6 December 1996

ELSEVIER

Chemical Physics Letters 263 (1996) 297–304

CHEMICAL PHYSICS LETTERS

Parameterization and efficient implementation of a solvent model for linear-scaling semiempirical quantum mechanical calculations of biological macromolecules

Darrin M. York [a,b,*], Tai-Sung Lee [b], Weitao Yang [b]

[a] *Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA*
[b] *Department of Chemistry, Duke University, Durham, NC 27708-0354, USA*

Received 10 June 1996; in final form 26 September 1996

Abstract

A method is developed to include solvation effects in linear-scaling semiempirical quantum calculations. Favorable scaling of computational effort for large molecules is achieved using a preconditioned conjugate gradient technique in conjunction with a linear-scaling recursive bisection method for evaluation of electrostatic interactions. The method requires approximately 30% computational overhead relative to gas-phase calculations. Effective atomic radii for biological macromolecules are derived from fitting to experimental and theoretical solvation energies for small molecules homologous to amino- and nucleic acid residues.

1. Introduction

Recent advances in computational resources and theoretical developments have greatly extended the realm of molecular problems accessible to quantum mechanical investigation. For the first time, macromolecular systems can begin to be studied with electronic structure methods. The field, however, is in its infancy, and faced with a host of new challenges. Many of the well-established methods for small molecules cannot be utilized in the limit of very large number of particles because of computational scaling barriers. In this letter we describe the parameterization and implementation of a solvation model [1] within a linear-scaling electronic structure theory framework [2] that allows biological macromolecules to be studied with semiempirical quantum mechanical methods [3].

2. Methodology

In conventional electronic structure methods, including semiempirical molecular orbital methods, the variational parameter is the single-particle density matrix defined by

$$\rho_{ij} = \sum_m n_m c_{im}^* C_{jm} , \tag{1}$$

where n_m and $\{C_m\}$ are the molecular orbital occupation numbers and basis set expansion coefficients.

* Corresponding author.

PII S0009-2614(96)01198-0

The expansion coefficients are obtained as eigenvectors of the Hartree–Fock equation

$$\mathbf{F}\cdot\mathbf{C}=\mathbf{S}\cdot\mathbf{C}\cdot\epsilon\,, \tag{2}$$

where $\mathbf{F}$ and $\mathbf{S}$ are the usual Fock and overlap matrices, respectively. Direct solution of Eq. (2) via matrix diagonalization or constrained minimization procedures lead to algorithms that require computational effort proportional to (at least) the cubic of the number of particles. This $\mathcal{O}(N^3)$ scaling behavior ultimately limits the size of systems that can be handled by conventional quantum mechanical methods. Several 'linear-scaling' electronic structure methods have been recently proposed that overcome the $\mathcal{O}(N^3)$ bottleneck (for example, see [2] and and references therein). Here we employ a 'divide-and-conquer' approach [4,5] toward the quantum mechanical treatment of biological macromolecules in solution [3].

For linear combination of atomic orbitals (LCAO) or other localized basis-set methods, the Fock, overlap, and density matrices are inherently sparse. It is in the exploitation of the nature of this sparsity pattern in conjunction with partitioning methods that allow the divide-and-conquer method to overcome the $O(N^3)$ diagonalization procedure while achieving very high accuracy [2,5]. The first step is to divide the molecule into localized (nonoverlapping) groups of atoms termed *subsystems*. A localized set of (overlapping) *weight functions* for each subsystem are then introduced to partition the electron density [4] or density matrix [2,5]. For Hartree–Fock methods, it is convenient to use a Mulliken-type partition for the density matrix. We define the symmetric weight matrix P_{ij}^{α} for each subsystem α by

$$P_{ij}^{\alpha}=1 \quad \forall\, i,j\in\alpha\,,$$

$$P_{ij}^{\alpha}=\tfrac{1}{2} \quad \forall\, i\in\alpha\; j\notin\alpha\,,$$

$$P_{ij}^{\alpha}=0 \text{ otherwise}. \tag{3}$$

Note the weight functions satisfy the condition $\sum_{\alpha}P_{ij}^{\alpha}=1\;\forall\, i,j$. The density matrix in the region of the subsystem can be approximated to high accuracy by a projection of the Fock matrix in a basis set localized in the neighborhood of that subsystem. The nature of the wavefunctions that extend beyond the boundary of the subsystem necessitate the use of a *buffer region*, in which basis functions on atoms are included in the projection. The approximate density matrix is expressed as

$$\rho_{ij}\approx\sum_{\alpha}P_{ij}^{\alpha}\rho_{ij}=\sum_{\alpha}P_{ij}^{\alpha}\sum_{m}n_{m}^{\alpha}C_{im}^{\alpha\,*}C_{jm}^{\alpha} \tag{4a}$$

and

$$n_{m}^{\alpha}=\left\{1+\exp\left[\left(\varepsilon_{m}^{\alpha}-\mu\right)/kT\right]\right\}^{-1}, \tag{4b}$$

where P_{ij}^{α} is the locally projected density matrix in the α^{th} subsystem, n_{m}^{α}, ϵ_{m}^{α}, and $\{C_{m}^{\alpha}\}$ are the corresponding occupation numbers, eigenvalues, and eigenvectors, respectively, k is the Boltzmann constant, and T is the absolute temperature. The occupation numbers are taken to be Fermi functions ($T=$ 298 K) with chemical potential μ chosen so that normalization of the electron density is maintained. The procedure proceeds iteratively as in conventional methods until self-consistency is achieved.

For a realistic quantum mechanical representation of biological macromolecules, inclusion of solvent effects are crucial. There is a rich literature on the incorporation of the solvent reaction field into quantum mechanical calculations [6–10]. A recently proposed solvent model that has several advantages for quantum mechanical methods is the conductor-like screening model (COSMO) [1]. The method has proven useful for modeling solvent effects of high dielectric media for small molecules [11–14]. The method is based on a variational principle for the electrostatic energy of a charge distribution contained in a cavity surrounded by a conductor ($\varepsilon=\infty$):

$$E_{\text{el}}=\tfrac{1}{2}\boldsymbol{q}^{\mathrm{T}}\cdot\mathbf{A}\cdot\boldsymbol{q}+\boldsymbol{q}^{\mathrm{T}}\cdot\mathbf{B}\cdot\boldsymbol{Q}+\tfrac{1}{2}\boldsymbol{Q}^{\mathrm{T}}\cdot\mathbf{C}\cdot\boldsymbol{Q}\,, \tag{5}$$

where the vectors $\boldsymbol{q}$ and $\boldsymbol{Q}$ represent the reaction field surface charge distribution, and the solute charge density, respectively, and the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ describe electrostatic interactions between the vectors. The surface charge distribution q is determined by the condition that the total electrostatic energy is minimized, $\partial E_{\text{el}}/\partial q=0$, which leads to the linear equation

$$\mathbf{A}\cdot\boldsymbol{q}=-\mathbf{B}\cdot\boldsymbol{Q}\,. \tag{6}$$

The conventional approach is to solve explicitly for $\boldsymbol{q}$ through inversion of the matrix $\mathbf{A}$ [1]. This

leads to a convenient solution of the Green's function for the problem

$$E_{el} = \tfrac{1}{2}\boldsymbol{Q}^{T} \cdot \mathbf{G} \cdot \boldsymbol{Q}, \tag{7a}$$

$$\mathbf{G} = \mathbf{C} - \mathbf{B}^{T} \cdot \mathbf{A}^{-1} \cdot \mathbf{B}. \tag{7b}$$

For finite dielectric, the surface charges are scaled by a factor $f(\varepsilon) = (\varepsilon - 1)/(\varepsilon + \delta)$ [1], where δ is a number between 0 and 2 which we take here to be 0 in accord with the Gauss theorem. The use of Eq. (7b) is convenient for self-consistent field (SCF) calculations of small molecules where the inverse can be computed once at the beginning of the iterative procedure and stored. Subsequent iterations require only matrix multiplications. Unfortunately, for large molecules, this approach is not possible. This results from the fact that the computational effort of the matrix inversion is $\mathcal{O}(M^3)$ and storage requirement is $\mathcal{O}(M^2)$, where M is the dimensionality of the surface charge vector $\boldsymbol{q}$. For biological macromolecules, M will vary as $N^{2/3}$ to N, where N is the number of atoms. For 3-dimensional structures such as globular proteins, the accessible surface area (an hence M) will generally vary as $N^{2/3}$, whereas linear molecules such as DNA or two-dimensional structures such as lipid bilayers will vary as N. Clearly, the above procedure will become limiting for these systems.

We have developed an alternate approach that has favorable computational and memory scaling properties, and has been demonstrated to be accurate and efficient for semiempirical calculations of biological macromolecules. The strategy that we adopt is to minimize the quadratic function Eq. (5) directly using conjugate gradient methods [15]. We are guaranteed convergence after M iterations since the matrix $\mathbf{A}$ is symmetric, nonsingular, and positive definite. This is in contrast to the matrices involved in other boundary element methods that are neither symmetric nor nonsingular [6,7]. The idea is to perform a set of successive line minimizations in directions that are orthogonal. For a quadratic functional such as Eqs. (7), this can be accomplished by the following iterative procedure:

At the first iteration $k = 0$, a starting guess for the solution vector $\boldsymbol{q}_0$ is provided (possibly the null vector), and the initial search direction $\boldsymbol{\delta q}_0$ is chosen in the direction of the downhill gradient,

$$\boldsymbol{\delta q}_0 = -\nabla_q E(\boldsymbol{q}_0) = \left(\boldsymbol{\phi}^{Q} - \mathbf{A} \cdot \boldsymbol{q}_0\right), \tag{8}$$

where $\boldsymbol{\phi}_0$ is the vector representing the potential $\mathbf{B} \cdot \boldsymbol{Q}$ due to the solute charge $\boldsymbol{Q}$ at the surface charge positions. At the k^{th} iteration, an improved solution vector $\boldsymbol{q}_{k+1} = \boldsymbol{q}_k + \alpha_k \cdot \boldsymbol{\delta q}_k$ is determined by the condition $\partial E_{el}(\boldsymbol{q}_k + \alpha_k \cdot \boldsymbol{\delta q}_k)/\partial \alpha_k = 0$ that leads to the expression for α_k

$$\alpha_k = -\frac{\boldsymbol{\delta q}_k^{T} \cdot \left(\mathbf{A} \cdot \boldsymbol{q}_k - \boldsymbol{\phi}^{Q}\right)}{\boldsymbol{\delta q}_k^{T} \cdot \mathbf{A} \cdot \boldsymbol{\delta q}_k}. \tag{9}$$

The new direction is taken as the residual $\boldsymbol{\delta q}_{k+1} = \boldsymbol{\phi}^{Q} - \boldsymbol{A} \cdot \boldsymbol{q}_{k+1}$. The procedure is repeated iteratively until a converged solution is reached.

The rate of convergence of the procedure will be increased when the matrix $\mathbf{A}$ resembles the unit matrix. Consequently, a preconditioner matrix can be used to multiply both side of Eq. (6) to obtain a slightly modified set of Eq. (9):

$$\alpha_k = -\frac{\boldsymbol{\delta q}_k^{T} \cdot \tilde{\mathbf{A}}^{-1} \cdot \left(\mathbf{A} \cdot \boldsymbol{q}_k - \boldsymbol{\phi}^{Q}\right)}{\boldsymbol{\delta q}_k^{T} \cdot \mathbf{A} \cdot \tilde{\mathbf{A}}^{-1} \cdot \boldsymbol{\delta q}_k}, \tag{10}$$

where $\tilde{\mathbf{A}}$ is the preconditioner that presumably satisfies

$$\tilde{\mathbf{A}}^{-1} \cdot \mathbf{A} \approx \mathbf{1}$$

In order for the preconditioned conjugate gradient method to have a computational scaling advantage, the required matrix multiplications of the form $\mathbf{A} \cdot \boldsymbol{x}$ must be reduced from $\mathcal{O}(M^2)$ procedures. If these multiplications remain $\mathcal{O}(M^2)$, no formal scaling advantage is achieved since M iterations are required in principle to arrive at the exact minimum. In practice we observe that for the molecular systems considered here the number of iterations required for a given level of accuracy varies much less severely than $\mathcal{O}(M)$, and hence some advantage may still be achieved. Although the Coulomb interaction matrix $\mathbf{A}$ is not sparse, the matrix multiplication $\mathbf{A} \cdot \boldsymbol{q}$ that represents the Coulombic potential of the surface charge vector, $\boldsymbol{q}$, can be accomplished in $\mathcal{O}(M)$ or $\mathcal{O}(M * \log(M))$ effort using fast-multipole methods [16]. Similar strategies have been adopted for computation of the electrostatic field in boundary element methods using grid-based multipole expansions [17]

and fast multipole methods [18]. Since the surfaces of biological macromolecules are highly irregular, standard fast-multipole methods that require building a hierarchy of cells is cumbersome. An ideal method for this purpose is the simple recursive bisection method proposed by Pérez-Jordá and Yang [19]. This method is adaptive in the sense that sets of particles are recursively divided by splitting planes along the axis of minimal rotational inertial. The method has been demonstrated to be highly efficient, linear-scaling, and require essentially no overhead relative to direct methods even with very small particle distributions.

To accelerate convergence, the preconditioner matrix of Eq. (10) was chosen as the block diagonal Coulomb interaction matrix consisting of surface elements belonging to the same atom. The inverse of this matrix and storage requirements are linear. Additional speed up can be achieved by taking advantage of a short-range neighbor list of surface charge elements and using the preconditioned conjugate gradient method recursively. The matrix multiplications are simple using a short-range cutoff in real space, and hence scale linearly. The disadvantage of this approach is that additional memory is required to store the lists of neighbor charges, although the requirement is still linear.

3. Parameterization of atomic radii

The solvent reaction field in the COSMO model will depend heavily on the solute cavity. A frequently employed convention is to define the cavity as the volume enclosed by the accessible surface defined by a set of effective atomic radii [20]. We employ the surface segmentation strategy described in the original COSMO formulation [1] and implemented in the MOPAC software package [21]. The surface is defined as the surface accessible to the center of a solvent probe with radii R^{solv} around the van der Waals surface formed by the effective atomic radii, minus a distance δ^{sc} that relates to the distance from the probe center where the dielectric screening effect begins. We take R^{solv} to be the radius of a water molecule (1.4 Å), and choose δ^{sc} equal to R^{solv} as is the default in MOPAC and the convention employed previously [1]. The cavity surface was constructed at a discretization level of 60 surface elements per atomic sphere for all molecules. In all calculations the AM1 Hamiltonian [22] was used with an external dielectric of 78.4. For calculations of small molecules, the SCF convergence criterion was 10^{-5} kcal/mol for the energy. Minimization of Eq. (7) was carried out until the solvation energy was converged to 10 significant figures; a detailed analysis of accuracy and timings for the semiempirical divide-and-conquer method have been reported elsewhere [2].

To determine the atomic radii appropriate for describing solvation of biological macromolecules, we fit calculated solvation free energies to experimental and theoretical data for a series of molecules representing amino acid backbone and side chain components [20,23–25], nucleic acid bases [26], and phosphate groups [27]. For most of the neutral molecules, it is possible to measure partition coefficients of the vapor to water transfer, from which solvation free energies can be derived. The process of solvation defined in this way includes a term that reflects the free energy necessary to form a cavity in the solvent. Frequently this term is assumed to be proportional to the accessible surface of the molecule [20],

$$\Delta G_{\mathrm{cav}} = \gamma \cdot \mathrm{SA} + b\,, \tag{11}$$

where SA is the accessible surface area in $\mathrm{Å}^2$, and γ and b are regression parameters (other more sophisticated models for the cavitation term are possible, for example see Ref. [7]). We take γ to be 5 cal/mol · Å and b to be 1 kcal/mol as given by others [20]. Atomic radii were fit to the experimental and theoretical data shown in Table 1. For neutral molecules, one parameter per atom (H,C,N,O,S) was sufficient to obtain a reliable fit (Fig. 1). For ionic guanidinium, carboxylate, and phosphate groups, additional parameters were needed, as in other work [20,28]. The overall fit for neutral molecules is 0.8 kcal/mol, and for DNA bases is 0.9 kcal/mol. The largest errors occur for ethanol (1.88 kcal/mol) and N-methylacetamide (1.82 kcal/mol). It is likely that for ethanol and methanol additional stabilization in solution arises from specific hydrogen bonding interactions with water. The solvation free energy of N-methylacetamide is a topic of some controversy, since the experimental solvation free energy be-

comes less favorable with the removal of a methyl group to form acetamide [23]. This result has recently been brought into question by theoretical free energy perturbation simulations that suggest removal of the methyl group makes solvation more favorable [29] (for additional discussion, see Refs. [7–10]).

Because of the low volatility of nucleic acids, little experimental solvation free energy data is currently available for these compounds. Consequently, we have fit results to theoretical data provided from free energy perturbation simulations [26]. The overall fit to the nucleic acid bases is very close to that for the neutral molecules, suggesting that the data and fit are most likely reasonable. It is noteworthy that in all cases the fitted solvation free energies are slightly greater than the data for nucleic acid bases, the

Table 1
Solvation free energies of amino and nucleic acid components

Molecule	Residue	ΔG_{cav}	ΔG_{el}	ΔG_{calc}	ΔG_{exp}	err
Neutral amino acids						
methane	Ala	1.69	−0.13	1.56	1.93	0.38
propane	Val	2.03	−0.07	1.96	1.98	0.01
isobutane	Leu	2.15	−0.09	2.06	2.28	0.22
butane	Ile	2.19	−0.08	2.11	2.15	0.04
toluene	Phe	2.31	−3.47	−1.16	−0.89	0.27
4-cresol	Tyr	2.39	−6.87	−4.48	−6.13	1.65
3-methylindole	Trp	2.56	−8.92	−6.36	−5.91	0.45
methanol	Ser	1.78	−5.53	−3.75	−5.11	1.37
ethanol	Thr	1.96	−5.09	−3.13	−5.01	1.88
methanethiol	Cys	1.93	−3.56	−1.63	−1.24	0.39
methylethyl sulfide	Met	2.22	−3.08	−0.86	−1.49	0.63
acetic acid	Asp	1.98	−10.09	−8.11	−6.70	1.42
propionic acid	Glu	2.13	−9.28	−7.15	−6.47	0.68
acetamide	Asn	2.01	−12.54	−10.53	−9.71	0.82
propionamide	Gln	2.16	−11.33	−9.17	−9.41	0.24
N-butylamine	Lys	2.30	−5.19	−2.90	−4.29	1.39
N-propylguanidine	Arg	2.48	−12.77	−10.29	−10.91	0.62
4-methylimidazole	His	2.19	−12.99	−10.80	−10.25	0.55
N-methylacetamide [a]	bb	2.17	−10.44	−8.26	−10.08	1.82
Ionized amino acids						
acetate ion	Asp (−)	1.95	−83.19	−81.24	−80.65	0.59
propionate ion	Glu (−)	2.09	−80.58	−78.49	−79.12	0.63
N-butylammonium	Lys (+)	2.32	−70.18	−67.86	−69.24	1.38
N-p-guanidinium	Arg (+)	2.30	−68.07	−65.77	−66.07	0.30
Nucleic acids [b]						
9-methyladenine	A	2.56	−18.06	−15.50	−13.6	1.9
1-methylthymine	T	2.52	−15.55	−13.03	−12.4	0.6
9-methylguanine	G	2.65	−26.08	−23.43	−22.4	1.0
1-methylcytosine	C	2.42	−21.27	−18.84	−18.4	0.4
1-methyluracil	U	2.39	−16.86	−14.47	−14.0	0.5
H_2PO_4 (−)	PO_4 (−)	1.97	−112.97	−111	−111	0

Free energies are in Kcal/mol. Atomic radii used to determine the accessible surface and $\epsilon = 1$ cavity were: H = 1.11, C = 1.42, N = 1.54, O = 1.46, S = 2.07 Å; acetate group CO_2^-: C = 1.63, O = 1.52 Å; guanidinium NH^+: H = 0.78 Å; and phosphate group PO_4^-: O = 1.35, P = 1.67 Å.

[a] Calculations are for trans-N-methylacetamide.

[b] Solvation free energies for methylated nucleic acid bases were obtained from free energy perturbation simulations, except for 9-methyladenine which was estimated from solubility data (Ref. [26]).

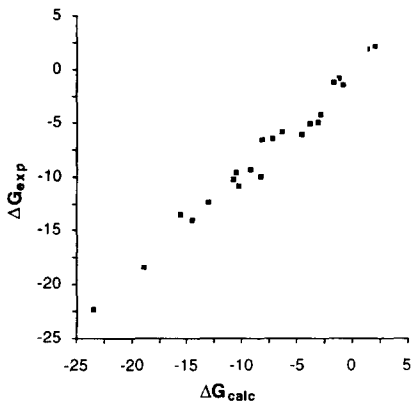

Fig. 1. Regression of experimental and calculated solvation free energies (kcal/mol) for neutral molecules listed in Table 1.

largest difference being with the value for 9-methyladenine estimated from solubility data [26].

4. Calculations of macromolecules

The parameterized radii were applied to calculations of several protein systems using the divide-and-conquer MOPAC (DAC/MOPAC) program described by Lee, York, and Yang [2], modified to use the preconditioned conjugate gradient/recursive bisection method (PCGRB/COSMO) described here. The program has been further modified to read in density-matrix fragments from an amino- and nucleic acid fragment library in order to construct an approximate macromolecular density matrix that can be used as an initial guess to the SCF density matrix, or else to obtain an estimate of the solvation energy in the absence of electronic relaxation (see Appendix). Application of the methods described here to the study of aqueous polarization effects on biological macromolecules is the subject of further work [3]. Here, timings for several large protein systems ranging from over 438 to 4380 atoms both in gas-phase and in solution are presented (Fig. 2). The CPU overhead of the solvation calculations is about 30% relative to the corresponding gas-phase calculations. The CPU effort for a given number and type of particles depends on the molecular shape. The computational effort of the DAC/MOPAC method and the PCGRB/COSMO model are oppositely affected by shape. The efficiency of the DAC/MOPAC method is favored for linear systems that have small buffer regions, and slowest for compact 3-dimensional structures. The efficiency of the PCGRB/COSMO model, on the other hand, is favored for compact structures that have the least surface area. The opposite shape dependence of the methods to some extent balance one another making the overall CPU time for the solvation calculations in Fig. 2 appear more uniformly linear than the corresponding gas-phase values.

It is noteworthy to point out that the timing in Fig. 2 are in terms of the CPU per iteration. It is in general difficult to estimate the scaling behavior of the number of SCF iterations as a function of system size. The total number of iterations will depend on factors such as the conditioning of the Fock matrix, the density of states near the Fermi level, the molecular configuration and charge state of the system, and the minimization or mixing scheme that is employed. We observe no systematic trend in the number of SCF iterations with system size for the macromolecules studied here. For example, bpti (892 atoms), subtilisin BPN (3837 atoms), and superoxide dismutase (4380 atoms) required 12, 10, and 11 iterations in solution, respectively (the typical range was 10 to 14). It is noteworthy that inclusion of solvation tends to stabilize the valence states and

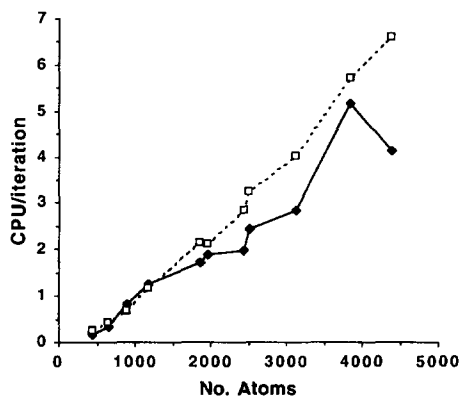

Fig. 2. CPU time per SCF iteration (10^3 s) for several protein molecules: amyloid β-peptide (438 atoms), crambin (642 atoms), bpti (892 atoms), calbindin P43G (1195 atoms), ribonuclease A (1856 atoms), lysozyme (1960 atoms), interleukin 1-β (2437 atoms), C-MYB/DNA complex (2512 atoms), hiv-1 protease dimer (3118 atoms), subtilisin BPN (3837 atoms), and superoxide dismutase (4380 atoms). Timings are given in the gas phase (solid line), and in solution (dashed line). All calculations were performed on a dedicated IBM RISK6000 workstation (128 MB memory).

results in faster convergence than a gas-phase SCF calculation starting from the same density matrix.

5. Conclusion

We have presented an efficient method for inclusion of solvent effects in linear-scaling electronic structure calculations of biological macromolecules. The method is demonstrated to have favorable scaling behavior, and requires approximately 30% CPU overhead relative to gas-phase calculations. The methods developed here and elsewhere [2,3] allow quantum mechanical calculations of systems over 4000 atoms in solution to be treated on a typical workstation. A simple set of atomic radii parameters were also developed that reproduce to reasonable accuracy (less than 1 kcal/mol on average) solvation free energies of small molecules homologous to amino- and nucleic acid residues. We hope this work stimulates the application of quantum chemical methods to the study of biological macromolecules in solution.

Acknowledgements

The authors gratefully acknowledge P.A. Kollman for providing free energy data prior to publication. The authors acknowledge financial support from a subcontract agreement with the NIH Research Resource Program at the University of North Carolina at Chapel Hill. This work was supported by the U.S. Environmental Protection Agency, the National Science Foundation, and the Exxon Education Foundation. DY acknowledges partial support through an NSF postdoctoral fellowship at Duke, and a current NIH postdoctoral fellowship at Harvard. WY acknowledges support from the Alfred P. Sloan Foundation.

Appendix A

A.1. Fragment density matrices

It is often useful to construct an approximate density matrix that represents a first order approximation to the macromolecular charge distribution [30–33]. Here we adopt a simple procedure for constructing an approximate density matrix for biological macromolecules based on a fragment library. The method is similar in spirit to conventional molecular mechanical approaches that employ empirical point charge representations for the gas-phase charge distribution of isolated residues or fragments, and assemble these fragments to obtain the macromolecular charge density and potential. These empirical methods are routinely used to provide information about solvation energies, pK_a shifts, and electrostatic potential surfaces of biological macromolecules [34]. In the present method, we abandon the empirical point charge description and instead use directly the fragment density matrix elements. For the purposes of solvation energies, only the 'atom-diagonal' density-matrix elements (elements on the same center) are used, since in the current implementation of MOPAC and our modified program, only these elements are coupled directly to the reaction field. Hence, the macromolecular fragment potential is represented by a set of atomic monopole, dipole, and quadrupole components.

A.2. Construction of the fragment library

For biological macromolecules, the basic set of building blocks consists of the amino acids and ribo- and deoxyribonucleotides. Calculations were performed on individual residues with end groups designed to mimic neighboring residues. In the case of the amino acids, ends were capped with CH_3CO- and CH_3NH- groups. Charged amino- and carboxy-terminal groups were derived from the alanine zwitterion. For each of the nucleic acids (dA,dT,dC,dG, rA,rU,rC,rG), a three-unit single strand of stacked bases were used to derive density matrix elements for nucleotides at the 3′ and 5′ ends as well as the stacked (non terminal) nucleotides. The diagonal elements of the density matrix for each residue were scaled to enforce residue normalization and reduce systematic errors. The atom and residue type names, coordinates, and density matrix elements for each molecule were stored in a fragment library.

A.3. Construction of the fragment density

For each atom i, the fragment library was searched until a atom/residue match was found, i_{frag}. The

orthogonal transformation (rotation) that orients the fragment density matrix elements into the reference frame of the macromolecule was then determined. The rotation was defined by the requirement that a set of nearest neighbor *reference* atoms for i_{frag} were best fitted to the corresponding reference set for atom i in the macromolecular structure. Usually the set of first nearest neighbors is sufficient to define a unique best-fit rotational mapping. However, in certain cases, such as a carbonyl oxygen that has only one covalently bonded neighbor, the rotation is not unique (as is the case for any linear system). In these instances, second nearest neighbors were included in the fitting set. For biological macromolecules, we did not observe a need to go beyond second nearest neighbors. Once a set of fitting neighbors was defined, the fragment library atom i_{frag} and its associated reference atoms were translated so that atoms i and i_{frag} are superposed. The rotation matrix $\mathbf{R}_i$ was then determined that minimized the rms positional deviation of the macromolecular and fragment library reference atoms using the algorithm of Mclachlan [35].

For most semiempirical methods, the density matrix consists of elements formed from s- and p-type basis function products. We denote the transformed and untransformed density matrix elements consisting of x-type basis function products on centers i and j as $\rho_{ij}^{x_i x_j}$ and $\tilde{\rho}_{ij}^{x_i x_j}$, respectively. The prescription for obtaining the transformed density matrix elements is given by

$$\rho_{ij}^{\text{ss}} = \tilde{\rho}_{ij}^{\text{ss}}\,, \tag{12a}$$

$$\boldsymbol{\rho}_{ij}^{\text{sp}} = \tilde{\boldsymbol{\rho}}_{ij}^{\text{sp}} \cdot \mathbf{R}_j^T\,, \tag{12b}$$

$$\boldsymbol{\rho}_{ij}^{\text{pp}} = \mathbf{R}_i \cdot \tilde{\boldsymbol{\rho}}_{ij}^{\text{pp}} \cdot \mathbf{R}_j^T\,. \tag{12c}$$

In this way, a macromolecular density-matrix charge distribution can be assembled from a fragment density-matrix library. This representation generally overestimate the charge separation of interacting polar and ionic groups that if allowed to relax quantum mechanically, would usually tend to equalize.

References

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

[2] T.-S. Lee, D.M. York and W. Yang, J. Chem. Phys. 105 (1996) 2744.

[3] D.M. York, T.-S. Lee and W. Yang, J. Am. Chem. Soc., in press.

[4] W. Yang, Phys. Rev. Lett. 66 (1991) 438.

[5] W. Yang and T.-S. Lee, J. Chem. Phys. 103 (1995) 5674.

[6] J. Tomasi and M. Persico, Chem. Rev. 94 (1994) 2027.

[7] C.J. Cramer and D.G. Truhlar, in: Reviews in Computational Chemistry, Vol. VI, eds. K.B. Lipkowitz and D.B. Boyd (VCH Publishers, New York, 1995).

[8] D. Tannor, B. Marten, R. Murphy, R.A. Friesner, D. Sitkoff, A. Nicholls, M. Ringnalda, W.A. Goddard, III, and B. Honig, J. Am. Chem. Soc. 116 (1996) 11875.

[9] J.L. Chen, L. Noodleman, D.A. Case and D. Bashford, J. Phys. Chem. 98 (1994) 10059.

[10] B. Marten, K. Kim, C. Cortis, R.A. Friesner, R.B. Murphy, M.N. Ringnalda, D. Sitkoff and B. Honig, J. Phys. Chem. 100 (1996) 11775.

[11] A. Klamt, J. Phys. Chem. 99 (1995) 2224.

[12] J. Andzelm, C. Kölmel and A. Klamt, J. Chem. Phys. 103 (1995) 9312.

[13] T.N. Truong and E.V. Stefanovich, Chem. Phys. Lett. 240 (1995) 253.

[14] H.S. Rzepa and G.A. Suner, J. Chem. Soc., Perkin Trans. 7 (1994) 1397.

[15] W.H. Press, S.A. Teukolski, W.T. Vetterling and B.P. Flannery, Numerical Recipes in FORTRAN: the Art of Scientific Computing, 2nd Ed. (Cambridge Univ. Press, 1992).

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

[17] R.J. Zauhar and A. Varnek, J. Comput. Chem. 17 (1996) 864.

[18] R. Bharadwaj, A. Windemuth, S. Sridharan, B. Honig and A. Nicholls, J. Comput. Chem. 16 (1995) 898.

[19] J.M. Pérez-Jordá and W. Yang, Chem. Phys. Lett. 247 (1995) 484.

[20] D. Sitkoff, K.A. Sharp and B. Honig, J. Phys. Chem. 98 (1994) 1978.

[21] J.J. P. Stewart, J. Computer-aided Molecular Design 4 (1990) 1.

[22] M.J. S. Dewar, E.G. Zoebisch, E.F. Healy and J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902.

[23] R. Wolfenden, L. Andersson, P.M. Cullis and C.C.B. Southgate, Biochemistry 20 (1981) 849.

[24] S. Cabani, P. Gianni, V. Mollica and L. Lepori, J. Solution Chem. 10 (1981) 563.

[25] A. Ben-Naim and Y. Marcus, J. Chem. Phys. 81 (1984) 2016.

[26] J.L. Miller and P.A. Kollman, J. Phys. Chem. 100 (1996) 8587.

[27] Y. Marcus, J. Chem. Soc. Faraday Trans. 87 (1991) 2995.

[28] E.V. Stefanovich and T.N. Truong, Chem. Phys. Lett. 244 (1995) 65.

[29] P.A. Kollman and P.-Y. Morgantini, J. Am. Chem. Soc. 117 (1995) 6057.

[30] J.-G. Lee and R.A. Friesner, J. Phys. Chem. 97 (1993) 3515.

[31] P.D. Walker and P.G. Mezey, J. Am. Chem. Soc. 115 (1993) 12423.

[32] L. Massa, L. Huang and J. Karle, Int. J. Quant. Chem. 29 (1995) 371.

[33] S.R. Gadre, R.N. Shirsat and A.C. Limaye, J. Phys. Chem. 98 (1994) 9165.

[34] B. Honig and A. Nicholls, Science 268 (1995) 1144.

[35] A.D. Mclachlan, J. Mol. Biol. 128 (1979) 49.