolecular interactions, to model the weak intermolecular interactions, and to avoid the costly $O(N^3)$ scaling as a result of matrix diagonalization. In Sec. III, the results of applying this electronic-structure-based method to the 10 basepair poly(dG)·poly(dC) DNA double helix are discussed.

III. APPLICATION OF THE METHOD TO THE 10 BASEPAIR POLY(DG)·POLY(DC) DNA DOUBLE HELIX

The combination of the above-described techniques allow us to perform a quantum-molecular dynamics calculations of a deoxyribonucleic acid (DNA) double helix. Using a 0.2-fs time step, a simulation was done to relax a DNA segment composed of ten guanine-cytosine basepairs. The relaxation was accomplished via a method known as dynamical quenching, where the velocities are set to zero as the kinetic energy reaches a maximum; thus the system's geometry seeks the nearest minimum-energy configuration. This relaxation was computed on a DEC 3000/600 Alpha workstation, requiring 691 CPU minutes for the first time step and averaging approximately 18 CPU minutes for each additional time step. As a comparison, use of a direct diagonalization method, instead of the order-*N* method used here, is estimated to take approximately 70 CPU minutes per time step.

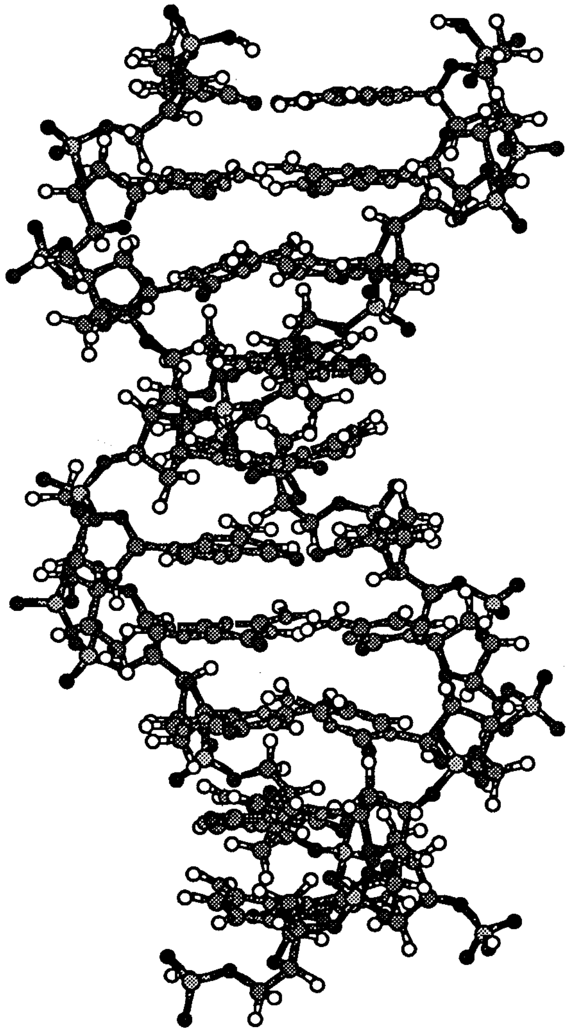

FIG. 1. This DNA segment consists of ten guanine-cytosine basepairs. Each basepair is just an $n\times 36°$ and an $n\times 3.3728$ Å translation of the first. This structure was fully relaxed to the nearest local minimum, and the electronic and vibrational DOS's were calculated.

Figure 1 shows the geometry of the relaxed structure for the nearest local-energy minimum. A comparison between this final structure and the initial structure and a quantitative analysis will be published elsewhere.[45] This structure is considered in particular because of its simplicity—each basepair is just an $n\times 36°$ and an $n\times 3.3728$-Å translation of the original basepair. The initial coordinates were obtained from a structure based on x-ray-diffraction studies of microcrystalline fibers and refined via a least-squares method using ''standard'' bond lengths.[46] Currently, water molecules surrounding the DNA segment are not included in the simulation, but will be added in future work. In natural DNA, the phosphate groups along the backbone chain are negatively charged by one e^-, and counterions exist to compensate for this change. We treat the DNA molecule taking into account this extra electron at each phosphate group, but we do not include the counterions in the simulation. Since this would give rise to long-range Coulomb repulsions between the negatively charged phosphate groups, a common approximation is to mimic the effect of the counterions by neutralizing the molecule. We accomplish this by distributing a positive electronic charge smeared evenly over the phosphorous and oxygen atoms located in the phosphate group. This smeared charge only appears in the calculation of the long-range Coulomb intermolecular interactions.

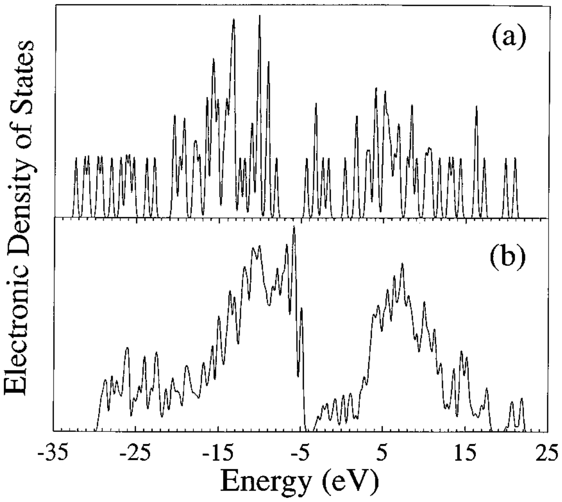

FIG. 2. The calculated electronic DOS for (a) an isolated GC basepair and (b) the poly(dG)·poly(dC) structure. The band gap ϵ_g is found to be 1.40 eV for the poly(dG)·poly(dC) structure, and 3.37 eV for the isolated basepair.

The electronic density of states (DOS) for an isolated grand canonical (GC) basepair and the poly(dG)·poly(dC) structure are shown in Figs. 2(a) and 2(b), respectively. These electronic DOS's are calculated based on the local relaxed minimum-energy configurations of the two systems. It is interesting to note that the electronic eigenvalue spectrum for the poly(dG)·poly(dC) system has qualitatively similar features as that of the isolated GC basepair. The energy-band gap ϵ_g is taken as the difference between the highest occupied molecular-orbital (HOMO) level and the lowest unoccupied molecular-orbital (LUMO) level, and is found to be 1.40 eV for the poly(dG)·poly(dC) structure, and 3.37 eV for the isolated basepair. The reduction from the band gap of the isolated basepair to the poly(dG)·poly(dC) structure is 1.97 eV, due to the addition of the backbone and phosphate groups as well as due to the broadening of the HOMO and LUMO energy levels from the coupling of the basepairs.

It is well known that LDA does not yield quantitatively accurate results for the energy values of band gaps, but rather generally underestimates these values. In addition to the LDA errors differences between our results and experimental results mainly occur due to two other important factors. First, the presence of water will structurally support the DNA double helix as well as screen charges in electrostatic interactions, resulting in an increased value for the energy gap. Second, more accurately including the effects of counterions will yield a structural difference in the DNA double helix, which will affect the HOMO-LUMO gap.

On the basis of several experiments, which measured the resistivity as a function of temperature, the energy-band gap ϵ_g of some DNA compounds was found to range from 1.8 to

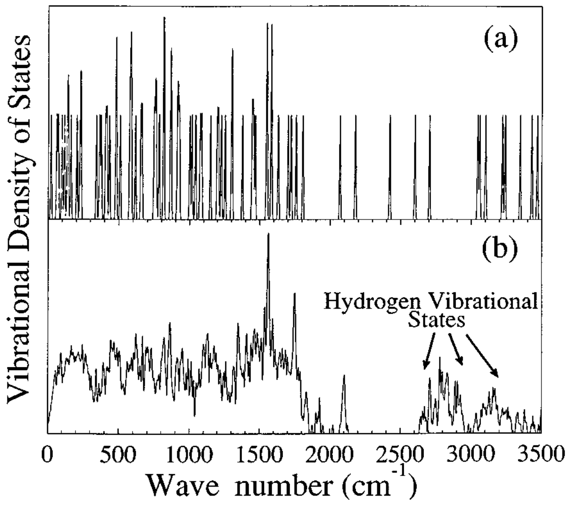

FIG. 3. The calculated vibrational DOS for (a) an isolated GC basepair and (b) the poly(dG)·poly(dC) structure.

2.4 eV.[4] In addition, a few theoretical calculations, based on semiempirical methods, were completed and determined the energy-band gap of DNA systems. Most all of these calculations also neglect the effects of water. One calculation for both the *A* and *B* forms of a DNA molecule, containing only guanine or cytosine bases, yields results for the energy band gap of approximately 2.0 eV.[47] Other semiempirical results for poly(dG)·poly(dC), find energy band gaps in the range of 6.0–6.5 eV,[48] and one result yields a value of 11.7 eV.[49] These latter results do not take into account effects due to the backbone structure or counterions, and are relatively high compared to experiment. In addition, calculations where a Mg^{2+} counterion is included find energy-band gaps of 8.7 and 2.0 eV, dependent on the location of the counterion.[50]

The calculated vibrational DOS for an isolated G·C basepair and the poly(dG)·poly(dC) structure are shown in Figs. 3(a) and 3(b), respectively. Similar features can be found within the two spectra, although there are some frequency shifts in the DNA molecule [Fig. 3(b)] due to the addition of the sugar backbone and the phosphate groups. Both spectra were obtained by a diagonalization of the dynamical matrix which was constructed by finite differences. In this method each atom is displaced, once at a time, in each direction of space by 0.0125 Å. The forces are computed on all atoms, and dividing by the displacement (assuming a harmonic approximation) gives one column of the force constant matrix. For the poly(dG)·poly(dC) structure, this involved 1932 displaced atom calculations. Cubic anharmonic terms are removed by averaging the dynamical matrix using positive and negative displacements. In the process of diagonalizing the dynamical matrix, we unfortunately obtained several negative eigenvalues. This is a common problem using finite differences, in which small errors in the force constants can break the rotational invariance of the exact dynamical matrix, and produce imaginary eigenfrequencies for the lowest-energy modes (the translational invariance is explicitly built in the calculated dynamical matrix). Also, small frequency ''floppy'' modes (which are certainly present in this large system) may not have been completely relaxed in the process of the structure optimization, which would again produce a few imaginary frequencies.

For the poly(dG)·poly(dC) structure, the local density of vibrational states (LDOS) were calculated to determine the contributions of the guanine, cytosine, sugar backbone, and phosphate components on each of the modes. Although the number of modes is quite large and most experimental work does not assign displacement patterns to the results of their spectra, this LDOS information, along with our calculated eigenvectors, is used to compare our results with those of experiment. First, from an examination of the LDOS and the eigenvectors, we find a strongly coupled sugar backbone-phosphate mode peaked at 739.1 cm^{-1}. Experimental work finds a 790-cm^{-1} band for a phosphate-sugar vibration of *B*-DNA forms.[51] Second, we located the presence of symmetric and antisymmetric stretch modes peaked at 1068.1 and 1282.4 cm^{-1}, respectively, compared to experimental values of 1094 and 1215 cm^{-1}, respectively.[52] Third, we find strong LDOS cytosine-sugar backbone modes peaked at 342.9 cm^{-1} and two other LDOS cytosine-sugar backbone modes peaked nearby at 224.9 and 235.9 cm^{-1}. Comparisons can be made qualitatively to experimental results, also showing a strong cytodine mode at 317 cm^{-1} and two nearby modes at 248 and 264 cm^{-1}, where these modes involve the ribose ring.[53] Fourth, additional comparisons of our LDOS cytosine-sugar backbone and LDOS guanine-backbone modes to that of experiment show good qualitative agreement.[54]

IV. SUMMARY

In summary, we have successfully completed an *ab initio* molecular dynamics simulation for a DNA molecule. We demonstrate that electronic structure methods have matured, and that computational resources have allowed simulations of large biological systems within a feasible amount of time. Our results for the band gap and the vibrational modes of the dehydrated poly(dG)·poly(dC) DNA structure are comparable to experimental results and the theoretical results of others.

In future work, we propose to perform simulations of the hydrated poly(dG)·poly(dC) DNA structure. In addition, more accurate modeling of the cations will be included by using a completely self-consistent version of the electronic-structure-based method presented here. Comparisons will be made to see how the effects of hydration will change the resulting electronic structure and the vibrational modes. Due to the inclusion of the water molecules, the electronic structure and vibrational properties are expected to be more accurate when compared with the experimental data.

ACKNOWLEDGMENTS

We thank the Office of Naval Research (Grant No. ONR 00014-90-J-1304) and DOE (Grant No. DEFG 02-91ER45439) for support. It is our pleasure to thank Alex Demkov, José Ortega, Wolfgang Windl, and Stuart Lindsay for useful discussions about this work. One of us (P.O.) gratefully acknowledges support and encouragement from R. M. Martin during his stay at the University of Illinois at Urbana-Champaigne, when this work was started. Some of the computations were performed on a SGI Power Challenge at NCSA.

[*] Present address: Department of Biochemistry and Biophysics, CB#7260, School of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599-7260. Electronic address: lewis@femto.med.unc.edu

[1] P. K. Weiner and P. A. Kollman, J. Comput. Chem. **2**, 287 (1981).

[2] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).

[3] J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nucleic Acids* (Cambridge University Press, New York, 1987).

[4] R. Pethig, *Dielectric and Electronic Properties of Biological Materials* (Wiley, New York, 1983).

[5] J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. Chem. **12**, 4409 (1979).

[6] J. Harris, Phys. Rev. B **31**, 1770 (1985).

[7] M. C. Payne, M. P. Teter, D. C. Allen, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).

[8] O. F. Sankey and D. J. Niklewski, Phys. Rev. B **40**, 3979 (1989).

[9] J. Ortéga, J. P. Lewis, and O. F. Sankey, Phys. Rev. B **50**, 10 516 (1994).

[10] J. P. Lewis and O. F. Sankey, Biophys. J. **69**, 1068 (1995).

[11] W. Yang, Phys. Rev. Let. **66**, 1438 (1991).

[12] D. A. Drabold and O. F. Sankey, Phys. Rev. Let. **70**, 3631 (1993).

[13] X.-P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10 891 (1993).

[14] M. S. Daw, Phys. Rev. B **47**, 10 895 (1993).

[15] P. Ordejón, D. A. Drabold, M. P. Grumbach, and R. M. Martin, Phys. Rev. B **48**, 14 646 (1993).

[16] E. B. Stechel, A. R. Williams, and P. J. Feibelman, Phys. Rev. B **49**, 10 088 (1994).

[17] L.-W. Wang, Phys. Rev. B **49**, 10 154 (1994).

[18] S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).

[19] F. Mauri and G. Galli, Phys. Rev. B **50**, 4316 (1994).

[20] W. Hierse and E. B. Stechel, Phys. Rev. B **50**, 17 811 (1994).

[21] P. Ordejón, D. A. Drabold, R. M. Martin, and M. P. Grumbach, Phys. Rev. B **51**, 1456 (1995).

[22] E. Hernandez and M. J. Gillan, Phys. Rev. B **51**, 10 157 (1995).

[23] P. Ordejón, E. Artacho, and J. M. Soler, Phys. Rev. B **53**, 10 441 (1996).

[24] R. G. Gordon and Y. S. Kim, J. Chem. Phys. **56**, 3122 (1972).

[25] D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).

[26] D. A. Drabold, P. Ordejón, J. J. Dong, and R. M. Martin, Solid State Commun. **96**, 833 (1995).

[27] A. A. Demkov, J. Ortega, O. F. Sankey, and M. P. Grumbach, Phys. Rev. B **52**, 1618 (1995).

[28] F. J. Garcia-Vidal, A. Martín-Rodero, F. Flores, J. Ort'ega, and R. Perez, Phys. Rev. B **44**, 11 412 (1991).

[29] W. J. Hehre, R. F. Stewart, and J. A. Pople, J. Chem. Phys. **51**, 2657 (1969).

[30] W. A. Harrison, *Electronic Structure and the Properties of Solids* (Dover, New York, 1989).

[31] A. D. Becke, Phys. Rev. A **38**, 3098 (1988).

[32] A. D. Becke, J. Chem. Phys. **96**, 2155 (1988).

[33] J. P. Perdew, Phys. Rev. B **33**, 8822 (1986).

[34] J. P. Perdew, Phys. Rev. B **34**, 7406 (1986).

[35] P. O. Löwdin, J. Chem. Phys. **18**, 365 (1950).

[36] F. Flores, A. Martín-Rodero, E. C. Goldberg, and J. C. Durán, Nuovo Cimento **10**, 303 (1988).

[37] J. Bardeen, Phys. Rev. Lett. **6**, 57 (1961).

[38] J. Langlet, P. Claverie, F. Caron, and J. C. Boeuve, Int. J. Quantum. Chem. **19**, 299 (1981).

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

[40] T. A. Halgren, J. Am. Chem. Soc. **114**, 7827 (1992).

[41] W. Kohn, Phys. Rev. **115**, 809 (1959).

[42] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, New York, 1986).

[43] H. Hellmann, *Einfuhrung in die Quantumchemie* (Franz Duetsche, Leipzig, 1937).

[44] R. P. Feynman, Phys. Rev. **56**, 340 (1939).

[45] J. P. Lewis, P. Ordejón, and O. F. Sankey (unpublished).

[46] R. Chandrasekaran and S. Arnott, in *Biophysics. Nucleic Acids. Crystallographic and Structural Data II,* edited by W. Saenger, Landolt-Börnstein, New Series, Group VII, Vol. I, Pt. b (Springer-Verlag, New York, 1989).

[47] T. Shinoda, N. Shima, and M. Tsukada, J. Theor. Biol. **151**, 433 (1991).

[48] B. F. Rozsnyai, F. Martino, and J. Ladik, J. Chem. Phys. **52**, 5708 (1970).

[49] P. Otto, E. Clementi, and J. Ladik, J. Chem. Phys. **78**, 4547 (1983).

[50] B. F. Rozsnyai and J. Ladik, J. Chem. Phys. **52**, 5711 (1970).

[51] S. C. Erfurth, P. J. Bond, and W. L. Peticolas, Biopolymers **14**, 1245 (1975).

[52] E. B. Brown and W. L. Peticolas, Biopolymers **14**, 1259 (1975).

[53] J. C. P. Beetz and G. Ascarelli, Spectrochim. Acta **36A**, 525 (1980).

[54] S. C. Erfurth, E. J. Kiser, and W. L. Peticolas, Proc. Natl. Acad. Sci. U.S.A. **69**, 938 (1972).