Chapter 18

Application of Linear-Scaling Electronic Structure Methods to the Study of Polarization of Proteins and DNA in Solution

Darrin M. York[1]

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Semiempirical quantum calculations of biomolecular systems in solution were performed using recently developed linear-scaling methods to examine the role of solute polarization in the process of solvation. The solvation free energy of several protein and DNA molecules and complexes were computed and decomposed to asses the relative magnitude of electrostatic and polarization contributions. The effect of solvation and complex formation on the electronic density of states was also studied.

Over the past several decades, the theoretical treatment of large biomolecular systems in solution has been restricted almost exclusively to use of molecular mechanical models[1]. These models typically neglect explicit electronic polarization terms due to the need for very rapid energy and force evaluations in, for example, molecular dynamics calculations. Complementary to molecular dynamics calculations with explicit solvent have been the application of implicit solvation methods for estimation of free energies of hydration of biomolecules[2]. These methods often employ a dielectric continuum approximation and involve solving, for example, the Poisson or Poisson-Boltzmann equation. As with molecular mechanics models, these methods do not treat explicitly the electronic degrees of freedom.

In the present study, we apply recently developed linear-scaling electronic structure methods to the calculation of solvation free energies and electronic density of states (DOS) distributions of several protein and DNA systems in solution. In this way, quantum mechanical many-body polarization effects are assessed directly.

THEORY

The development of new methods for computing the electronic structure of molecular systems with computational effort that scales

[1]Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, 4, rue Blaise Pascal, Strasbourg 67000, France

Gao and Thompson; Combined Quantum Mechanical and Molecular Mechanical Methods
ACS Symposium Series; American Chemical Society: Washington, DC, 1998.

approximately linearly with system size[3-7] has recently allowed very large molecules (10^3-10^4 atoms) to be considered for the first time. Conventional single-determinant non-perturbative wave-function theories such as Hartree-Fock and density-functional methods scale as the cube of the system size due to the orthonormality constraint on the molecular orbitals that make up the wave function, or equivalently the idempotency condition for the single-particle density matrix. Here we employ a linear-scaling semiempirical approach, summarized below, to electronic structure, the details of which have been presented elsewhere[5-7].

In Hartree-Fock molecular orbital and Kohn-Sham density functional methods, the electronic energy of the system is governed by the single-particle density matrix

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

where φ_i are basis functions for the expansion of the molecular orbitals, and $\hat{\rho}$ is the density operator defined as

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

where n_m are the occupation numbers of the molecular orbitals ψ_m that are solutions of

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

In equations (2) and (3), $\hat{H}$ refers generically to either the Fock or Kohn-Sham Hamiltonian operators. The second equality in equation (2) follows from the assumption that the orbital occupation numbers are taken from a Fermi distribution $f_\beta(\varepsilon)$ with inverse temperature β (taken here to correspond to 300K), and Fermi level μ. For localized basis set methods, the density matrix can be partitioned using a set of normalized, symmetric weight matrices $\mathbf{w}^\alpha$ that are localized in real space. A convenient choice is to employ a Mulliken-type partition[8]

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

where

$$\begin{aligned} w_i^\alpha &= \tfrac{1}{2} \; \forall \, i \in \alpha \\ &= 0 \; otherwise \end{aligned} \tag{5}$$

The global density matrix can then be approximated by a superposition of partitioned elements:

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

and

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

where $\hat{H}^{\alpha}$ is a local projection of $\hat{H}$ in a set of basis functions that are localized in the region of the subsystem α. This local basis set typically is made up of basis functions centered on the atoms contained in the subsystem α in addition to basis functions centered on nearby *buffer* atoms contained in other subsystems. Normalization of the total electron density at each step of the self-consistent field procedure is enforced through adjustment of the chemical potential μ in equation (7). In the case of the Hartree-Fock methods, the total energy is given by

$$E = \tfrac{1}{2}\sum_{ij}\tilde{\rho}_{ij}\left(F_{ij}(\tilde{\rho}) + H_{ij}^{core}\right) \tag{8}$$

where $\boldsymbol{F}$ is the Fock matrix and $\boldsymbol{H}^{core}$ is the one-electron core Hamiltonian matrix. In the above formulation, there is no need for construction or diagonalization of the global Fock matrix, and hence the cubic scaling 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[5-7].

Solvent effects are critical to the behavior and stability of biomolecules in solution; hence, inclusion of solvent effects in electronic structure calculations is important. A recently developed model for high dielectric solvents in quantum mechanical calculations is the conductorlike screening model[9]. This model employs a variational principle based on a conductor, the results of which are subsequently corrected for finite dielectric media. The model has been shown to give accurate results for small polar and ionic solutes in high dielectric media such as water.

In this model, the classical electrostatic energy of a charge distribution q contained in a cavity (ε_{cav}=1) surrounded by a continuum of constant dielectric ε can be written as:

$$E_{el} = \tfrac{1}{2}\int_{\Omega}\int_{\Omega}\frac{\sigma(\mathbf{r})\sigma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\Omega d\Omega' + \int_{V}\int_{\Omega}\frac{q(\mathbf{r})\sigma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}dVd\Omega' + \tfrac{1}{2}\int_{V}\int_{V}\frac{q(\mathbf{r})q(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}dVdV' \tag{9}$$

where σ is the reaction field surface charge at the boundary Ω between different dielectric regions. For a conductor, the surface charge distribution can be determined by minimization of the total electrostatic energy with q fixed. In matrix notation, this leads to the solution

$$E_{el} = \tfrac{1}{2}\mathbf{q}^{T}\cdot(\mathbf{C}-\mathbf{B}^{\mathrm{T}}\cdot\mathbf{A}^{-1}\cdot\mathbf{B})\cdot\mathbf{q} = \tfrac{1}{2}\mathbf{q}^{T}\cdot\mathbf{G}\cdot\mathbf{q} \tag{10}$$

where σ and $\mathbf{q}$ are column vectors of the reaction field and solute charge distributions, respectively, the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ represent Coulombic interactions between σ:σ, σ:$\mathbf{q}$, and $\mathbf{q}$:$\mathbf{q}$, respectively, and $\mathbf{G}$ is the Green's function matrix for the problem. For a finite dielectric, the surface charge distribution is scaled by a factor of (ε-1)/ε in accord with the Gauss theorem,

and leads to an error on the order of $1/(2\varepsilon)$[9], which is small for a high dielectric medium such as water (ε=80). The main advantages of the method is that direct computation of the Green's function matrix makes analytic derivatives facile, and affords high efficiency with conventional methods that incorporate **G** directly into the Hamiltonian at the beginning the self-consistent field procedure. A disadvantage of the method is the reduced reliability of results for low dielectric media compared to exact solution of the Poisson equation.

The solution equation (10) cannot be applied as written for large molecules since the process of matrix inversion scales as M^3, where M is the dimension of the surface charge vector. However, this problem can be overcome by use of a preconditioned conjugate gradient minimization technique and fast multipole method[10] for linear-scaling evaluation of electrostatic interactions[7]. The method is an iterative minimization technique that requires evaluation of matrix-vector products of the form $\mathbf{A}\cdot\mathbf{x}=\mathbf{b}$ (an M^2 procedure) at each iteration. This operation corresponds to evaluation of the electrostatic potential of the surface charge vector **x**, and can be realized with order $M\cdot log(M)$ effort using fast multipole methods[10]. The number of iterations can be decreased using a preconditioner matrix $\mathbf{A}_{pc}$ such that $\mathbf{A}\cdot\mathbf{A}_{pc}\approx 1$. We have chosen $\mathbf{A}_{pc}$ as the inverse of the block diagonal matrix constructed by the elements of **A** corresponding to common atomic surface area patches. With this preconditioner, the number of iterations to reach a fixed convergence level does not appear to grow with system size for the molecules considered here. The method has been shown to be accurate and efficient for biological macromolecules[5-7].

METHODS

Structures for DNA $(CG)_8$ helices in canonical A, B, and Z-forms were generated from ideal monomer subunits obtained from fiber diffraction experiments[11]. Initial structures for proteins, and protein-protein and protein-DNA complexes were obtained from nuclear magnetic resonance data in solution, and refined with 50 steps of steepest descents energy minimization to relax the structures on the quantum mechanical energy surface. Quantum mechanical calculations were performed using the self-consistent linear-scaling electronic structure methods described previously[5-7,12] with the semiempirical AM1 Hamiltonian[13]. Subsystem partitions were chosen to be the amino- and nucleic acid biopolymer subunits. Buffer regions were determined using a 8Å distance criterion R_b, and core Hamiltonian, Fock, and density matrix elements were evaluated using a 9Å cutoff R_m (see below). Solvent effects were included self-consistently using a linear-scaling solvation method for macromolecules with atomic radii parameterized to reproduce solvation free energies of amino acid backbone and side-chain homologues and modified nucleic acid bases[7].

Table I. Convergence of energetic quantities (eV) in solution with buffer and matrix element cutoffs R_b/R_m (Å)*.

	ΔH_f	$\Delta G_{el,sol}$	ε_{homo}	ΔE_{gap}
B-DNA (1006 atoms, total charge -30)				
4/7	-506.092	-444.647	-5.768	5.747
6/7	-506.707	-444.259	-6.755	6.541
8/9	-506.764	-444.255	-7.619	7.405
10/11	-506.765	-444.255	-7.619	7.405
12/13	-506.765	-444.255	-7.619	7.405
crambin (642 atoms, total charge 0)				
4/7	-124.951	-11.853	-5.964	1.830
6/7	-116.040	-11.536	-7.470	5.682
8/9	-116.101	-11.534	-8.692	6.910
10/11	-116.103	-11.534	-8.704	6.923
12/13	-116.103	-11.534	-8.705	6.923

*Quantities are the heat of formation ΔH_f, electrostatic component of the solvation free energy $\Delta G_{el,sol}$, highest occupied molecular orbital eigenvalue ε_{homo} and energy gap ΔE_{gap} (see text).

RESULTS

Convergence

Table I summarizes the convergence of energetic quantities with buffer size and matrix cutoff for a canonical $(CG)_8$ B-DNA helix, and crambin in solution. The matrix element cutoff (R_m) was chosen to be slightly larger than the buffer cutoff (R_b) to insure inclusion of all off-diagonal subsystem-buffer matrix elements. A matrix element cutoff below 7Å was observed to lead to larger errors (data not shown). All energetic quantities are well converged using the 8/9Å (R_b/R_m) scheme. Energetic quantities converge slightly faster in the case of B-DNA (~10^{-3} eV with the 8/9Å scheme) than for crambin, particularly for the energy gap (~10^{-2} eV with the 8/9Å scheme),

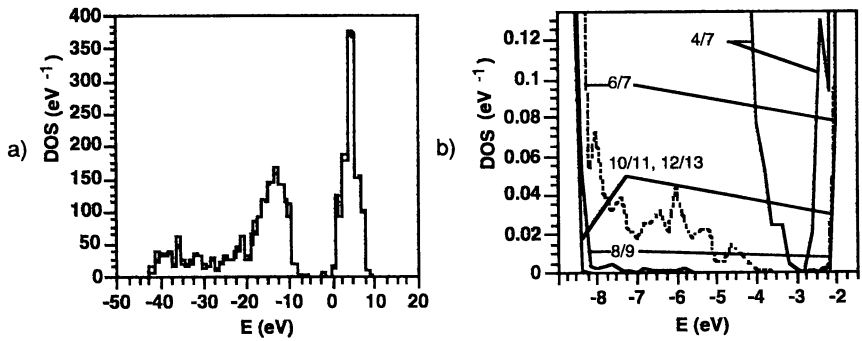

Fig. 1. Convergence of the electronic DOS of crambin in solution with different R_b/R_m schemes: a) overall DOS, b) DOS near the Fermi level (Note: the area of the graph corresponds to 1 electron).

although the difference might not be significant. Figures 1(a,b) illustrate the variation of the DOS for crambin as a function of (R_b/R_m). The overall DOS profile is similar in all cases. A detailed examination of the DOS in the region of the energy gap reveals the low order 4/7Å scheme results in "leakage" of electron density into the region of the gap (Figure 1b). This effect is significantly reduced using the 6/7Å scheme, and becomes almost negligible with larger cutoffs. In what follows, we employ the 8/9Å scheme for determination of electronic properties of protein and DNA systems in solution.

Solvation energies

The process of solvation can be decomposed into different paths, two of which are considered here (Figure 2). The first path involves 1) freezing the electronic degrees of freedom of the solute in the gas phase and allowing this charge distribution to induce an electrostatic solvent reaction field ($\Delta G_{el,gas}$), followed by 2) electronic relaxation (polarization) of the gas phase charge distribution to the final solution phase charge distribution, at the same time allowing the induced reaction field to adjust accordingly (ΔG_{pol}). Both $\Delta G_{el,gas}$ and ΔG_{pol} involve relaxation processes and are stabilizing (lower the total energy). The second path in Figure 2 involves 1) an internal perturbation of the gas phase electronic charge distribution to the solution phase distribution in the absence of a solvation effect (ΔG_{int}), followed by 2) adding the reaction field ($\Delta G_{el,sol}$) that is induced by the solution phase charge distribution. The former is a destabilizing internal energy reorganization contribution, whereas the latter consists of a stabilizing solvent reaction field response to the solution phase charge distribution (analogous to $\Delta G_{el,gas}$ for the gas phase charge distribution). In this study, only electronic energy contributions are evaluated (no cavitation terms) with fixed geometries, and the effect of solvation is approximated by a

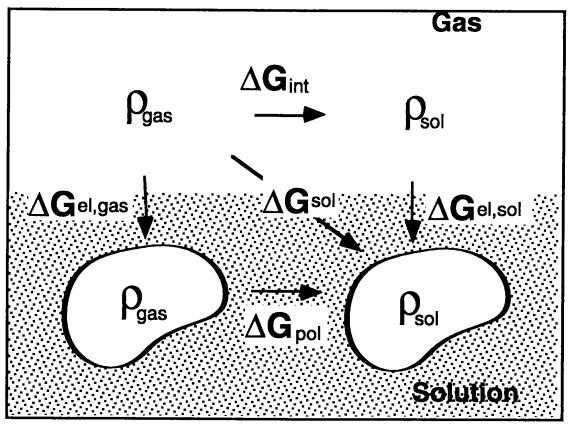

Fig. 2. The solvation process decomposed into different paths.

classical electrostatic dielectric continuum model as described above. The main advance in this study consists of the explicit consideration of the contribution of quantum mechanical many-body effects of the macromolecular solute in the process of solvation.

Table II shows the solvation free energy components for different canonical forms of DNA and several proteins derived from NMR. By their definition, the electrostatic components of the solvation free energy ($\Delta G_{el,gas}$ and $\Delta G_{el,sol}$) bracket the total free energy of solvation ΔG_{sol}; i.e., $-\Delta G_{el,gas} \leq -\Delta G_{sol} \leq -\Delta G_{el,sol}$, or equivalently $\Delta G_{pol} \leq 0 \leq \Delta G_{int}$. The magnitude of the values of $\Delta G_{el,gas}$ and $\Delta G_{el,sol}$ differ from the values for ΔG_{pol} and ΔG_{int} by an order of magnitude or more for the molecules studied here. The ΔG_{pol} contributes approximately 10% for the proteins and 2% for the DNA. This is consistent with results of hybrid quantum mechanical/molecular mechanical simulations of peptides in solution[14]. Although the magnitude of the polarization term is greater (per atom) for DNA than for the proteins, the relative percentage is significantly less due to the dominant $\Delta G_{el,gas}$ term. The ΔG_{int} term reflects the internal energy penalty the solute pays in order to adopt the ideal solution phase charge distribution. Within the linear response regime, this term is equal to minus the *solute polarization* energy (see below).

The contributions due to solvation and solute polarization to the free energy of binding of myosin to calmodulin[15] and the DNA binding domain of Myb with DNA[16] are summarized in Table III. In both cases, the process of binding has a neutralizing effect as oppositely charged species come together. Consequently, the change in the solvation free energy strongly disfavors complexation, although the overall enthalpy of formation is predicted to be favorable. In these examples $\Delta\Delta G_{pol}$ is positive. This arises from the separated species being more polarized due to both greater exposed surface area and a larger charge induced reaction field than in the

TABLE II: Solvation free energies, enthalpies of formation, and Fermi gap energies (eV) of proteins and DNA*.

Molecule	atoms	ΔG_{sol}	$\Delta G_{el,gas}$	ΔG_{pol}	ΔG_{int}	$\Delta G_{el,sol}$	ΔH_f		ΔE_{gap}	
Proteins										
crambin	642	-10.0	-8.7	-1.3	1.6	-11.6	-116.1	*(-106.1)*	6.91	*(5.14)*
bpti	892	-44.2	-41.2	-2.9	3.5	-47.6	-115.0	*(-70.9)*	6.48	*(2.26)*
lysozyme	1960	-68.3	-62.8	-5.6	6.7	-75.0	-296.8	*(-228.5)*	5.99	*(3.46)*
DNA										
A-DNA	1006	-447.6	-440.2	-7.4	8.4	-455.9	-503.1	*(-55.5)*	7.75	*(2.75)*
B-DNA	1006	-437.3	-430.9	-6.4	7.0	-444.3	-506.8	*(-69.5)*	7.41	*(0.40)*
Z-DNA	1006	-457.7	-446.6	-11.1	13.3	-471.0	-507.3	*(-49.6)*	7.78	*(1.06)*

*Unrefined protein coordinates (see text) were obtained from solution NMR data taken from the Brookhaven Protein Data Bank for crambin (1CCN), bovine pancreatic trypsin inhibitor (1PIT), and lysozyme (2LYM). $(CG)_8$ sequences of duplex A, B, and Z-form DNA were constructed from idealized subunits derived from fiber diffraction data (see text). Values for ΔH_f and ΔE_{gap} are given for calculations in solution (plain text) and in the gas phase *(italics)*.

TABLE III: Solvation free energies, enthalpies of formation, and Fermi gap energies (eV) of biomolecular complexes in solution*.

Molecule	atoms	ΔG_{sol}	$\Delta G_{el,gas}$	ΔG_{pol}	ΔG_{int}	$\Delta G_{el,sol}$	ΔH_f	ΔE_{gap}
Myb-DNA								
complex	2512	-101.2	-86.8	-14.4	12.1	-113.3	-459.3	6.95
Myb	1815	-125.3	-115.1	-10.2	9.6	-134.8	-159.4	7.77
DNA	697	-248.2	-242.7	-5.5	6.1	-254.3	-319.7	7.79
ΔΔ		*272.3*	*271.0*	*1.3*	*-3.6*	*275.8*	*-19.8*	*-0.38*
myosin-calmodulin								
complex	2700	-101.2	-86.8	-14.4	12.1	-113.3	-580.1	6.49
myosin	2259	-65.6	-52.1	-13.6	8.5	-74.1	-546.0	7.33
calmodulin	441	-310.8	-285.8	-24.9	19.1	-329.8	-28.5	8.21
ΔΔ		*168.8*	*153.7*	*15.1*	*-9.7*	*178.4*	*-5.5*	*0.17*

*Differences between values for the complexed and separated molecules (ΔΔ) are shown in *italics*. The ΔE_{gap} values for ΔΔ were obtained from a DOS distribution that was a superposition of DOS distributions of the separated molecules. Unrefined coordinates were obtained from solution NMR data taken from the Brookhaven Protein Data Bank for calmodulin-myosin (2BBM), and Myb-DNA (1MSE).

complexed forms. Conversely, $\Delta\Delta G_{int}$ is negative, indicating there is less of an internal energy penalty for adopting the solution phase charge distribution in the absence of a reaction field for the complexes than in the separated species.

It is of interest to determine whether the electronic response of the solute in the process of solvation is a linear response[17]. This can be addressed by considering the process of perturbing the gas phase system by an applied field, and calculating the energetic stabilization that results from the electronic relaxation of the perturbed system. In this case, the external field is taken as the solvent reaction field v_{RF} for the solution phase charge distribution. The process we are interested in is thus:

$$\left[\rho_{gas}\right]_o \xrightarrow{v_{RF}(\mathbf{r})} \left[\rho_{gas}\right]_{v_{RF}(\mathbf{r})} \xrightarrow{\Delta E_{pol,solute}} \left[\rho_{sol}\right]_{v_{RF}(\mathbf{r})} \tag{11}$$

If the electronic response of the solute is a linear response, then the solute polarization energy is given by

$$\Delta E_{LR,solute} = \tfrac{1}{2}\int \delta\rho(\mathbf{r}) v_{RF}(\mathbf{r}) d^3r \tag{12}$$

where $\delta\rho=\rho_{sol}-\rho_{gas}$ is the *solute polarization density*. The solute polarization energy $\Delta E_{pol,solute}$ of equation (11) is calculated as

$$\Delta E_{pol,solute} = \Delta G_{sol} - \Delta G_{el,sol} + 2\Delta E_{LR,solute} = \Delta E_{LR,solute} + (\Delta G_{int} + \Delta E_{LR,solute}) \tag{13}$$

From equation (13) the quantity $(\Delta G_{int}+\Delta E_{LR,solute})$ is the difference between the solute polarization energy and its ideal linear response value. Figure 3

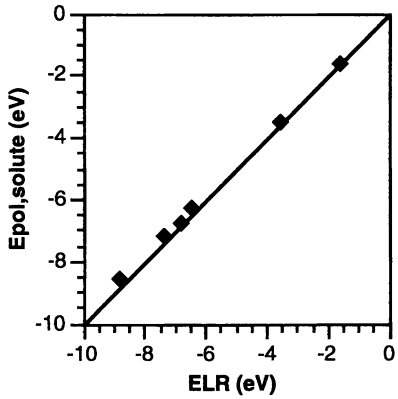

Fig. 3. Regression of the solute polarization energy and ideal linear-response energy (see text).

shows a linear regression of the calculated values of $\Delta E_{pol,solute}$ and $\Delta E_{LR,solute}$. For the systems considered here, the macromolecular response of the solutes is very nearly a linear response. This lends further credence to the use of linear response models for modeling polarization in biomolecular simulations in solution[17].

Energy gaps and electronic DOS

In the present method, the number of electrons for a given set of subsystem molecular orbitals is determined from the Fermi level ε, and given by

$$N(\varepsilon) = \sum_{ij} \sum_{\alpha}^{sub-systems} W_{ij}^{\alpha} S_{ij}^{\alpha} \sum_{m} f_{\beta}(\varepsilon_m^{\alpha} - \varepsilon) C_{im}^{*} C_{jm} \qquad (14)$$

where $\mathbf{S}^{\alpha}$ is the subsystem overlap matrix, and $\mathbf{C}^{\alpha}$ is the matrix of molecular orbital expansion coefficients, and $f_{\beta}(\varepsilon)$ and $\mathbf{W}^{\alpha}$ are the Fermi function and partition weight matrix, respectively, as defined earlier. The electronic density of states is defined as

$$g(\varepsilon) = \left(\frac{\partial N}{\partial \varepsilon}\right)\Bigg|_{\varepsilon} = -\sum_{ij} \sum_{\alpha}^{sub-systems} W_{ij}^{\alpha} S_{ij}^{\alpha} \sum_{m} f'_{\beta}(\varepsilon_m^{\alpha} - \varepsilon) C_{im}^{*} C_{jm} \qquad (15)$$

where the derivative of the Fermi function is $f'_{\beta}(\varepsilon) = -\beta exp(\beta\varepsilon)/(1+exp(\beta\varepsilon))^2$. It is often convenient, and more numerically stable, to calculate the DOS by finite differences, especially with a large value of the inverse temperature β. Use of a finite β establishes a unique mapping between the number of electrons and corresponding Fermi level, i.e. $\varepsilon \leftrightarrow N$ and therefore $\varepsilon = \varepsilon(N)$.

The function $\varepsilon(N)$ can be easily solved numerically. This allows a definition of electronic structural quantities analogous to conventional molecular orbital methods, in particular we define:

$$\varepsilon_{homo} = \varepsilon(N-\delta),\ \varepsilon_{lumo} = \varepsilon(N+\delta),\ \textit{and}\ \ \Delta E_{gap} = \varepsilon_{homo} - \varepsilon_{lumo} \qquad (16)$$

where δ is a small number, taken here to be 0.05. Here δ is used to avoid numerical instabilities associated with the Fermi energy in the region of energy gaps, especially in the low temperature (large β) limit. To illustrate this, we note that if one uses the standard orbital population conventions (occupation numbers 1 for the N lowest lying orbitals and 0 otherwise, corresponding to the $\beta\rightarrow\infty$ limit), the "Fermi level" for an insulator, as defined by the normalization condition, is not unique; i.e. the Fermi level can take on any value between the highest occupied and lowest unoccupied molecular orbital eigenvalues. As values of β become large, determination of the Fermi level becomes numerically unstable. The definitions (16) avoid this instability, and result in rapidly convergent quantities as shown earlier in Table I.

The electronic density of states and energy gap at the Fermi level are useful quantities for describing molecular electronic structure. Here we consider the electronic DOS of the biomolecules and complexes discussed previously in the gas phase and in solution (Tables II & III). In the case of $(CG)_8$ DNA in canonical A, B, and Z-forms, the energy gap ranges from 7.41-7.78 eV. 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 energy gap results here suggest that a free electron conduction mechanism is unlikely. The gas phase energy gap values are listed for comparison to illustrate the dramatic effect of solvent stabilization on the electronic structure of these molecules. In the case of the protein systems examined here (all of which are neutral or cationic at neutral pH), the energy gaps in solution are slightly smaller than for the DNA (5.99-6.91 eV), and the electronic DOS are shifted toward more negative values in solution[12]. We note that the energy gaps for both DNA and proteins increase in solution resulting from preferential solvent stabilization of the occupied valence states relative to the virtual states; the smallest change occurs in the case of the hydrophobic protein crambin (5.14 eV to 6.91 eV).

Figure 4 compares the electronic DOS for the myosin-calmodulin and Myb-DNA complexes relative to the uncomplexed species in solution. The value of $\Delta N(E)$ is seen to be non-positive in these calculations. This results from a slight shift in the electronic levels toward more negative values in the uncomplexed molecules, and is more pronounced in the case of Myb-DNA complex. The energy gaps at the Fermi level are similar to those of the proteins, and do not change significantly upon complex formation (Table III).

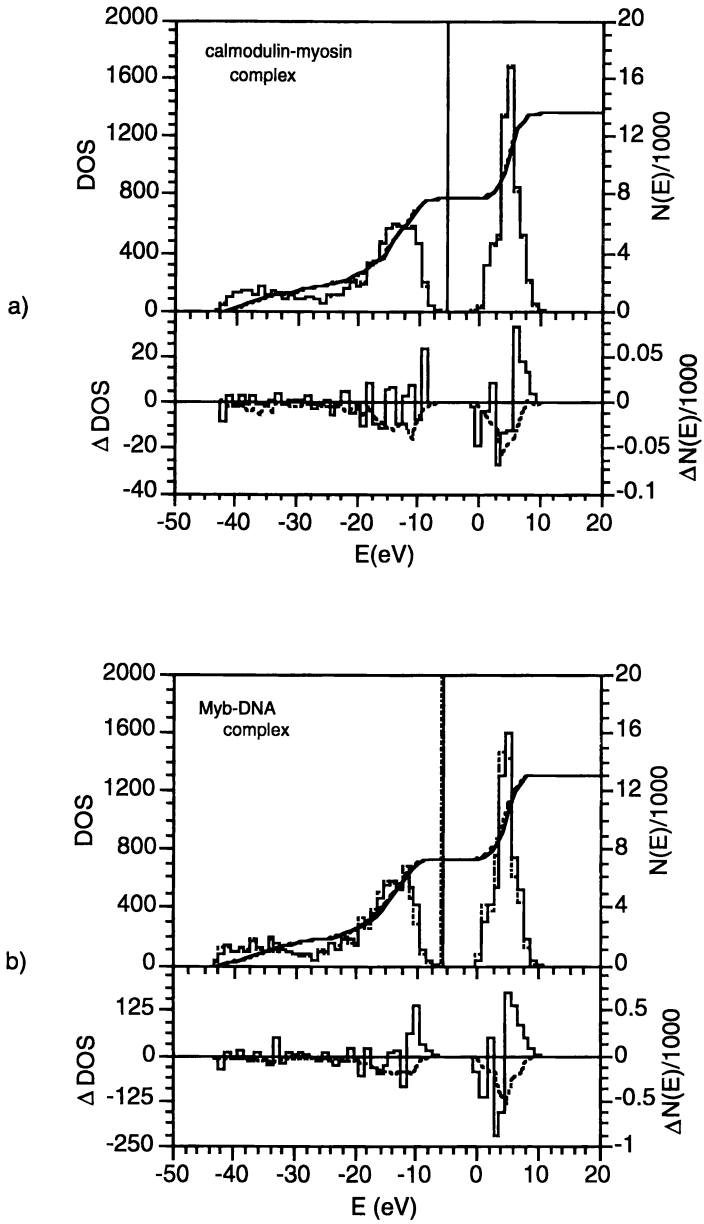

Fig. 4. Electronic DOS (eV^{-1}) and $N(\varepsilon)$ (see text) in solution for a) calmodulin-myosin, and b) Myb-DNA. Shown are values for the complexes (solid line), superposition of states of the isolated species (dotted line), and the difference (shown immediately below). Vertical lines indicate the Fermi levels.

CONCLUSION

Linear-scaling electronic structure calculations have been performed for several biomolecules in solution at the semiempirical level to investigate the effects of solute polarization on solvation free energies and electronic density of state distributions. Results in the gas phase and in solution are compared. It is demonstrated that polarization contributes on the order of 10% for proteins and 2% for DNA of the total solvation free energy. The electronic response of the solute in the process of solvation is well approximated by a linear response model. In the case of binding between highly charged protein and DNA molecules, the overall $\Delta\Delta G_{sol}$ strongly disfavors binding, as expected, with $\Delta\Delta G_{pol}$ and $\Delta\Delta G_{int}$ making positive and negative contributions, respectively. Solvation has a pronounced effect on the electronic DOS, especially of highly charged biomolecules, causing a shift and broadening of the spectrum. The energy gaps at the Fermi level are observed to significantly increase upon solvation. These results are a first step toward the study of biological macromolecules in solution using self-consistent field methods to treat explicitly quantum mechanical many-body effects of the solute.

ACKNOWLEDGMENTS

The author acknowledges financial support through and NIH postdoctoral fellowship at Harvard University, Cambridge, MA, USA and an EMBO postdoctoral fellowship at Université Louis Pasteur, Strasbourg, France. The author thanks Prof. Martin Karplus for helpful discussions and a critical reading of the manuscript.

REFERENCES AND NOTES

1. C. L. Brooks III, M. Karplus and B. M. Pettitt *Proteins: A theoretical perspective of dynamics, structure, and thermodynamics* (John Wiley & Sons, New York, 1988). A. Warshel, *Computer Simulation of Chemical Reactions in Enzymes and Solutions* (John Wiley & Sons, New York, 1991).
2. Gilson, M. K. *Curr. Opinion in Struct. Biol.* **1995**, *5*, 216. Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144.
3. Yang, W. *Phys. Rev. Lett.* **1991**, *66*, 438. Yang, W.; Lee, T.-S. *J. Chem. Phys.* **1995**, *103*, 5674.
4. Cortona, P. *Phys. Rev. B* **1991**, *44*, 8454. Baroni, S.; Giannozzi, P. *Europhys. Lett.* 1992, *17*, 547. Galli, G.; Parinello, M. *Phys. Rev. Lett.* **1992**, *69*, 3547. Li, X.-P.; Nunes, W.; Vanderbilt, D. *Phys. Rev. B.* **1993**, *47*, 10891. Mauri, F.; Galli, G.; Car, R. *Phys. Rev. B* **1993**, *47*, 9973. Ordejón, P.; Drabold, D. A.; Grumback, M. P.; Martin, R. M. *Phys. Rev. B* **1993**, *48*, 14646. Stechel, E. B.; Williams, A. P.; Feibelman, P. J. *Phys. Rev. B* **1993**, *49*, 3898. Daw, M. S.*Phys. Rev. B* **1993**, *47*, 10899. Drabold, D. A.; Sankey, O. F. *Phys. Rev. Lett.* **1993,** *70*, 3631. Kohn, W. *Chem. Phys. Lett.* **1993**, *208*, 167. Kohn, W. *Phys. Rev. Lett.* **1996**, *76*, 3168. Stewart, J. J. P. *Int. J. Quant. Chem.* **1996**, *58*, 133. S. L. Dixon, S. L.; Merz, K. M.; *J. Chem. Phys.* **1996**, *107*, 6643. Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1997**, *107*, 879. Millam, J. M.; Scuseria, G. E. *J. Chem. Phys.* **1997**, *106*,

5569. Daniels, A. D.; Millam, J. M.; Scuseria, G. E. *J. Chem. Phys.* **1997**, *107*, 425.
5. York, D. M.; Lee, T.-S.; Yang, W. *J. Am. Chem. Soc.* **1996**, *118*, 10940.
6. Lee, T.-S.; York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *105*, 2744.
7. York, D. M.; Lee, T.-S.; Yang, W. *Chem. Phys. Lett.* **1996**, *263*, 297.
8. Mulliken, R. S.; Politzer, P. *J. Chem. Phys.* **1971**, *55*, 5135.
9. Klamt, A.; Schüürmann, G. *J. Chem. Soc. Perkin Trans.* **1993**, *2*, 799.
10. Pérez-Jordá, J.; Yang, W. *Chem. Phys. Lett.* **1995**, *247*, 484. Pérez-Jordá, J.; Yang, W. *J. Chem. Phys.* **1995**, *104*, 8003.
11. Arnott, S.; Hukins, D. W. L. *Biochem. Biophys. Res. Commun.* **1972**, *47*, 1504.
12. York, D. M.; Lee, T.-S.; Yang, W. *Phys. Rev. Lett.*, in press.
13. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902. Dewar, M. J. S.; Healy, E. F.; Holder, A. J.; Yuan, Y. C.; *J. Comput. Chem.* **1990**, *11*, 541.
14. Gao, J.; Xia, X.*Science* **1992**, *258,* 631. Gao, J. *Biophys. Chem.* **1994**, *51*, 253. Orozco, M.; Luque, F. J.; Habibollahzadeh, D.; Gao, J. *J. Chem. Phys.* **1995**, *102*, 6145. Thompson, M. A.; Schenter, G. K. *J. Phys. Chem.* **1995**, *99*, 6374.
15. Ikura, M.; Clore, G. M.; Gronenborn, A. M.; Zhu, G.; Klee, C. B.; Bax, A. *Science* 1992, *256*, 632.
16. Ogata, K.; Morikawa, S.; Nakamura, H.; Sekikawa, A.; Inoue, T.; Kanai, H.; Sarai, A.; Ishit S.; Nishimura, Y. *Cell* **1994**, *79*, 639.
17. Rappé, A. K.; Goddard, W. A. *J. Phys. Chem.* **1990**, *94*, 4732. Rick, S. E.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141. Bernardo, D. N.; Ding, Y.; Krogh-Jespersen, K.; Levy, R. M. *J. Comput. Chem.* **1995**, *16*, 1141. Gao, J *J. Phys. Chem. B* **1997**, 101, 657. York, D. M.; Yang, W. *J. Chem. Phys.* **1995**, *104*, 159.
18. Murphy, C. J.; Arkin, M. R.; Jenkins, Y.; Ghatlia, N. D.; Bossmann, S. H.; Turro, N. J.; Barton, J. K. *Science* **1993**, *262*, 1023. Arkin, M. R.; Stemp, E. D. A.; Holmlin, R. E.; Barton, J. K.; Hörmann,m A.; Olson, E. J. C.; Barbara, P. F. *Science* **1996**, *273*, 475.