Quantum Mechanical Computations on Very Large Molecular Systems: The Local Self-Consistent Field Method

VINCENT THÉRY, DANIEL RINALDI, JEAN-LOUIS RIVAIL,* BERNARD MAIGRET, and GYÖRGY G. FERENCZY†

Laboratoire de Chimie Théorique UA CNRS 510, Université de Nancy I, Domaine Scientifique Victor Grignard B.P. 239, 54506 Vandœuvre les Nancy cedex France

Received 18 June 1993; accepted 26 September 1993

ABSTRACT

Quantum chemical computations on a subset of a large molecule can be performed, at the neglect of diatomic differential overlap (NDDO) level, without further approximation provided that the atomic orbitals of the frontier atoms are replaced by parametrized orthogonal hybrid orbitals. The electrostatic interaction with the rest of the molecule, treated classically by the usual molecular mechanical approximations, is included into the self-consistent field (SCF) equations. The first and second derivatives of energy are obtained analytically, allowing the search for energy minima and transition states as well as the resolution of Newton equations in molecular dynamics simulations. The local self-consistent field (LSCF) method based on these approximations is tested by studying the intramolecular proton transfer in a Gly-Arg-Glu-Gly model tetrapeptide, which reveals an excellent agreement between a computation performed on the whole molecule and the results obtained by the present method, especially if the quantum subsystem includes the side chains and the peptidic unit in between. The merits of the LSCF method are examplified by a study of proton transfer in the Asp^{69}—Arg^{71} salt bridge in dihydrofolate reductase. Simulations of large systems, involving local changes of electronic structure, are therefore possible at a good degree of approximation by introducing a quantum chemical part in molecular dynamics studies. This methodology is expected to be very useful for reactivity studies in biomolecules or at the surface of covalent solids.

*Author to whom all correspondence should be addressed.
†Current address: Chemical Works of Gedeon Richter Ltd., P.O. Box 27, H 1475, Budapest, Hungary.

Journal of Computational Chemistry, Vol. 15, No. 3, 269–282 (1994)

CCC 0192-8651/94/030269-14

Introduction

It often happens that one needs quantitative information on a part of a large molecular system, which can only be obtained by a quantum chemical computation.[1] For instance, in a chemical reaction the electronic properties of the bonds are strongly modified all along the reaction path,[2,3] and this process can hardly be described by the usual molecular mechanics approach. This problem obviously comes under the heading of quantum chemistry, although it may happen that at least one of the reacting molecules is far too large to be treated as a whole by any quantum chemical program. The usual method consists of studying a small molecule which has the adequate minimal structure to be safely considered as a model of the large system. Nevertheless, this approach is difficult since the long-range interactions with the rest of the macromolecule, in particular the electrostatic interactions, play an important role[4] which cannot be neglected. The interactions of a small molecule, treated quantum mechanically, with a large system described classically can be considered,[5] but when this molecule is a model of a fragment separated from the whole system by a mental process which consists of cutting at least one bond and saturating the corresponding valences by extra hydrogen atoms, the boundary between the classical and the quantum part of the system becomes quite artificial and is difficult to treat properly.

There have been many attempts to perform economic quantum chemical computations on large systems. Warshel and Levitt[6] attempted to perform, in the linear combination of atomic orbitals (LCAO) approximation, quantum chemical computations on a fragment interacting with the rest of a macromolecule, described classically, by an SCF approach using orthogonal hybrid atomic orbitals in a simple semiempirical scheme. Recently a mixed valence bond molecular mechanics scheme has been proposed.[7] An interesting method to describe two subsets of a large system at two different levels of approximation in the complete neglect of differential overlap (CNDO) scheme is due to Naray–Szabo et al.[8–10] It consists of describing the largest part of the system at a first level of approximation in the form of strictly localized bond orbitals (SLBOs) and performing a full SCF computation on the fragment of interest by a linear combination of SLBOs. This approach has been generalized recently[11] in the NDDO formalism,[12] leading to a computational scheme which has been tested extensively and appears adequate to describe correctly the electronic properties of a fragment embedded in a larger molecular system. Nevertheless, this method appears not well adapted to the computation of very large molecular systems, like enzymes, because the computational time requested by the SLBO increases rapidly with the size of the molecule. On the other hand, a useful method requires the variations of energy with atomic coordinates to be reproduced correctly in order to allow the search of stationary geometries or the computation of forces, and the SLBO approximation would probably need a long and risky parametrization process.

In the present study, we show that this approximation can be simplified further to propose an efficient mixed classical quantum mechanical method called LSCF. This method meets with the present-day requirements of computational chemistry, which not only include the possibility of computing the wave function and energy for a given molecular geometry but also give the derivatives of energy with respect to atomic coordinates in order to analyze the potential energy surface or to define the forces acting on the atoms in view of a molecular dynamics simulation of the whole system.

This article focuses on the quantum chemical part of the problem, especially in the energy variations in a chemical reaction occurring in large polypeptides. In the following sections, we first develop the principles of the method. We then give some practical information about its implementation. The method is then tested on a peptide, small enough to allow full quantum chemical computations used as reference data. Finally, we study a simple reaction occurring in a large protein as an example of application.

Principles of the Method

In this section, we develop the basic assumptions and the equations of the method. The principles have already been outlined briefly.[13] The system is assumed to be divided into two parts: the fragment or subsystem S, which must be treated quantum mechanically, and its environment E, which is usually the rest of the macromolecule. Let us now consider one pair of bonded atoms XY at the boundary between these two parts. For obvious

chemical reasons, X—Y is assumed to be a single bond. The atom belonging to S is called the frontier atom is assumed to have only s and p atomic orbitals in its valence shell denoted by $|s\rangle$, $|x\rangle$, $|y\rangle$, $|z\rangle$, respectively. Y is the atom of this bond belonging to E. The subsystem can share several such X—Y bonds with its environment.

In our previous approach,[11] the two electrons of the X—Y bond are represented by an SLBO, which is a linear combination of two hybrid orbitals, one $|l>$ defined on X, the other defined on Y.

If $|l>$ belongs to a set of four orthogonal hybrid orbitals defined on X, the three other hybrids—$|i>, |j>, |k>$, respectively—can be used in a basis set, together with the other atomic orbitals of the atoms belonging to S, to compute molecular orbitals describing the electrons of the subsystem. In the NDDO approximation, the molecular orbitals are orthogonal to all the SLBOs which are supposed to describe the valence electrons in the environment, including that which would be built with $|l>$.

The basic idea of this new method is to avoid the computation of the localized bond orbitals of the environment by assuming the transferability of the bond properties according to the basic principles of molecular mechanics. The SLBO corresponding to the X—Y bond is assumed to be transferable from a fully computed model system, provided that it is orthogonal to the molecular orbitals of the subsystem. This condition is fulfilled in the NDDO approximation if the orbitals of X entering the SCF computation remain, as above, orthogonal to $|l>$. One can further simplify the model by replacing the interactions of the electrons in S with the electronic population of atom Y by a purely electrostatic term computed with the net charge Q_Y of atom Y treated exactly like the other atoms in E.

The important step is the definition of the contribution of atom X to the XY SLBO (i.e., the hybrid orbital $|l>$ and its coefficient c_l or equivalently its diagonal density matrix element $P_{ll} = 2c_l^2$). The hybrid orbital $|l>$ has the general form

$$|l> = a_{l1}|s> + a_{l2}|x> + a_{l3}|y> + a_{l4}|z> \quad (1)$$

The parameters a_{l1}, a_{l2}, a_{l3}, and a_{l4} are compelled to fulfill the following conditions:

1. The hybrid has to be directed toward Y so that a_{l2}, a_{l3}, and a_{l4} are proportional to the cosines of the X—Y direction.
2. The s contribution to the normalized hybrid is assumed to be a transferable property of the X—Y bond so that the coefficient a_{l1} is a precalculated parameter.
3. The four coefficients are bound together by the normalization condition of $|l>$.

The only numerical parameter needed to define the hybrid is therefore the coefficient a_{l1} of the normalized combination shown in eq. (1). This quantity, as well as the matrix element P_{ll}, can be either extracted from a localized orbital computed on a model system or treated as a pre-defined parameter. Typically, a_{l1} is expected to be close to 0.5 for an sp^3 hybridization and P_{ll} close to 1 for a nonpolarized σ bond.

The hybrid $|l>$, which does not enter the SCF computation, will be denoted as the excluded orbital. In the present study, we limit ourselves to the case in which an atom has only one excluded orbital. In this case, the three other orbitals only need to be orthogonal to $|l>$ and can be defined by any convenient orthogonalization procedure.

The use of a basis set containing atomic orbitals and hybrid atomic orbitals simultaneously is not a formal problem. It only requires a transformation of the usual integrals, which is straightforward. Let a_{i1}, a_{i2}, a_{i3}, a_{i4} be the coefficients of the hybrid orbital $|i>$ of X in the basis of atomic orbitals denoted $|\theta> |\zeta> \ldots$, and $|\mu> |\nu>$ two atomic orbitals of any atom of S. Then all the integrals involving a one-electron operator $\hat{A}$ are obtained by the following relationships:

$$\langle i|\hat{A}|\mu\rangle = \sum_{\theta=1}^{4} a_{i\theta}\langle\theta|\mathbf{A}|\mu\rangle \quad (2)$$

$$\langle i|\hat{A}|j\rangle = \sum_{\theta=1}^{4}\sum_{\zeta=1}^{4} a_{i\theta}a_{j\zeta}\langle\theta|\mathbf{A}|\zeta\rangle \quad (3)$$

The transformation of two-electron integrals can be given similarly:

$$(ij|\mu\nu) = \sum_{\theta=\alpha}^{\delta}\sum_{\zeta=\alpha}^{\delta} a_{i\theta}a_{j\zeta}(\theta\zeta|\mu\nu) \quad (4)$$

After these transformations, the hybrid orbitals entering the SCF computation behave exactly like any atomic orbital, so that now we shall not distinguish them and shall adopt the general notation $|\lambda> |\mu> |\nu> |\eta>$ for the atomic orbitals. If one assumes, as in most of the usual force fields, that the electrostatic interactions of the environment

can be represented by point charges Q_A located on the atoms A of E, the $(\mu\nu)$ element of the local Fock operator takes the general form

$$F^S_{\mu\nu} = H_{\mu\nu} + \sum_\lambda \sum_\eta P_{\lambda\eta}\left[(\mu\nu|\lambda\eta) - \frac{1}{2}(\mu\eta|\lambda\nu)\right] + \sum_l P_{ll}\left[(\mu\nu|ll) - \frac{1}{2}(\mu l|l\nu)\right] + \sum_{A\in E} Q_A(\mu\nu|s_As_A) \quad (5)$$

In this equation, the last sum represents the electrostatic perturbation of the electron distribution in S by the atomic charges in E in the formalism of the current modified neglect of differential overlap (MNDO),[14] AM1,[15] or PM3[16] methods. At a short distance, it mimics the Coulomb interactions corrected from penetration effects. At a long distance, it is equal to the pure Coulomb interaction.

Under our basic assumptions, if Z_A stands for the number of valence electrons of atom A, P_A is its electron population, and if f_{KA} is the core-core repulsion function between atoms K and A, specific of each semiempirical method, the energy of the system is expressed by

$$U = \sum_\mu \sum_\nu P_{\mu\nu}\left\{H_{\mu\nu} + \frac{1}{2}\sum_\lambda \sum_\eta P_{\lambda\eta}\left[(\mu\nu|\lambda\eta) - \frac{1}{2}(\mu\eta|\lambda\nu)\right] + \sum_l P_{ll}\left[(\mu\nu|ll) - \frac{1}{2}(\mu l|\nu l)\right] + \sum_{A\in E} Q_A(\mu\nu|s_As_A)\right\} + \sum_l \left\{\sum_{l'} \frac{1}{2} P_{ll}P_{l'l'}(ll|l'l') + \sum_{A\in E} Q_A(ll|s_As_A) + \sum_{K\in S} Z_K P_{ll}(ll|s_Ks_K)\right\} \times \frac{1}{2}\sum_{K\in S}\sum_{L\neq K, L\in S} Z_KZ_Lf_{KL} + \sum_{K\in S}\sum_{A\in E} Z_K[Z_Af_{KA} - P_A(s_Ks_K|s_As_A)] + V(E) + V'(E, S) \quad (6)$$

In this expression, one recognizes successively the energy of all the electrons in S including those of the excluded hybrid orbitals, the nuclear interactions in S, the interactions between the nuclei in S and the effective charges in E, and, finally, two classical contributions. The first one, $V(E)$, is simply the interaction between the atoms in E given by the force field. The second one, $V'(E, S)$, is a modified force field, in which the terms corresponding to the electrostatic interaction between atoms in E and atoms in S have been suppressed, since this interaction is already included in the previous terms.

Implementation

The method has been implemented in the GEOMOS semiempirical package[17] using simple modifications of the code when expressed in the matrix formalism.

CONSTRUCTION OF THE HYBRID ORBITALS

On each frontier atom X of S assumed to be a carbon, nitrogen, or oxygen atom, the four orthogonal hybrid orbitals are defined in two steps. First, a normalized hybrid $|l'>$ of $2s$ and $2p_z$ having the $2s$ content a_{l1} defined for the excluded hybrid $|l>$ is built together with the orthogonal combination $|k'>$ of the same orbitals. Two other orthogonal orbitals $|i'>$ and $|j'>$ are simply identified to the $2p_x$ and $2p_y$, respectively. Hence one defines a matrix

$$[A'_x] = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ (1 - a^2_{l1})^{1/2} & 0 & 0 & -a_{l1} \\ a_{l1} & 0 & 0 & (1 - a^2_{l1})^{1/2} \end{bmatrix} \quad (7)$$

such that

$$\begin{bmatrix} |i'\rangle \\ |j'\rangle \\ |k'\rangle \\ |l'\rangle \end{bmatrix} = [A'_x]\begin{bmatrix} |s\rangle \\ |x\rangle \\ |y\rangle \\ |z\rangle \end{bmatrix} \quad (8)$$

Second, a rotation of the reference axes is applied on the four orbitals $|i'>$, $|j'>$, $|k'>$, $|l'>$ to bring the transformed z-axis of $|l'>$ in coincidence with the $\vec{XY}$ orientation. This defines the excluded $|l>$ hybrid, and the transformed of the three other orbitals are assimilated to the three other functions $|i>$, $|j>$, $|k>$, respectively. Let $[R_x]$ be this rotation matrix. The matrix $[A_x]$, which transforms the atomic orbitals of the frontier atoms X into the orthogonal hybrids, is

$$[A_x] = [R_x][A'_x] \quad (9)$$

CONSTRUCTION OF THE LOCAL FOCK MATRIX

Let us assume that the subsystem S has f frontier atoms with the environment E. The original basis set of atomic orbitals defined on the atoms of S contains N functions. The transformed basis set, in which the hybrid orbitals of the frontier atoms replace the atomic orbitals of these atoms, is deduced

from the original one by a simple linear transformation defined by a matrix [B] built by setting the $[A_x]$ matrixes of each frontier atom X at the proper positions on the diagonal (the other diagonal elements being 1 and the off-diagonal ones 0).

The N–f functions defining the fragment basis set, which is used for the molecular orbital computations, are obtained by removing from the former the excluded hybrid orbitals.

Let [P'] be the $N \times N$ density matrix corresponding to the original basis set and [P] the $(N-f) \times (N-f)$ density matrix in the fragment basis set. At a given iteration of the SCF computation, a Fock matrix [F'] is readily defined by applying the standard procedure of the SCF code to the original basis set by using the density matrix [P'] obtained at the previous iteration. The only modification is the addition of the perturbation due to the point charges of the environment, which can be included in the core Hamiltonian.

The Fock matrix $[F'_T]$ corresponding to the transformed basis set is obtained by the linear transformation [B] defined above:

$$[F'_T] = [B]^{-1}[F'][B] \qquad (10)$$

Finally, the $(N-f) \times (N-f)$ local Fock matrix $[F^S]$ defined by eq. (5) is obtained by removing from $[F'_T]$ the rows and the columns corresponding to the excluded orbitals. The diagonalization of $[F^S]$ yields the coefficients of the functions belonging to the fragment basis set which define the fragment molecular orbitals at this iteration. A new density matrix [P] is computed. It is transformed into [P'] by adding the proper P_{ll} values at each diagonal position corresponding to each excluded hybrid orbital and by applying the $[B]^{-1}$ transformation. This [P'] matrix is then used to compute the [F'] matrix for the next iteration, and so on until the convergence criterion is satisfied.

ENERGY AND ENERGY DERIVATIVES

The expression of the energy in eq. (6) is obtained in the standard SCF procedure by a similar matrix transformation and by adding the classical contributions due to the interaction with the environment E.

The first and second derivatives of energy are mainly obtained with the standard GEOMOS code,[17] except for the derivatives with respect to the coordinates of the frontier atoms X and/or the atoms Y bonded to them. In this case, the orientation of the X—Y bond changes and the four hybrids $|i>, |j>, |k>, |l>$ vary. The derivatives can be obtained readily by using the derivatives of the rotation matrixes $[R_x]$, which are currently computed by standard codes.

Test of the Method

The aim of this test is to investigate the capability of the method to reproduce the molecular properties of a fragment of a large system and to evaluate the errors introduced in truncating the system at a given bond. The test system is a model tetrapeptide Gly-Arg-Glu-Gly, in which the free OH of the end glycine has been replaced by an H atom to simulate a fragment of a longer peptidic chain. It is small enough to be investigated completely by a semiempirical quantum chemical technique. The computational method is AM1.[15]

The test consists of comparing the quantitative results obtained in treating the whole molecule and applying the LSCF method to two different partitions of this molecule into the subset S and the environment E. The chemical reaction studied is simply the proton transfer from the carboxyl group of glutamic acid to the guanidine group of arginine. Two stable structures are optimized. In the first one, denoted as neutral structure, the Arg and Glu sidechains are in their neutral form; while in the second one, the acidic proton has been transferred from glutamic acid to arginine (zwitterionic form). In addition, this proton transfer process has been simulated by scanning the OH bond length and optimizing all the other parameters of the sidechains. To obtain comparable results during the test, the geometry of the backbone has been maintained in a fixed position. This geometry is the result of a previous energy refinement on the complete molecule by means of the consistent valency force field (CVFF) of the Discover package.[18] It is represented in Figure 1.

The electrostatic properties of the classical part of the system are represented by point charges located on the atoms. In the first tests, these charges are set equal to the Mulliken charges of these atoms obtained after the computation on the whole molecule. In the second half of this section, we compare the result obtained with other sets of charges.

The reference computation performed on the whole molecule shows (Fig. 2) that the minimum energy corresponds to the neutral form with, for the hydrogen atom location $R_{OH} = 0.973$ Å and $R_{NH} = 2.523$ Å. The zwitterionic form ($R_{OH} = 2.201$

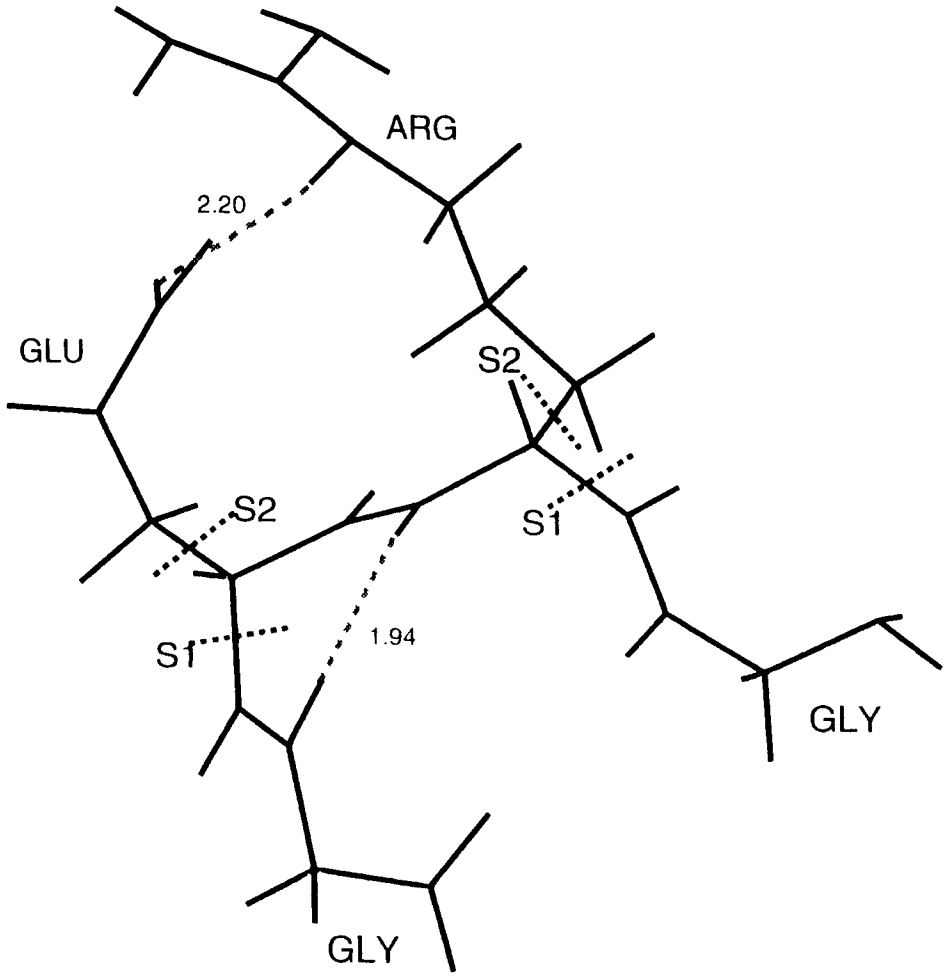

FIGURE 1. Geometry of the tetrapeptide Gly-Arg-Glu-Gly.

Å and R_{NH} = 1.013 Å) is 13.95 kcal · mol^{-1} above the neutral one, and the maximum energy is obtained for R_{OH} = 1.454 Å and R_{NH} = 1.143 Å. It is 19.01 kcal · mol^{-1} above the neutral form. In fact, this maximum is not a true transition state since the backbone has been considered rigid. This constraint explains a sudden change of slope visible in Figure 2 for R_{OH} = 1.750 Å.

INFLUENCE OF THE SIZE OF THE QUANTUM SUBSYSTEM

The use of the LSCF method requires the definition of the subsystem S. In our previous study,[11] we showed that, as expected, the subsystem S must be large enough to minimize the error introduced by truncating the delocalized orbital system. In addition, the frontier atoms must be far enough from the chemical group of interest. In the case of the tetrapeptide considered here, it seems reasonable to define the subsystem as including the two Arg and Glu sidechains and the central peptidic part which links these chains (subsystem S_1). The two frontier atoms are therefore the C_α atoms of arginine and glutamic acid, respectively; and the environment E is made of the two peptidic chains containing the glycine residues (see Fig. 1). The parameters used for the excluded hybrid orbitals are extracted from the computation on the whole molecule after a localization of the molecular orbitals by means of the Ruedenberg technique[19] and the proper normalization to simulate an SLBO. These values vary slightly from the neutral form to the zwitterionic one. We chose average quantities, which are the following:

- Excluded orbital close to arginine (C_α—N bond): α_{l1} = 0.5050; P_{ll} = 0.9100

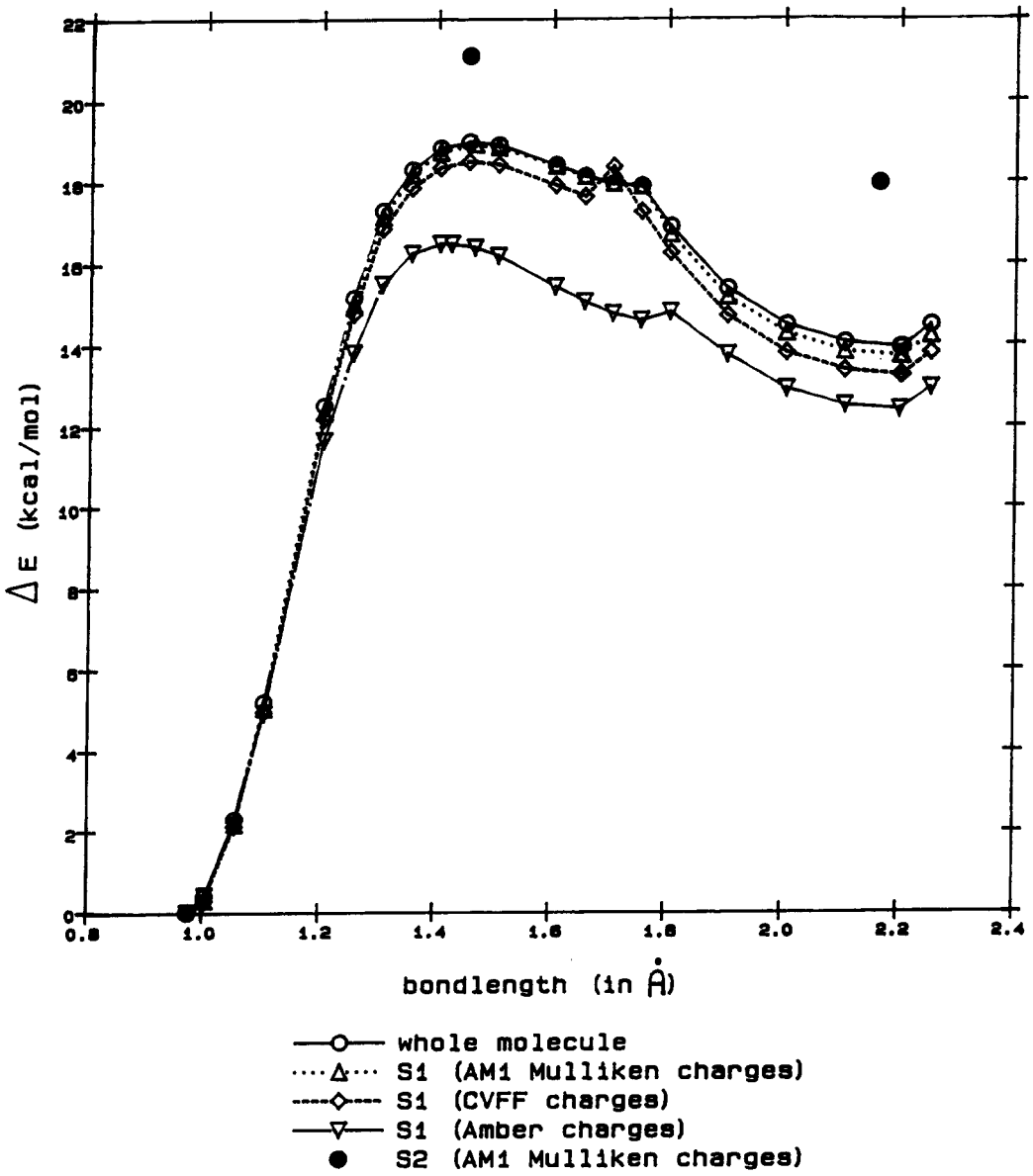

FIGURE 2. Variation of the energy during the proton transfer.

- Excluded orbital close to glutamic acid (C_α—CO bond): $a_{l1} = 0.5500$; $P_{ll} = 0.9300$

As above, we studied the two stable forms and the energy variation from one to the other. The curve plotted in Figure 2 is very close to the previous one but slightly below if the origin of energies is chosen at the lowest minimum. However, one notices that it reproduces all the features of the other one and that the distance between the two minima is now 13.74 kcal · mol^{-1} (i.e., only 0.21 kcal · mol^{-1} less than the "exact value"). The maximum is 18.94 kcal · mol^{-1} above the lowest minimum so that the error is only 0.07 kcal · mol^{-1}.

The interatomic distances are compared with the reference values (in parentheses):

For the neutral form: R_{OH} = 0.975 Å (0.973); R_{NH} = 2.270 Å (2.523)

For the zwitterionic form: R_{OH} = 2.200 Å (2.201); R_{NH} = 1.013 Å (1.013)

For the maximum: R_{OH} = 1.454 Å (1.454); R_{NH} = 1.143 Å (1.143)

TABLE I.
Table of the RMS Deviation of Bond Length, Bond Angle, and Dihedral Angle between the Optimized Parameters of Subsets S_1 and S_2 (Reference: AM1 Results for the Whole Molecule).

	State	Bond Length (pm = 10^{-2} Å)	Bond Angle (degree)	Dihedral Angle (degree)
Subsystem	Neutral	0.1340	0.2780	3.5323
S_1	Transition	0.1186	0.1750	0.1866
	Zwitterionic	0.1290	0.1866	0.8312
Subsystem	Neutral	0.2388	0.4945	4.9871
S_2	Transition	0.7740	4.0832	24.1736
	Zwitterionic	0.2758	0.4452	1.5906

The agreement is almost perfect except for the length of the OH—N hydrogen bond in the neutral form. This difference will be discussed below.

The good agreement between the optimized geometries holds for all the other coordinates. It is expressed by the root mean square (rms) deviations for the bond lengths, bond angles, and dihedral angles. These quantities are listed in Table I, first row.

To check the influence of the subsystem definition on these results, the same computations have been carried out on a reduced subsystem limited now to the sidechains themselves (subsystem S_2). The frontier atoms are thus the C_β atom of each sidechain, and the parameters for the C_β—C_α bond are, again, an average of the values determined after the Ruedenberg localized orbitals of the whole system. Their values are the following:

- Arginine chain: $a_{l1} = 0.5309$; $P_{ll} = 1.044$
- Glutamic acid chain: $a_{l1} = 0.5357$; $P_{ll} = 0.9060$

The rms deviations for the geometry parameters are given in Table I, second row, and the relative energies of the three characteristic points of the potential energy surface are visible in Figure 2. The agreement with the reference data is far less satisfactory than with the previous subsystem. This error may have different origins. The first explanation, inspired by the conclusions of our previous study,[11] is that the subsystem is too small now and that the perturbation introduced by truncating the fragment submitted to the quantum computation is not damped enough at the level of the reacting groups. In addition, this subsystem is made of two disconnected molecular systems, and the electron flow from one end to the other through the central peptidic bond is now impossible. One may also think that the good results obtained in the case of S_1 are just due to chance. To eliminate this explanation, we repeated the computation for a larger subsystem obtained by adding to S_1 the two neighbouring peptidic bonds. The environment is then reduced to the NH2 group at one end and to the COH group at the other end. As expected, the agreement with the reference results is excellent, even for the length of the H bond in the neutral form (2.520 Å).

The conclusion is therefore that the quantum subsystem must have a reasonable size and that limiting this system to the sidechains is not sufficient.

INFLUENCE OF THE ELECTROSTATIC REPRESENTATION OF THE ENVIRONMENT

One notices that in our previous studies the size of the environment is the largest in the case of subsystem S_2, and the poor results obtained in this case may also be due to the definition of the point charges used to represent this environment. This method is intended for use with a classical force field, in which the charges are usually scaled according to various criteria and differ from the Mulliken charges used in the previous sections.

To test the dependence of the results on the definition of the charges used to represent the electrostatic influence of the environment, we reproduced the previous computations on subsystem S_1 by replacing the Mulliken AM1 charges with the charges used in standard force fields: the CVFF force field which is in the Discover package[18] and the Amber force field.[20] The results are shown in Figure 2 and Table II. The general features of the energy variations are reproduced and the minima, as well as the maximum, appear at the same values of the bond lengths. However, the energy variations are noticeably underestimated, especially with the Amber charges. The curve obtained with the CVFF charges is closer to the reference, but one

TABLE II.
Main Geometric Features and Energies of the Charge Transfer Process Computed with Subset S_1 and S_2 and Various Sets of Charges (Reference Results: Whole Molecule Computed with AM1) (Bond Lengths in Å).

			Neutral Form		Intermediate State		Zwitterionic Form	
			Bond Length	Absolute Energy (au)	Bond Length	ΔE in kcal · mol^{-1}	Bond Length	ΔE in kcal · mol^{-1}
	Reference results	R_{OH}	0.973	(0)	1.454	19.005	2.201	13.950
		R_{NH}	2.523		1.143		1.013	
S_1	Zero charges	R_{OH}	0.976	−120.885631	1.454	18.489	2.200	14.219
		R_{NH}	2.232		1.142		1.013	
	AM1 Mulliken charges	R_{OH}	0.975	−120.903887	1.454	18.939	2.200	13.738
		R_{NH}	2.270		1.143		1.010	
	CVFF charges	R_{OH}	0.975	−120.906660	1.454	18.527	2.200	13.274
		R_{NH}	2.232		1.142		1.013	
	Amber charges	R_{OH}	0.973	−120.890534	1.454	16.499	2.199	12.416
		R_{NH}	2.569		1.154		1.013	
S_2	Zero charges	R_{OH}	0.976	−85.951327	1.424	16.601	1.808	14.355
		R_{NH}	2.282		1.137		1.040	
	AM1 Mulliken charges	R_{OH}	0.974	−85.958737	1.454	21.096	2.163	17.989
		R_{NH}	2.381		1.138		1.021	
	CVFF charges	R_{OH}	0.973	−85.952808	1.454	18.258	2.149	15.048
		R_{NH}	2.526		1.145		1.022	
	Amber charges	R_{OH}	0.974	−85.952508	1.424	16.774	2.147	13.951
		R_{NH}	2.392		1.158		1.025	

notices that the accident observed at R_{OH} = 1,750 Å is slightly displaced and strongly exaggerated. The length of the OH—N hydrogen bond in the neutral form is 2.569 Å for the Amber charges (close to the reference value) and 2.232 Å for the CVFF charges.

Due to the magnitude of the effects, an extensive study has been undertaken on subsystem S_2 with four different sets of point charges in E: a reference set with zero charges and the three previous sets (AM1 Mulliken charges, CVFF, and Amber). The results are listed in Table II. In the case of the neutral structure (which always corresponds to the lowest minimum chosen as origin of energies), we indicate, for the three sets considered here, the total energy which, by comparison with the reference set of zero charges, gives an idea of the magnitude of the electrostatic perturbation.

The results show that the magnitude of the perturbation may reach 5 kcal · mol^{-1} for the neutral form. The relative variations of energy from the neutral to the zwitterionic form are qualitatively comparable, although a difference of a few kcal · mol^{-1} may be very important for reactivity studies. One notices that the height of the energy barrier increases when the perturbation on the neutral form increases. The situation is slightly more complex for the zwitterionic form. One also notices that one finds two sets of values for the length of the OH—N hydrogen bond in the neutral form: one greater than 2.5 Å as in the reference results, the other one close to 2.3 Å as in the computation on the subset S_1 and the Mulliken AM1 charges. Obviously, this parameter is strongly dependent on the electrostatic interactions.

This study emphasizes the dependence of the

quantum chemical results on the representation of the electrostatic interaction. They are established by reference to the electrostatic potential created by the AM1 charge distribution, which is known not to be the best one.[21,22] Therefore, one cannot conclude anything about the quality of the various sets of charges. Let us only remark, at this early stage of the method, that regarding the electrostatic perturbation in the LSCF method, the Mulliken charges seem to overestimate this perturbation and that the CVFF charges seem rather consistent with the AM1 results obtained on the whole molecule. These remarks provide us with another reason for choosing a subsystem in which the atoms which strongly influence the reactive part of the system are described quantum mechanically.

Example of Application to a Protein

Due to the foregoing remarks, this first application of the LSCF method to a biochemical problem does not pretend to yield quantitative conclusions. In accordance with the test study, the reaction selected to illustrate the application of the method to a large system is a very simple one and is, again, an intramolecular proton transfer process. In view of testing the method on a real protein and analyzing the importance of the perturbation introduced by the electrostatic perturbation due to the presence of the whole protein, we chose to study the equilibria which are involved in the formation of the Asp^{69}—Arg^{71} salt bridge in dihydrofolate reductase. The crystallographic structure of the protein[23] is available in the Protein Data Bank (PDB).[24] In the crystal structure, the protein exists as a dimer, but in this study we only considered the monomer containing the salt bridge because the distance between this region and the other moiety of the protein is quite large. The system studied is made of 159 residues and is crystallized with many water molecules. To reduce slightly the number of degrees of freedom of the system, we only considered the four water molecules close to the salt bridge (numbered 733, 770, 783, and 784 in the PDB file).

The quantum *S* fragment is made of the sidechains or Asp^{69}, Asp^{70}, and Arg^{71}; of the two peptidic chains between them; and of the four water molecules. As in the first example considered above, the two frontier atoms are the C_α of Asp^{69} and Arg^{71}.

For this study, as in the case of the tetrapeptide test molecule, the whole system is first optimized with the CVFF force field. The scope of this work being to assess the performances of the LSCF method, the backbone of the fragment has not been reoptimized during the quantum chemical computation (to reduce the computer time).

The optimized geometries of the sidechains of Asp^{69}, Asp^{70}, and Arg^{71} in both the zwitterionic and the neutral forms are represented in Figures 3 and 4. The zwitterionic form is found, as expected, as the most stable one, but the geometry is quite different from the crystallographic one. In particular, the central Asp^{70} residue, which is not involved in the salt bridge, has moved from a geometry in which it was bonded to the backbone by a water molecule in a bridge position between one oxygen atom of this aspartate and that of the closest carbonyl of the backbone, to a rather free conformation close to what one would expect in aqueous solution. The solvatation by two water molecules is nearly optimal[25] and probably simulates what would happen in solution. The explanation for a very different structure in the crystal must be found in the intermolecular forces which favor the most compact configurations. The two other water molecules solvate the guanidinium group; the third residue, Asp^{69}, involved in the salt bridge is not solvated. The distance between the H atom of the guanidinium group and the oxygen atom of this aspartate is fairly short: 2.090 Å.

The corresponding neutral form appears as the less stable one: 70 kcal · mol^{-1} above the previous one. This difference is probably exaggerated and partly due to the fact that the backbone has not been reoptimized. The geometry is quite different from the zwitterionic form. In particular, the water molecules undergo an important displacement, but two of them are still solvating the Asp^{70} sidechain and the two others are solvating the guanidine group. In addition, one of them (labeled 733 in the PDB) bridges this group with the oxygen atom of Gly^{51}. The length of the hydrogen bond after this proton transfer in the salt bridge is now rather long: 2.610 Å.

This important stabilization of the zwitterionic form contrasts with the results obtained on the tetrapeptide. This difference can be attributed to the presence of the solvating water molecules, but the interaction with the whole protein probably plays a role too. To evaluate the importance of these effects, we first removed the interaction with the protein. The energy difference is still in favor of the zwitterionic form, but the difference is only 5.6

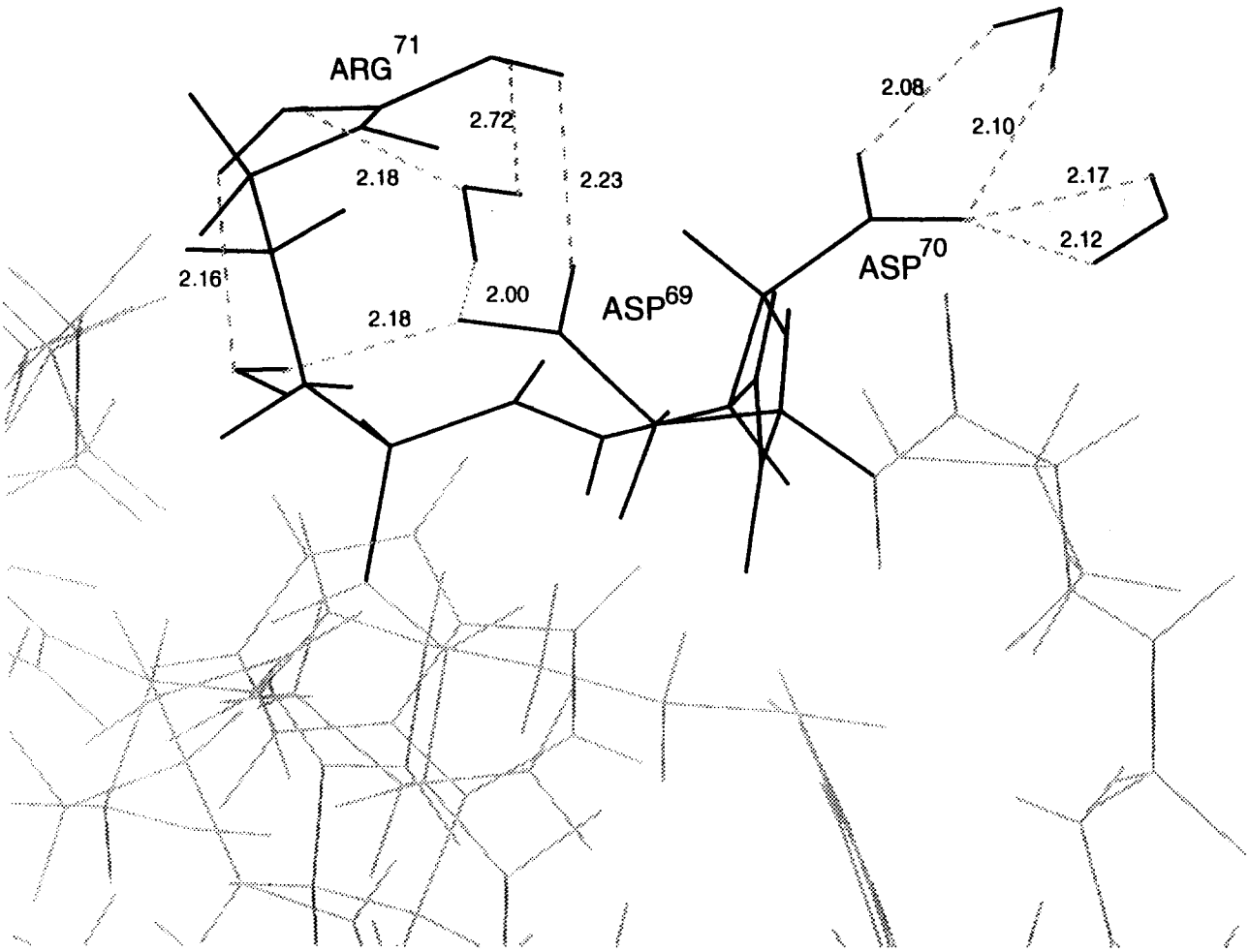

FIGURE 3. Zwitterionic form of the subsystem S_1 in the protein.

kcal · mol^{-1}. Finally, we removed the water molecules from the previous case and the energy difference became in favor of the neutral form for 5.0 kcal · mol^{-1}. The results then became consistent with the previous findings.

This preliminary study shows the importance of the electrostatic interactions between the fragment and the rest of the macromolecule in the energetics of a chemical process. The interactions are particularly important for the atoms which are very close to the quantum fragment (S). This is why it is so important to treat this boundary carefully, and it explains the difficulties which arise when extra hydrogen atoms are added to make the fragment computable with standard quantum chemical methods. The LSCF method, which avoids the use of these additional atoms, seems particularly well adapted to this important purpose.

Conclusion

This article is mainly addressed to the presentation of the LSCF method. The two systems treated have been chosen to show that it is now possible to perform a quantum chemical study of a fragment of a large system without worrying about the discontinuity between this fragment and the rest of the molecule. Many reactivity problems in large biomolecules can be tackled immediately with the help of this method, and studies of reactivity modification of a protein under the influence of mutations are in progress.

The study of the influence of conformational fluctuations or changes on enzyme reactivity is already feasible by means of the LSCF method cou-

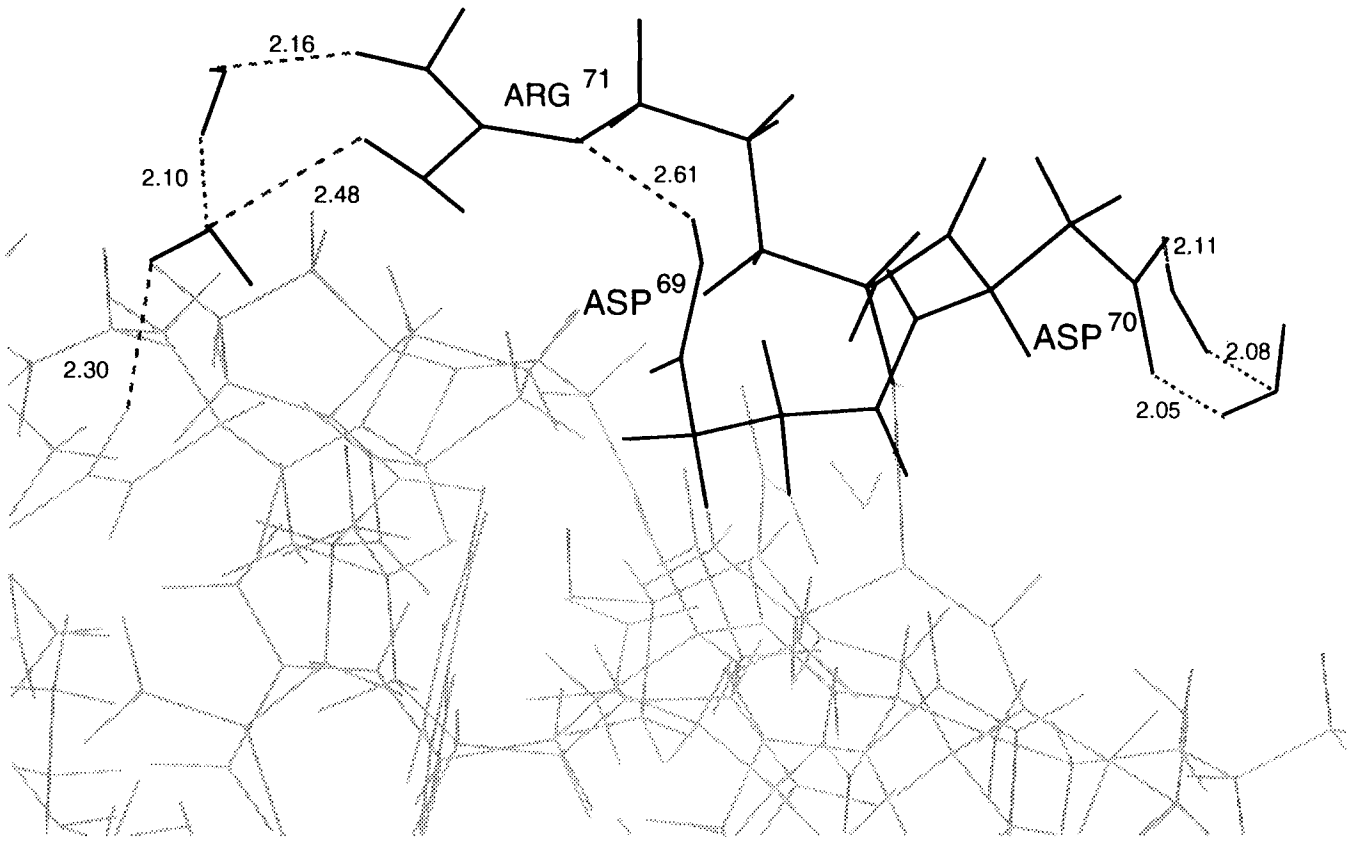

FIGURE 4. Neutral form of the subsystem S_1 in the protein.

pled with classical molecular dynamics as well as the case in which the subsystem is made of small molecules.[5] There is no doubt that such studies will become very important in the near future, since the computational aspect is no longer a problem and the possibility of entering the fascinating field of biochemical reactions[25,26] is a very strong motivation to use and improve a methodology intended for this purpose.

Among the probable improvements which we may take the risk to foresee, a new generation of force fields is expected.[27] One may consider that the current force fields have been unique tools to predict the energy variations induced by conformational changes and noncovalent interactions. These force fields, however, work at the level of atoms and bonds, and we have experienced in this study how much a modeling scheme which takes into account the behavior of electrons in the molecule may be demanding. A correct representation of the molecular electrostatic potential by a set of charges is still a matter of research.[21,22,28–36] It may need more than one point charge per atom[37] or a distribution of multipoles.[38–41] Any of these solutions can fit the LSCF methodology, and the corresponding algorithms have already been implemented in GEOMOS for various other purposes.

One of the most obvious differences between a quantum chemical computation and a molecular mechanical study based on a force field is that the former takes into account the modifications of the electron distribution under the influence of the interactions which give rise to the well-known induction term in the interaction potential.[42] This phenomenon can be modeled by introducing local[43] or distributed polarizabilities[41] and parameterizing the force field accordingly. This is not very common yet,[44,45] although it is possible to show that introducing induction effects ensures the conformational invariance of the charges which represent the electrostatic term,[46] as assumed previously.[27]

Again, the LSCF methodology can be adapted easily to incorporate the perturbation by distrib-

uted induced dipole or multipoles, or any other transferable model of electronic polarizability. More generally, one notices that the main feature of the LSCF method is to treat a fragment exactly like a molecule in a standard SCF computation; and the rest of the macromolecule can be compared, to a large extent, to a solvent. Therefore, all the models which have been devised to simulate the influence of a macroscopic environment on a molecular system[6,47-49] can be adapted in a very straightforward way to this methodology.

Another probable improvement deals with the LSCF method itself. In its present form, it only works with semiempirical methods since its basic assumptions are consistent with the simplifications introduced in these methods. One might feel frustrated to be obliged to resort to the semiempirical level when *ab initio* techniques are considered as the reference for modeling electronic properties. It is clear that the method can be extended to the *ab initio* level provided that some appropriate approximations are introduced. These approximations must be studied carefully to prevent spurious errors, which would make this sophisticated approach worse than a semiempirical one.

These remarks suggest that the LSCF method has a promising future in the areas of methodological improvements and applications. In its present form, it is perfectly well adapted to the study of biochemical reactions by means of molecular dynamics, at least at a qualitative level. We are also evaluating its use in chemical reactivity studies at the surface of covalent solids by considering a representative subsystem of the solid and simulating the rest of the structure with the proper charge distribution.

References

1. E. Haaksma, H. Timmerman, and H. Weinstein, *Isr. J. Chem.*, **31,** 409 (1991).
2. A. Warshel, *Curr. Opin. in Struct. Biol.*, **2,** 230 (1992).
3. A. Warshel, *Computer Modeling of Chemical Reactions in Enzymes and Solutions*, John Wiley & Sons, New York, 1991, p. 1.
4. A. Warshel and J. Åqvist, *Ann. Rev. Biophys. Biophys. Chem.*, **20,** 267 (1991).
5. M. J. Field, P. A. Bash, and M. Karplus, *J. Comp. Chem.*, **11,** 700 (1990).
6. A. Warshel and M. Levitt, *J. Mol. Biol.*, **103,** 227 (1976).
7. F. Bernardi, M. Olivucci, and M. A. Robb, *J. Am. Chem. Soc.*, **114,** 1606 (1992).
8. G. Naray-Szabo, *Acta Phys. Acad. Sci. Hung*, **40,** 261 (1976).
9. G. Naray-Szabo and P. R. Surjan, *Chem. Phys. Lett.*, **96,** 499 (1983).
10. G. Naray-Szabo, *Croat. Chim. Acta*, **57,** 901 (1984).
11. G. G. Ferenczy, J. L. Rivail, P. R. Surjan, and G. Naray-Szabo, *J. Comp. Chem.*, **13,** 830 (1992).
12. J. A. Pople, D. P. Santry, and G. A. Segal, *J. Chem. Phys.*, **43,** 125 (1965).
13. J. L. Rivail, M. Loos, and V. Théry, *Trends in Ecological Physical Chemistry*, L. Bonati et al., Eds., Elsevier, Amsterdam, 1993, p. 17.
14. M. J. S. Dewar and W. Thiele, *J. Am. Chem. Soc.*, **99,** 4899 (1977).
15. M. J. S. Dewar, E. G. Zoebish, E. F. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.*, **107,** 3902 (1985).
16. J. J. P. Stewart, *J. Comp. Chem.*, **10,** 209, 221 (1989).
17. D. Rinaldi, P. E. Hoggan, and A. Cartier, *GEOMOS*, QCPE 584.
18. Biosym Technologies, 9685 Scranton Road, San Diego, CA 92121-2777.
19. C. Edminston and K. Ruedenberg, *Rev. Mod. Phys.*, **35,** 457 (1963).
20. S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, *J. Comp. Chem.*, **7,** 230 (1986).
21. B. H. Besler, K. M. Merz, Jr., and P. A. Kollman, *J. Comp. Chem.*, **11,** 431 (1990).
22. K. M. Merz, Jr., *J. Comp. Chem.*, **12,** 749 (1992).
23. J. T. Bolin, D. J. Filman, D. A. Mattews, R. C. Hamlin, and J. Krant, *J. Biol. Chem.*, **257,** 13650 (1982).
24. J. R. Rodgers, O. Kennard, T. Schimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112,** 535 (1977).
25. N. Thanki, J. M. Thornton, and J. M. Goodfellow, *J. Mol. Biol.*, **202,** 637 (1988).
26. P. A. Kollman, *Curr. Opin. in Struc. Biol.*, **2,** 765 (1992).
27. K. M. Merz, Jr., *Curr. Opin. in Struc. Biol.*, **3,** 234 (1993).
28. L. E. Chirlian and M. M. Francl, *J. Comp. Chem.*, **8,** 894 (1987).
29. G. G. Ferenczy, C. A. Reynolds, and W. G. Richards, *J. Comp. Chem.*, **11,** 159 (1990).
30. F. Colonna, J. G. Ángyán, and O. Tapia, *Chem. Phys. Lett.*, **55,** 172 (1990).
31. G. G. Ferenczy, *J. Comp. Chem.*, **12,** 913 (1991).
32. M. Aida, G. Corongiu, and E. Clementi, *Int. J. Quant. Chem.*, **42,** 1353 (1992).
33. F. Colonna, E. Evleth, and J. G. Ángyán, *J. Comp. Chem.*, **13,** 1234 (1992).
34. C. Chipot, B. Maigret, J. L. Rivail, and H. A. Scheraga, *J. Phys. Chem.*, **96,** 10276 (1992).
35. C. Chipot, J. G. Ángyán, G. G. Ferenczy, and H. A. Scheraga, *J. Phys. Chem.*, **97,** 6628 (1993).
36. C. Chipot, J. G. Ángyán, B. Maigret, and H. A. Scheraga, *J. Phys. Chem.*, **97,** 9788, 9797 (1993).
37. G. Rauhut and T. Clark, *J. Comp. Chem.*, **14,** 503 (1993).
38. F. Vigné-Maeder and P. Claverie, *J. Chem. Phys.*, **88,** 4934 (1988).
39. P. Claverie, In *Intermolecular Interactions from Diatomics to*

Biopolymers., B. Pullman, Ed., John Wiley & Sons, Chichester, 1978, p. 69.

40. A. J. Stone, *Chem. Phys. Lett.*, **83,** 233 (1981).
41. A. J. Stone, *Mol. Phys.*, **56,** 1065 (1985).
42. A. D. Buckingham, In *Intermolecular Interactions from Diatomics to Biopolymers*. B. Pullman, Ed., John Wiley & Sons, Chichester, 1978, p. 1.
43. C. Voisin and A. Cartier, *J. Mol. Struc. (THEOCHEM)*, **286,** 35 (1993).
44. D. A. Pearlman, D. A. Case, J. C. Cadwell, G. L. Seibel, U. C. Singh, P. Weiner, and P. A. Kollman, *AMBER 4.0*, University of California, San Francisco, 1991.
45. N. Gresh, P. Claverie, and A. Pullman, *SIBFA*, QCPE 614.
46. F. Colonna and E. M. Evleth, *Chem. Phys. Lett.*, **212,** 665 (1993).
47. O. Tapia and G. Johannin, *J. Chem. Phys.*, **75,** 3624 (1981).
48. B. T. Thole and P. T. van Duijnen, *Biophys. Chem.*, **18,** 53 (1983).
49. J. G. Ángyán, M. Allavena, M. Picard, A. Potier, and O. Tapia, *J. Chem. Phys.*, **77,** 4723 (1982).