Quantum Mechanical Treatment of Biological Macromolecules in Solution Using Linear-Scaling Electronic Structure Methods

Darrin M. York*

Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France and Department of Checmistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

Tai-Sung Lee and Weitao Yang

Department of Chemistry, Duke University, Durham, North Carolina 27708-0354

(Received 24 December 1997)

A linear-scaling self-consistent field method for calculation of the electronic structure of biological macromolecules in solution is presented. The method is applied at the semiempirical Hartree-Fock level to the determination of heats of formation, solvation free energies, and density of electronic states for several protein and DNA systems. [S0031-9007(98)06157-2]

PACS numbers: 87.15.–v

The development of electronic structure methods capable of treating molecular systems up to several thousands of atoms is a first step toward the study of quantum mechanical effects of biological macromolecules in solution. The main computational bottlenecks in standard Hartree-Fock or Kohn-Sham density-functional methods derive from two sources: construction of matrix elements, and orthogonalization of molecular orbitals. The former is dominated by the classical electrostatic energy [1] that scales as $O(N^2)$, where N is the number of particles [2]. A number of methods have been developed that overcome the scaling bottleneck for systems of point charge particles [3], and recently have been extended to continuous charge distributions encountered in electronic structure calculations [4]. The orthogonalization bottleneck formally scales as $O(N^3)$, and typically manifests itself in the form of matrix diagonalization procedures. Several methods have been recently introduced that address this problem [5–7]. Here, we adopt a "divide-and-conquer" approach [6,8]; however, methods based on direct minimization of the density matrix (under the restraint it remains approximately idempotent), localized orbital approaches, and pseudodiagonalization procedures provide alternatives to the present method. Nonetheless, application of linear-scaling electronic structure methods to macromolecules in solution have only just been realized [9].

In this Letter, a method for the determination of the electronic structure of biological macromolecules (up to several thousands of atoms) in solution is developed, and results of fully self-consistent semiempirical Hartree-Fock calculations of protein and DNA systems are presented.

In Hartree-Fock molecular orbital and Kohn-Sham density-functional methods, the variational parameter is the single particle density matrix defined by

$$\rho_{ij}\langle\varphi_i|\hat{\rho}|\varphi_j\rangle\,, \tag{1}$$

where

$$\hat{\rho} = \sum_m n_m|\psi_m\rangle\langle\psi_m| = f_\beta(\hat{H} - \mu)\,, \tag{2}$$

and

$$\hat{H}|\psi_m\rangle = \varepsilon_m|\psi_m\rangle\,. \tag{3}$$

Here, $\hat{H}$ refers to either the Fock or Kohn-Sham Hamiltonian operators in the case of Hartree-Fock and Kohn-Sham methods, respectively. The second equality in Eq. (2) follows from the assumption that the occupation numbers are given by a Fermi distribution $f_\beta(\varepsilon) = (1 + e^{\beta\varepsilon})^{-1}$ with inverse temperature β, taken here to correspond to 300 K. For localized basis set methods, the density matrix can be partitioned using symmetric weight matrices $\underline{\underline{W}}^\alpha$ that are localized in real space, and normalized such that $\sum_\alpha W_{ij}^\alpha = 1 \quad \forall\ i, j$. A convenient choice is a Mulliken-type partition [10]

$$W_{ij}^\alpha = w_i^\alpha + w_j^\alpha\,, \tag{4}$$

$$w_i^\alpha \begin{cases} = \frac{1}{2} & \forall\ i \in \alpha \\ = 0, & \text{otherwise.} \end{cases} \tag{5}$$

The partitioned density matrix elements in a localized region of real space can be approximated by a local projection of the Hamiltonian in a basis set in the neighborhood of that region. The global density matrix can then be approximated by

$$\rho_{ij} = \sum_\alpha W_{ij}^\alpha \rho_{ij} \approx \sum_\alpha W_{ij}^\alpha \rho_{ij}^\alpha = \tilde{\rho}_{ij}\,, \tag{6}$$

where

$$\rho_{ij}^\alpha = \langle\varphi_i|f_\beta(\hat{H}^\alpha - \mu)|\varphi_j\rangle\,. \tag{7}$$

Here, $\hat{H}^\alpha$ is a local projection of the Hamiltonian in a basis set localized in the region of the subsystem α. In practice, this set would typically include basis functions centered on the atoms contained in the subsystem in addition to basis functions centered on nearby *buffer* atoms. The chemical potential μ in Eq. (7) is determined from the normalization requirement of the total density

TABLE I. Convergence of energetic quantities (eV) in solution with buffer and matrix element cutoffs R_b/R_m Å.[a]

	ΔH_f	ΔG_{el}	ΔE_{gap}
B-DNA			
4/7	−506.092	−444.647	5.747
6/7	−506.707	−444.259	6.541
8/9	−506.764	−444.255	7.405
10/11	−506.765	−444.255	7.405
12/13	−506.765	−444.255	7.405
Crambin			
4/7	−124.951	−11.853	1.830
6/7	−116.040	−11.536	5.682
8/9	−116.101	−11.534	6.910
10/11	−116.103	−11.534	6.923
12/13	−116.103	−11.534	6.923

[a]Quantities are the heat of formation ΔH_f, electrostatic component of the solvation free energy ΔG_{el} (see text) and energy gap ΔE_{gap} [17].

$\mathrm{Tr}(\hat{\rho}) = N$. Since the operator $\hat{H}$ itself depends on the density matrix, the procedure proceeds iteratively until self-consistency is achieved. In this formulation, there is no need for construction or diagonalization of the global Fock or Kohn-Sham Hamiltonian matrices, and, hence, the N^3 bottleneck associated with orthogonalization of the molecular orbitals is avoided. With proper choice of buffer region, the method has been demonstrated to be highly accurate and efficient [8,11].

Inclusion of solvent effects is crucial for a realistic description of the electronic structure of biological macromolecules. A particularly convenient solvation model for quantum mechanical calculations is the conductorlike screening model [12]. This model is based on a variational principle for the dielectric response of a conductor that is then scaled for finite dielectric media. This leads to an error on the order of $1/(2\varepsilon)$ [12] that is small for high dielectric media such as water ($\varepsilon \approx 80$).

For large molecules, the conventional matrix inversion solution is not feasible, since it is an $O(M^3)$ procedure, where M is the dimensionality of the surface charge vector (proportional to the molecular surface area that typically varies as $N^{2/3}$ to N for biosystems). To overcome this problem, a method has been introduced to directly minimize the electrostatic energy using a preconditioned conjugate gradient/fast multipole method [11]. The minimization procedure requires multiple evaluations of matrix-vector products that each require $O(M^2)$ effort by conventional methods. These operations, which represent the Coulombic potential of surface charge vectors, can be evaluated rapidly in $O[M \log(M)]$ effort using a recursive bisection fast multipole technique [13]. The method has been shown to be highly accurate and efficient for biological macromolecules [8,11].

Initial structures of $(CG)_8$ DNA helices in canonical *A*, *B*, and *Z* forms were generated from ideal monomer subunits obtained from fiber diffraction experiments [14], and those of proteins and complexes were obtained from nuclear magnetic resonance data in solution. Refinement of the atomic positions was accomplished in two stages: Initial geometry optimization was performed using an empirical "molecular mechanical" potential function tailored for biomolecules [15], followed by 50 steps of steepest descents minimization to relax the structures on the quantum mechanical potential energy surface.

Quantum mechanical calculations were performed using a fully self-consistent linear-scaling Hartree-Fock method [8,11] with the semiempirical AM1 Hamiltonian [16]. The amino and nucleic acid biopolymer subunits were used to define the subsystems in the divide-and-conquer procedure. Buffer atoms were determined using a distance criterion R_b, and Hamiltonian, Fock, and density matrix elements were evaluated using a cutoff R_m. Solvent effects were included self-consistently using a linear-scaling solvation method for macromolecules with atomic radii parametrized to reproduce solvation free energies of amino acid backbone and side-chain homologues and modified nucleic acid bases [11].

Table I summarizes the convergence of energetic properties with respect to the real-space cutoffs R_b and R_m. Relative energetic quantities converge rapidly with buffer size [17]. With the 8/9 Å (R_b/R_m) cutoff scheme, the heat of formation has converged to better than 10^{-5} eV/atom, representing eight significant digits in the electronic energy. Examination of the electronic density of states (DOS) in the gap region (data not shown) reveals artificial leakage of the states into the gap with small buffer size ($R_b = 4$ Å). This artifact rapidly diminishes with increasing buffer, and is essentially eliminated at $R_b = 8$ Å. Subsequently, we employ the 8/9 Å (R_b/R_m) cutoff scheme in the remainder of this Letter.

TABLE II. Energetic quantities (eV) for biological macromolecules in solution and in the gas phase.[a]

	No. atoms	ΔH_f	ΔG_{el}	$\varepsilon_{\mathrm{homo}}$	ΔE_{gap}
A-DNA	1006	−503.123 (*−55.544*)	−455.941 (*−440.160*)	−7.77 (*19.30*)	7.75 (*2.75*)
B-DNA	1006	−506.764 (*−69.489*)	−444.255 (*−430.859*)	−6.62 (*19.852*)	7.41 (*0.40*)
Z-DNA	1006	−507.308 (*−49.595*)	−471.017 (*−446.633*)	−7.79 (*19.71*)	7.78 (*1.06*)
Crambin	642	−116.101 (*−106.120*)	−11.577 (*−8.662*)	−8.69 (*−6.82*)	6.91 (*5.14*)
bpti	892	−115.017 (*−70.853*)	−47.639 (*−41.219*)	−8.86 (*−11.51*)	6.48 (*2.26*)
Lysozyme	1960	−296.839 (*−228.525*)	−75.050 (*−62.755*)	−8.10 (*−12.23*)	5.99 (*3.46*)

[a]Gas phase values are in *italics*. Protein coordinates were derived from solution NMR data obtained from the Brookhaven Protein Data Bank for crambin (1CCN), bovine pancreatic trypsin inhibitor (1PIT), and lysozyme (2LYM).

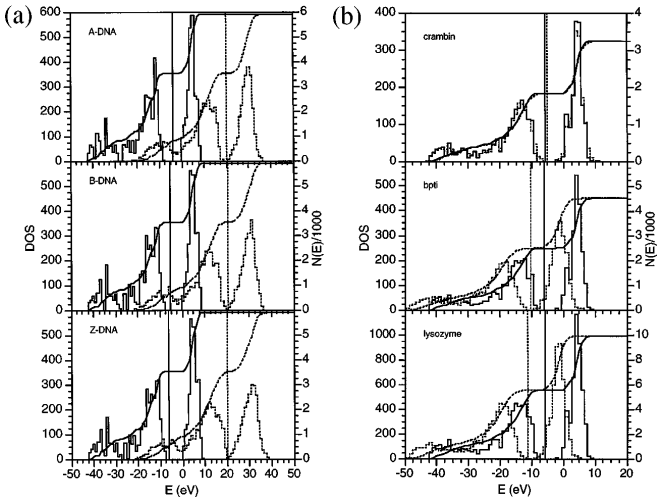

FIG. 1. Electron DOS for (a) $d(CG)_8$ in the *A*, *B*, and *Z* forms in the gas phase (dotted line) and in solution (solid line), and (b) proteins in the gas phase (dotted line) and in solution (solid line). Vertical lines indicate the Fermi levels.

Solvation effects.—Solvation has a profound effect on the electronic structure of biomolecules, and in most cases is essential for stable energies. The electrostatic contribution to the solvation free energy ΔG_{el} is defined as the electrostatic interaction between the solute charge density with its induced reaction field plus the self-energy of the reaction field. Results for DNA and proteins are summarized in Table II. For DNA, this interaction represents a tremendous stabilization energy that results from the large net negative charge. The solvation energy data indicates that *Z*-DNA is most stabilized by the dielectric medium, in qualitative agreement with the experimental observation that *Z* form DNA generally prefers high salt conditions. It should be pointed out that alternating pyrimidine/purine sequences such as $d(CG)_n$ are also characteristic of *Z*-DNA. The solvation energies of the proteins are considerably smaller than the DNA polyanions.

Density of states.—The energy of the highest occupied eigenstates give information about the rate of change of the energy with respect to the valence orbital occupations. For solvated DNA, the ΔG_{el}, $\varepsilon_{\mathrm{homo}}$, and ΔE_{gap} values follow the same relative order (Table II). Recently, there has been experimental evidence that long-range electron transfer can occur through the DNA base stack [18]; however, the mechanism of this process is not yet understood, and the subject remains controversial. The ΔE_{gap} values for the solvated DNA and proteins range from 7.4–7.8 eV and 5.99–6.91 eV, respectively. These results suggest that conduction via a free electron mechanism is unlikely for the systems considered here.

Inclusion of the solvent response greatly increases the energy gap for both DNA and proteins. For DNA, this increase is accompanied by a slight broadening and shift of the density of states toward more positive values [Fig. 1(a)]. The increased energy gap in DNA from ~0.4–2.8 eV in the gas phase to ~7.4–7.8 eV in solution reflects the preferential solvent stabilization of the occupied valence versus the unoccupied virtual orbitals. For the proteins, the increase of the gap and broadening of the DOS peaks are less pronounced than for DNA [Fig. 1(b)]. The gas phase and solution DOS of crambin, the most hydrophobic protein considered, are very similar. For bpti and lysozyme, the gas phase DOS are shifted toward more negative values as a result of the net positive charge.

Protein-protein and protein-DNA interactions.—We consider two macromolecular complexes mediated by highly ionic interactions: the protein-protein complex of myosin with calmodulin [19], and the protein-DNA complex of Myb with its DNA binding sequence [20] (Table III). It is clear that, although the change in heat of formation upon binding is highly favorable, the effect of solvation is to oppose binding of the opposite charged ionic species.

Figure 2 illustrates the effect of complex formation on the electronic DOS. The overall difference is more pronounced in the case of protein-DNA binding, indicating strong coupling of electronic states. The difference in running integration numbers $\Delta N(E)$ are almost exclusively nonpositive for both complexes. This results from a slight overall shift of the eigenstates of the separated species toward more negative values. In the case of calmodulin binding myosin, complex formation widens the energy gap, whereas the gap is narrowed in the case of Myb binding DNA.

In conclusion, we report the development and application of a linear-scaling fully self-consistent method for electronic structure of biological macromolecules in solution. The method is applied to the determination of heats of formation, solvation energies, and electronic density of states of *A*, *B*, and *Z* form DNA helices and several protein molecules. The nature of solvent effects on the binding of protein-protein and protein-DNA complexes are also examined. These results demonstrate the feasibility of applying quantum mechanical techniques toward the study of large biological systems in solution.

TABLE III. Relative binding energetics (eV) of protein-protein and protein-DNA complexes in solution.[a]

	No. atoms	ΔH_f	ΔG_{el}	$\varepsilon_{\mathrm{homo}}$	ΔE_{gap}
Calmodulin/myosin	2700	−459.309 (*−19.800*)	−113.272 (*269.538*)	−7.148 (*0.568*)	6.953 (*−0.376*)
Myb/DNA	2512	−580.081 (*−5.452*)	−225.541 (*179.944*)	−6.814 (*0.056*)	6.485 (*0.170*)

[a]Differences between values for the complexed and separated molecules are shown in *italics*. Coordinates were derived from solution NMR data obtained from the Brookhaven Protein Data Bank for calmodulin-myosin (2BBM), and Myb-DNA (1MSE).

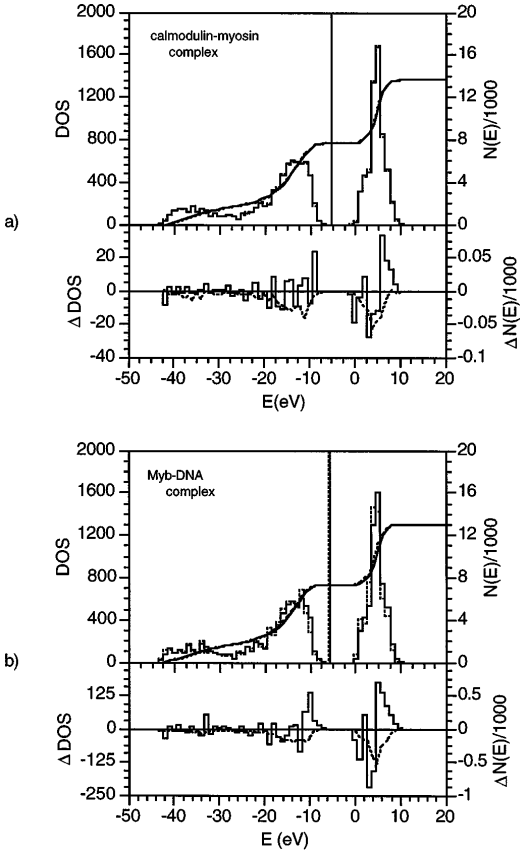

FIG. 2. Electronic DOS in solution for the complexes of (a) calmodulin with myosin, and (b) Myb with DNA. Shown are the DOS of the complex (solid line) and superposition of states of the isolated species (dotted line). The difference between the complex and superposition DOS is shown on a smaller scale immediately below.

*Corresponding author.

[1] J. Almlöf, K. J. Faegri, and K. Korsell, J. Comput. Chem. **3**, 385 (1982); M. Häser and R. Ahlrichs, J. Comput. Chem. **10**, 104 (1989); D. L. Strout and G. E. Scuseria, J. Chem. Phys. **102**, 8448 (1995).

[2] Construction of the Coulomb matrix involves a four index contraction over basis functions. For localized basis sets methods, in the limit of large numbers of particles, the non-negligible one-electron terms in the sum grow linearly with the system size. Using a supermatrix formulation and threshold criteria, the four index contraction over basis functions can be replaced by a two index contraction over non-negligible basis function products, and the formal $O(N^4)$ procedure is reduced to $O(N^2)$ in practice.

[3] L. Greengard, Science **265**, 909 (1994).

[4] M. C. Strain, G. E. Scuseria, and M. J. Frisch, Science **271**, 51 (1996); R. Kutteh, E. Aprà, and J. Nichols, Chem. Phys. Lett. **238**, 173 (1995); M. Challacombe, E. Schwegler, and J. Almöf, J. Chem. Phys. **104**, 4685 (1996); C. C. White and M. Head-Gordon, J. Chem. Phys. **104**, 2620 (1996); E. Schwegler, M. Challacombe, and M. Head-Gordon, J. Chem. Phys. **106**, 9708 (1997); J. Pérez-Jordá and W. Yang, J. Chem. Phys. **107**, 1218 (1997).

[5] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).

[6] W. Yang and T.-S. Lee, J. Chem. Phys. **103**, 5674 (1996).

[7] P. Cortona, Phys. Rev. B **44**, 8454 (1991); S. Baroni and P. Giannozzi, Europhys. Lett. **17**, 547 (1992); G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992); X.-P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10 891 (1993); F. Mauri, G. Galli, and R. Car, Phys. Rev. B **47**, 9973 (1993); P. Ordejón *et al.*, Phys. Rev. B **48**, 14 646 (1993); E. B. Stechel, A. R. Williams, and P. J. Feibelman, Phys. Rev. B **49**, 10 088 (1994); M. S. Daw, Phys. Rev. B **47**, 10 895 (1993); D. A. Drabold and O. F. Sankey, Phys. Rev. Lett. **70**, 3631 (1993); W. Kohn, Chem. Phys. Lett. **208**, 167 (1993); Phys. Rev. Lett. **76**, 3168 (1996); J. M. Millam and G. E. Scuseria, J. Chem. Phys. **106**, 5569 (1997); S. L. Dixon and K. M. Merz, Jr., J. Chem. Phys. **107**, 879 (1997); A. D. Daniels, J. M. Millam, and G. E. Scuseria, J. Chem. Phys. **107**, 425 (1997); J. J. P. Stewart, Int. J. Quantum Chem. **58**, 133 (1996).

[8] T.-S. Lee, D. M. York, and W. Yang, J. Chem. Phys. **107**, 2744 (1996).

[9] D. M. York, T.-S. Lee, and W. Yang, J. Am. Chem. Soc. **118**, 10 940 (1996).

[10] R. S. Mulliken and P. Politzer, J. Chem. Phys. **55**, 5135 (1971).

[11] D. M. York, T.-S. Lee, and W. Yang, Chem. Phys. Lett. **263**, 297 (1996).

[12] A. Klamt and G. Schüürmann, J. Chem. Soc. Perkin Trans. **2**, 799 (1993).

[13] J. Pérez-Jordá and W. Yang, Chem. Phys. Lett. **247**, 484 (1995); J. Chem. Phys. **104**, 8003 (1995).

[14] S. Arnott and D. W. L. Hukins, Biochem. Biophys. Res. Commun. **47**, 1504 (1972).

[15] S. J. Weiner *et al.*, J. Am. Chem. Soc. **106**, 765 (1984); J. Comput. Chem. **7**, 230 (1986).

[16] M. J. S. Dewar *et al.*, J. Am. Chem. Soc. **107**, 3902 (1985).

[17] For a given DOS profile, the Fermi energy is determined by the electron normalization $\varepsilon = \varepsilon(n)$. For insulating systems, the Fermi energy becomes unstable in the gap region in the low temperature limit. To avoid this instability, we define the quantities $\varepsilon_{\text{homo}} = \varepsilon(N - \delta)$ and $\Delta E_{\text{gap}} = \varepsilon(N + 1 + \delta) - \varepsilon_{\text{homo}}$, where δ is a small number taken here as 0.05.

[18] M. R. Arkin *et al.*, Science **273**, 475 (1996).

[19] M. Ikura *et al.*, Science **256**, 632 (1992).

[20] K. Ogata *et al.*, Cell **79**, 639 (1994).