Journal of Computational Physics **151,** 242–263 (1999)
Article ID jcph.1999.6219, available online at http://www.idealibrary.com on IDEAL®

A Linear-Scaling Quantum Mechanical Investigation of Cytidine Deaminase

James P. Lewis,*,[1] Shubin Liu,† Tai-Sung Lee,‡,[2] and Weitao Yang‡

**CB7260, Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599-7260, and Department of Chemistry, Duke University, Durham, North Carolina 27708-0348; †CB7260, Department of Biochemistry and Biophysics, School of Medicine, and Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-7260; and ‡Department of Chemistry, Duke University, Durham, North Carolina 27708-0348*

Received August 21, 1998; revised February 2, 1999

We describe the divide-and-conquer technique for linear-scaling semiempirical quantum mechanical calculations. This method has been successfully applied to study cytidine deaminase. Large-scale simulations were performed for optimizing geometries surrounding the active site of the enzyme and obtaining related energetics. The results of the minimizations provide a significant complement to experimental efforts and aid in the understanding of the enzymatic profile of cytidine deaminase. More specifically, we present our predictions about the structure of the active species and the structure of the active site for low pH. Finally, we present our results for the structure of the zinc ion coordination for different substrates which represent points along the reaction profile. In particular, we find that our results for the Zn-S_γ132 and the Zn-S_γ129 bondlengths yield similar trends compared to x-ray crystallography data as the enzyme structure changes from the ground-state to the transition-state analog and from the transition-state analog to the product. © 1999 Academic Press

I. INTRODUCTION

Considerable effort has been devoted to study the catalytic mechanisms of the nucleoside deaminases cytidine deaminase (CDA) and adenosine deaminase [1–13]. The enzyme CDA is an efficient catalyst which accelerates the rate of hydrolytic deamination of cytidine to uridine. Understanding the catalytic mechanisms for CDA could have important potential therapeutic uses [14–16]. Available x-ray crystallographic studies include several crystal structures of stable intermediate CDA states achieved using different analog complexes to

[1] Present address: Department of Chemistry, University of Utah, Salt Lake City, UT 84112-0850.

[2] Present address: Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0446.

represent the ground-state [12], pre-transition-state [10], transition-state [8], and product [13]. The experimental data from these studies have begun to suggest important subtleties in the configuration of a catalyzed active zinc ion and in conformational differences between complexes. The C^4 position on the pyrimidine ring moves about 1.5 Å closer to the Zn-activated nucleophile while the amino group moves away from the nucleophile, with the N^4–C^4 distance increasing from 1.33 Å in the ground-state to approximately 2.80 Å in the product-state [13]. The enzyme active site apparently changes structure to accommodate each successive ligand in the sequence of structures.

However, many of these subtleties lie at or beyond the limit normally considered to be the noise level of an x-ray diffraction experiment. To verify their significance, higher resolution must be achieved or some theoretical tool must be employed to complement the experimental procedures. A quantum mechanical (QM) description of the sequence of structures formed during catalysis has long been a goal to help mechanistic enzymologists understand the reaction pathway. Until recently, progress toward this goal has been frustrated by inadequate computational methods and by the lack of a good experimental system with which to calibrate the calculations. The problem is that any theoretical technique employed needs to give a valid description for much of the protein environment surrounding the active site.

Quantum mechanical methods for determining the electronic structure of atoms and molecules are crucial for a reliable description of complex chemical processes that are inaccessible to conventional empirical models, e.g., bond formation and cleavage in chemical reactions, polarization, and chemical bonding of metal ions. However, standard QM techniques are limited to fairly small molecular systems because of a variety of theoretical and technical difficulties. One of the several limitations of applying QM methods to large systems is the high order scaling properties. For example, in Hartree–Fock (HF) type calculations, the computational requirement, including CPU time and memory, scale as N^2 to N^4, where N is the number of electron orbitals in the system. Density functional theory (DFT) offers more computational efficiency, without loss of accuracy, but has similar scaling problems due to the N^3 diagonalization of the Hamiltonian matrix [17, 18].

Several linear-scaling methods have been proposed to circumvent this N^3 diagonalization constraint [19–37]. In all the methods, the localization of the electronic degrees of freedom is the key for achieving linear-scaling in QM calculations. One can efficiently calculate the electronic structure of a large system with just the local variables, localized molecular orbitals, or the density matrix. There is also the scaling issue of numerical integration and matrix construction in DFT (and HF) calculations. As the focus here is on our use of semiempirical approaches, the reader is referred to a recent review article on the linear-scaling approach for the first-principle methods [38].

The divide-and-conquer (DAC) approach proposed by Yang was the first linear-scaling method used to carry out QM calculations [19, 39]. The basic strategy of this method is as follows: divide a large system into many subsystems, determine the electron density of each subsystem separately, and sum the corresponding contributions from all subsystems to obtain the total electron density and the energy of the system. This approach is based on the fact that the electron density is a local property and, in DFT, the ground-state of a system can be obtained solely from the electron density. A density-matrix reformulation of the method by Yang and Lee further simplifies the algorithm [20]. Merz *et al.* and St-Amant *et al.* have successfully applied and extended the DAC approach to several systems [40–43].

The organization of this paper is as follows. In the next section (Section II), we review the DAC approach as it has been implemented into a semiempirical framework.

After discussing this implementation, we will discuss the application of the DAC approach within this semiempirical framework to the cytidine deaminase enzyme system in Section III.

II. THEORY

A. *Semiempirical Quantum Chemical Theory*

In semiempirical methods, the matrix elements of the Hamiltonian matrix are efficiently calculated because of the judicious approximations made. The Hamiltonian matrix constructed from semiempirical (or tight-binding like) approximations is usually very sparse and the computational effort in practice is limited only by the cubic-scaling diagonalization processes. The sparseness of the Hamiltonian matrix implies that semiempirical QM methods [44–46] are naturally conducive for the application of linear-scaling algorithms to "diagonalize" the Hamiltonian matrix. The combination of an efficient determination for the matrix elements and a linear-scaling algorithm for "diagonalization" have allowed feasible QM calculations for very large systems. The density matrix version of the DAC approach by Yang and Lee [20] accommodates a density matrix description so that it can be applied to HF and semiempirical methods. It has been implemented into semiempirical methods by Lee *et al*. [47] and Dixon and Merz [40, 41]. Stewart has implemented a pseudodiagonalization method to obtain localized molecular orbitals [48]. Daniels *et al*. have used a conjugate gradient density matrix search to replace diagonalization in semiempirical calculations [49]. These methods all have shown the feasibility of simulating molecules with several thousand atoms within the semiempirical QM framework.

In the HF method, the system wavefunction is obtained from the eigenvalue equation

$$(\mathbf{F} - \varepsilon_m \mathbf{S})\mathbf{C}_m = 0, \tag{1}$$

where **F** is the Fock matrix, **S** is the basis overlap matrix, and $\{\mathbf{C}_m\}$ and $\{\varepsilon_m\}$ are the molecular orbitals and the corresponding eigenenergies of those orbitals. In terms of the density matrix **P**, the Fock matrix can be rewritten as

$$F_{ij} = H_{ij} + \sum_{kl} \left\{ P_{kl} \langle ij \mid kl \rangle - \frac{1}{2} P_{kl} \langle ik \mid jl \rangle \right\}, \tag{2}$$

where **H** is the one-electron core Hamiltonian matrix and $ijkl$ are the indices of basis functions. The four-center two-electron repulsion integral, $\langle ij \mid kl \rangle$, is defined by

$$\langle ij \mid kl \rangle = \int d\mathbf{r}_1 \, d\mathbf{r}_2 \phi_i(\mathbf{r}_1) \phi_j(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \phi_k(\mathbf{r}_2) \phi_l(\mathbf{r}_2), \tag{3}$$

where $\{\phi_i\}$ are the basis functions. The density matrix is obtained from the molecular orbitals $\{C_m\}$ of Eq. (1) by the following summation over the occupied orbitals

$$P_{ij} = 2 \sum_m^{occ} C_{im} C_{jm}. \tag{4}$$

Finally, the total electronic energy of the system is

$$E = \frac{1}{2}\sum_{ij} P_{ij}(H_{ij} + F_{ij}). \tag{5}$$

To achieve self-consistent results, Eqs. (1)–(5) must be solved iteratively.

One of the main practical difficulties for HF methods is the high-order scaling property of calculating the Fock matrix and the density matrix. Obtaining the Fock matrix could scale as N^4 because of the evaluation of four-center integrals. Furthermore, to get accurate results, one should use a better basis set which leads to a very time-consuming evaluation of integrals. Because of these difficulties, the application of HF methods is limited for application to large systems.

Semiempirical calculations are set up with the same general structure as a HF calculation, but the required integrals are approximated or completely omitted. The approximated integrals are calculated from a set of parameters which are fit to give the best possible agreement with experimental data, as such the evaluation of integrals becomes much faster and larger systems can be calculated. We only focus on the MNDO type semiempirical methods, since not only is this type used to generate the results presented in Section III, but also these methods are the most widely used and give acceptable results in most cases [50]. The approximations in the MNDO type semiempirical calculations, which include MNDO [50], AM1 [51] and PM3 [52, 53], are described as follows.

(1) The overlap matrix in Eq. (1) is set to the identity matrix and Eq. (1) reduces to

$$(\mathbf{F} - \varepsilon_m)\mathbf{C}_m = 0. \tag{6}$$

(2) All three- and four-center two-electron integrals are set to zero, i.e., $\langle ij \mid kl\rangle = 0$, except for integrals with ij on the same atom and kl on the same atom which will be calculated.

(3) The two-center/two-electron integrals are approximated by the interactions of classical multipoles. The basis pair ij is treated as an electron density distribution and approximated by point charges, such that the two-center integrals are represented by interactions between two sets of multipole distributions. Of course, this approximation fails when the distance between the two centers approaches zero, since consequently the integrals will approach infinity. In such a case, the two-center integrals should reduce to the corresponding one-center integrals, but they do not. To correct this problem, a modified formula is used for the Coulomb interactions [54]. For example, in the integral where $\langle ij|$ reduces to two s-states on atom A and $|kl\rangle$ reduces to two s-states on atom B, i.e., $\langle s_A s_A \mid s_B s_B\rangle$, the factor $1/R$ is set equal to $1/\sqrt{(R + C_A + C_B)^2 + (G_A + G_B)^2}$, where C_A, C_B, G_A, and G_B are derived parameters so that the two-center integrals reduce to the corresponding one-center integrals at $R = 0$.

(4) The one-center two-electron integrals are fit to experimental data.

(5) The off-diagonal elements of the Hamiltonian matrix, **H**, are approximated by

$$H_{ij} = \frac{S_{ij}}{2}(\beta_i + \beta_j), \tag{7}$$

where S_{ij} is an overlap matrix element, assuming that i and j are Slater orbitals, and the $\{\beta_i\}$ are parameters. The overlap matrix here is used only for calculation of the **H** matrix and it is not the equivalent overlap matrix as used in approximation (1). The diagonal elements of **H** are attributed to two types of contributions. For an atomic orbital i on atom A, the contribution from atom A is fit, but the contribution from other nuclei, namely the electron–nuclear interaction, is approximated by

$$H_{ii}^{e,n} = -\sum_{B} Z_B \sum_{kl} \langle ii \mid kl \rangle, \tag{8}$$

where Z_B is the nuclear charge of atom B and $\langle ii \mid kl \rangle$ is the two-center/two-electron integral. The indices k and l signify that the atomic orbitals are centered on atom B, any atom different from atom A.

(6) In simple semiempirical methods, the nuclear–nuclear interaction between atom A and atom B is modified to become the core–core repulsion interaction as

$$E_N(A, B) = Z_A Z_B \langle s_A s_A \mid s_B s_B \rangle. \tag{9}$$

However, in the MNDO method, the nuclear–nuclear interaction is modified such that

$$E_N^{\mathrm{MNDO}}(A, B) = Z_A Z_B \langle s_A s_A \mid s_B s_B \rangle \left(1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}\right) \tag{10}$$

or for O–H or N–H pairs,

$$E_N^{\mathrm{MNDO}}(A, B) = Z_A Z_B \langle s_A s_A \mid s_B s_B \rangle \left(1 + R_{AB} e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}\right), \tag{11}$$

where α_A and α_B are parameters. In the AM1 and PM3 methods, extra terms are added, yielding

$$E_N(A, B) = E_N^{\mathrm{MNDO}}(A, B) + \frac{Z_A Z_B}{R_{AB}} \left(\sum_k a_{kA} e^{-b_{kA}(R_{AB} - c_{kA})^2} + \sum_k a_{kA} e^{-b_{kA}(R_{AB} - c_{kA})^2} \right), \tag{12}$$

where a, b, and c are all parameters.

Considering the above approximations, semiempirical methods are much faster than HF methods in evaluating the the matrix elements needed for constructing the Hamiltonian matrix. After this construction, the Hamiltonian matrix must be diagonalized to determine the band structure energy contribution of the total energy. In the next sections (Subsections II.B–II.D) we discuss the DAC approach to achieve linear-scaling "diagonalization" and its implementation to semiempirical QM methods.

B. *The Divide-and-Conquer Approximation*

In this section, we describe the original DAC method which is only applicable to DFT methods [19, 39]. In the next section (Subsection II.C), we will describe the density matrix version of the DAC method which can be applied to both HF and semiempirical methods.

The electron density is the fundamental variable in the DAC method and can be represented as the sum of contributions from subsystems. This is made possible through normalization of partition functions,

$$\sum p^{\alpha}(\mathbf{r}) = 1, \tag{13}$$

where $p^{\alpha}(\mathbf{r})$ is the partition function for the subsystem α. The total density is then expressed as a sum of subsystem density,

$$\rho(\mathbf{r}) = \sum_{\alpha} p^{\alpha}(\mathbf{r})\rho(\mathbf{r}) = \sum_{\alpha} \rho^{\alpha}(\mathbf{r}), \tag{14}$$

where $\rho^{\alpha}(\mathbf{r}) \equiv p^{\alpha}(\mathbf{r})\rho(\mathbf{r})$. By definition, each subsystem density has the proper contribution to the total density. Prescriptions have been defined for the partition functions [19, 39]. The resulting density and energy do not depend on the particular form of the partition function in any significant way.

A subsystem density defined in Eq. (14) is localized in only a small region of the physical space and therefore can be obtained efficiently with an approximation that depends on the local physical space,

$$\rho^{\alpha}(\mathbf{r}) = 2p^{\alpha}(\mathbf{r})\sum_{m} f_{\beta}\left(\mu - \varepsilon_{m}^{\alpha}\right)\left|\psi_{m}^{\alpha}(\mathbf{r})\right|^{2}, \tag{15}$$

where $f_{\beta}(x)$ is the Fermi function ($f_{\beta}(x) = [1 + \exp(-\beta x)]^{-1}$) with inverse temperature β, and $\psi_{m}^{\alpha}(\mathbf{r})$ and ε_{m}^{α} are local eigenfunctions and eigenvalues of the subsystem. The factor of 2 in Eq. (15) is for double occupancy in closed-shell systems. For each subsystem, the local eigenfunctions are given by the linear combinations of the local basis functions $\{\phi_{i}^{\alpha}\}$,

$$\psi_{m}^{\alpha}(\mathbf{r}) = \sum_{i} C_{im}^{\alpha}\phi_{i}^{\alpha}(\mathbf{r}), \tag{16}$$

where the linear coefficients are the solutions of the following generalized eigenvalue equation derived from the Rayleigh–Ritz variational principle,

$$\left(\mathbf{H}^{\alpha} - \varepsilon_{m}^{\alpha}\mathbf{S}^{\alpha}\right)\mathbf{C}_{m}^{\alpha} = 0. \tag{17}$$

The Hamiltonian matrix and the overlap matrix elements are given by $(\mathbf{H}^{\alpha})_{ij} = \langle\phi_{i}^{\alpha}|H|\phi_{j}^{\alpha}\rangle$ and $(\mathbf{S}^{\alpha})_{ij} = \langle\phi_{i}^{\alpha}|\phi_{j}^{\alpha}\rangle$, where the Kohn–Sham Hamiltonian H depends on the density. The chemical potential μ is set by the electron density normalization condition,

$$N = \int \rho(\mathbf{r})\,d\mathbf{r}^{3} = 2\sum_{\alpha}\sum_{m} f_{\beta}\left(\mu - \varepsilon_{m}^{\alpha}\right)\left\langle\psi_{m}^{\alpha}(\mathbf{r})|p^{\alpha}(\mathbf{r})|\psi_{m}^{\alpha}(\mathbf{r})\right\rangle. \tag{18}$$

Equations (15)–(18) need to be solved self-consistently, just as in the Kohn–Sham method. Finally, the total electronic energy can be calculated from $E = \varepsilon + Q[\rho]$, where ε is an approximation to the sum of the Kohn–Sham eigenvalues,

$$\varepsilon = 2\sum_{\alpha}\sum_{m}\varepsilon_{m}^{\alpha} f_{\beta}\left(\mu - \varepsilon_{m}^{\alpha}\right)\left\langle\psi_{m}^{\alpha}(\mathbf{r})|p^{\alpha}(\mathbf{r})|\psi_{m}^{\alpha}(\mathbf{r})\right\rangle, \tag{19}$$

The quantity $Q[\rho] = \int \rho[-\phi(\mathbf{r})/2 - V_{xc}(\mathbf{r})]\,d\mathbf{r} + E_{xc}[\rho]$, where $\phi(\mathbf{r})$ is the electrostatic potential due to the electrons, $V_{xc}(\mathbf{r})$ is the exchange-correlation potential, and $E_{xc}[\rho]$ is the exchange-correlation energy, can be determined from the density alone [19, 39].

The use of a set of basis functions localized in the relevant part of the space is what makes the DAC method have linear-scaling with system size. The partition functions of Eqs. (13)–(14) define the physical space division of a molecule. This concludes our discussion of the general DAC method. In the next section (Subsection II.C), we discuss the density matrix version of the DAC method.

C. *Density Matrix Version of Divide-and-Conquer Method*

The fundamental component of the DAC method is the density matrix. The Kohn–Sham one-electron density matrix is defined in terms of the Kohn–Sham orbitals $\{\psi_m(\mathbf{r})\}$ as

$$\rho(\mathbf{r},\mathbf{r}') = 2\sum_m^{N/2} \psi_m(\mathbf{r})\psi_m(\mathbf{r}') = \sum_{ij} \rho_{ij}\phi_i(\mathbf{r})\phi_j(\mathbf{r}'), \tag{20}$$

where the density matrix in the atomic orbital space is given by the linear coefficients in the expansion of the Kohn–Sham orbitals; namely,

$$\rho_{ij} = 2\sum_m^{N/2} C_{im}C_{jm}. \tag{21}$$

We can define a partition matrix P_{ij}^{α} for subsystem α in the space of atomic orbitals. Corresponding to Eq. (13), we implement the normalization condition

$$\sum_{\alpha} P_{ij}^{\alpha} = 1. \tag{22}$$

There is a simple way to construct such matrices; namely,

$$\begin{aligned} P_{ij}^{\alpha} &= 1 && \text{if } i \in \alpha \text{ and } j \in \alpha \\ &= 1/2 && \text{if } i \in \alpha \text{ and } j \notin \alpha \\ &= 0 && \text{if } i \notin \alpha \text{ and } j \notin \alpha. \end{aligned} \tag{23}$$

The density matrix can then be divided into subsystem contributions as

$$\rho_{ij} = \sum_{\alpha} P_{ij}^{\alpha}\rho_{ij} \equiv \sum_{\alpha} \rho_{ij}^{\alpha}, \tag{24}$$

which parallels Eq. (14).

We now make the approximation for each subsystem,

$$\rho_{ij}^{\alpha} = 2P_{ij}^{\alpha}\sum_m f_{\beta}\left(\mu - \varepsilon_m^{\alpha}\right)C_{im}^{\alpha}C_{jm}^{\alpha}. \tag{25}$$

This approximation corresponds to the one made for Eq. (15) in the original density approach. It uses a set of local eigenvectors to approximate the density matrix of a subsystem. This is the crux of linear-scaling in the computational effort, because the set of local eigenvectors for a subsystem is finite and independent of the size of the whole system. The chemical potential μ is determined by the normalization,

$$N = 2\sum_{ij}\rho_{ij}S_{ij} = 2\sum_{ij}\left\{\sum_{\alpha}P_{ij}^{\alpha}\sum_m f_{\beta}\left(\mu - \varepsilon_m^{\alpha}\right)C_{im}^{\alpha}C_{jm}^{\alpha}\right\}S_{ij}, \tag{26}$$

and the sum of eigenvalues becomes

$$\begin{aligned} \varepsilon &= 2\sum_{\alpha}\sum_m \varepsilon_m^{\alpha} f_{\beta}\left(\mu - \varepsilon_m^{\alpha}\right)\sum_{ij}P_{ij}^{\alpha}C_{im}^{\alpha}C_{jm}^{\alpha}S_{ij} \\ &= 2\sum_{ij}\left\{\sum_{\alpha}P_{ij}^{\alpha}\sum_m f_{\beta}\left(\mu - \varepsilon_m^{\alpha}\right)C_{im}^{\alpha}C_{jm}^{\alpha}\right\}H_{ij}, \end{aligned} \tag{27}$$

corresponding to Eq. (19). The second equality in Eq. (27) follows from the eigenequation Eq. (18) and the special construction of the partition matrix in Eq. (23) can also be written as $P_{ij}^{\alpha} = q_i^{\alpha} + q_j^{\alpha}$, where $q_i^{\alpha} = 1/2$ if $i \in \alpha$ and $q_i^{\alpha} = 0$ if $i \notin \alpha$.

The main difference between the density formulation and the general DAC method is that the division of the molecule in the former is accomplished in the space of the atomic orbitals, while the approximation of each subsystem by a set of local basis functions remains the same as described in Eqs. (16) and (17). When there is only one subsystem per atom, Eqs. (22)–(24) for the division of the density matrix correspond exactly to the Mulliken population analysis [55].

There are two advantages of this density matrix formulation. The first one is that we no longer need to calculate the integrals associated with the partition functions, that is, $\langle \phi_i^{\alpha}(\mathbf{r}) | p^{\alpha}(\mathbf{r}) | \phi_j^{\alpha}(\mathbf{r}) \rangle$. This makes the new formulation more efficient, as three-dimensional numerical integration is time consuming. The second advantage is that the density matrix formulation can be applied to other *ab initio* methods such as HF and semiempirical methods. A drawback in dividing the molecule in atomic orbital space is that the division becomes less localized in the physical space as we use more diffuse atomic functions—a well-known problem of the Mulliken population analysis. In contrast, the Hirshfeld-type partition of the density, Eqs. (13) and (14), is much less dependent on the basis functions.

The energy gradients with respect to the nuclear coordinates can be calculated by a "divide-and-conquer" approximation to the exact force expression, as has been done with the original density formulation [56]. However, we can have more efficient force calculations within the semiempirical approach. See the next section (Subsection II.D) where we discuss the implementation of the DAC method for semiempirical QM methods.

D. *The Divide-and-Conquer Implementation to Semiempirical Quantum Chemical Methods*

The density matrix DAC method has been implemented into the MOPAC semiempirical method [57]. In semiempirical calculations, the electronic energy is expressed by

$$E = \frac{1}{2}\sum_{ij} \rho_{ij}(H_{ij} + F_{ij}), \tag{28}$$

where $\mathbf{H}$ is the one-electron core Hamiltonian matrix, $\mathbf{F}$ is the Fock matrix, and ρ is the density matrix. In the DAC approach, and using Eq. (14), the electronic energy can be rewritten as

$$E = \frac{1}{2}\sum_{\alpha}\sum_{ij} \rho_{ij}^{\alpha}(H_{ij} + F_{ij}). \tag{29}$$

The energy gradient expressions for the DAC approach have been previously derived and shown to be accurate [56, 20]. In the MOPAC package, the energy gradients are calculated with the frozen density approximation. With this approximation, the DAC energy gradient with respect to the α nucleus position $\mathbf{R}_a$ is expressed by

$$\nabla_{\mathbf{R}_a} E = \frac{1}{2}\sum_{\alpha}\sum_{ij} \rho_{ij}^{\alpha} \nabla_{\mathbf{R}_a}(H_{ij} + F_{ij}). \tag{30}$$

Procedures for calculating the gradients are similar to those in MOPAC, except that the total density matrix is approximated by the DAC approach. The MOPAC package uses the BFGS optimization procedure for geometry optimization [57]. This procedure requires constructing the Hessian Matrix which has an $O(N^2)$ scaling requirement for memory usage; it cannot be used for large molecules. Instead, we choose a conjugate gradient method for geometry optimization [58].

The gradients can be calculated by analytical methods [59] or by the finite difference method. The finite difference method is faster than the analytical method, since the formulas to calculate matrix elements are much simpler than those to calculate the derivatives of matrix elements. Previous work showed that the finite difference method gives very close agreement to the analytic method, thus the finite difference method has been used to calculate gradients, as it is used in the MOPAC package [57]. However, the finite difference method scales at least quadratically and it would become the computational bottleneck when the system size becomes very large.

Several tests were performed to demonstrate the efficiency of the DAC implementation to the semiempirical quantum chemical methods [47]. The results of these tests demonstrate that the DAC implementation is computationally efficient and accurate. In all previously performed tests and for the results presented here, the following computational criteria are used.

A subsystem is defined as one amino acid residue for protein molecules and one nucleotide unit for DNA molecules. Instead of the entire set of atomic orbitals, each subsystem is described by a set of local basis functions, which enhances the accuracy of the description for neighboring atoms (*buffer atoms*). Buffer atoms are selected by a distance criterion, R_b; if an atom is within a distance R_b of a subsystem, this atom will be included as a buffer atom for that subsystem. The diagonalization for a subsystem is performed with atomic basis functions of the subsystem atoms and buffer atoms, and the computational effort scales as N_α^3, where N_α is the number of basis functions in the α subsystem and its buffer region. Studies using density functional theory have shown that the buffer region size needed for a given accuracy is independent of the size of the whole molecule [39, 60]. Hence, one can choose N_α as roughly a constant; for example, each subsystem consists of a single amino acid. The accuracy for different buffer sizes was previously determined elsewhere [47]. An accuracy criterion of 5×10^{-3} per atom was chosen for the energy calculations and 0.1 kcal/(mol Å) in gradient calculations. With this accuracy criterion, we found that the buffer size should have no less than a 6.0 Å cutoff ($R_b = 6.0$ Å).

While the DAC method overcomes the $O(N^3)$ scaling problem in the diagonalization process, the $O(N^2)$ scaling of memory storage must be addressed. Since most matrix elements in quantum calculations are negligibly small for large molecules, sparse matrix storage methods can be employed. Because of density matrix locality in real space, we choose to truncate the matrix elements using a distance criterion, R_h. Only the matrix elements corresponding to atom pairs with interatomic distance less than R_h are evaluated and stored. This cutoff reduces the memory storage so that it becomes linear with respect to the size of the system. In addition, the CPU time used for matrix element evaluation is significantly reduced since fewer of them are evaluated. Similarly, a smaller number of matrix elements for the one-electron core Hamiltonian and Fock matrices are evaluated due to this judicious cutoff criteria.

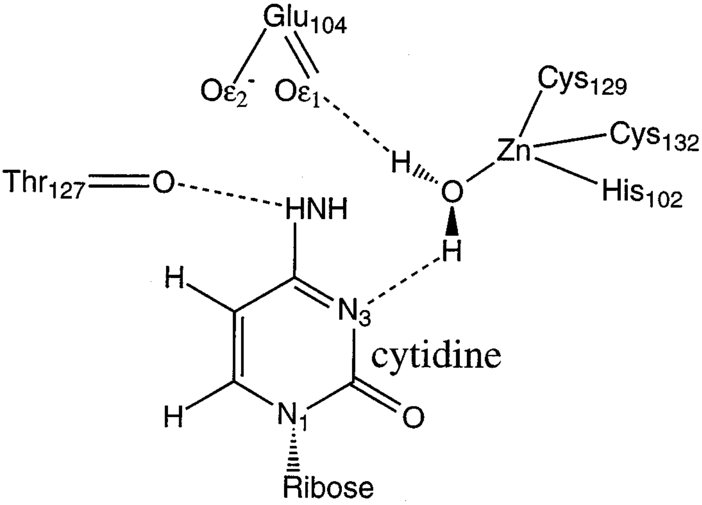
For solution phase calculations, a dielectric continuum model of the solvent (COSMO) is used. The solute charge distribution is represented by a set of atomic charges, dipole moments, and quadrupole moments, that induces a reaction field charge density on the solvent accessible surface of the solute [47, 61]. The solvent polarization effects on the solvation energetics of biological molecules have been calculated with the DAC approach [62, 63]. The error of screen energy in COSMO for the non-infinite dielectric constant solvent has been estimated as $1/2\epsilon$, where ϵ is the dieletric constant [64]. For water, whcih has $\epsilon = 78.5$, the error is less than 1%. For the worst case, vacuum has $\epsilon = 1.0$ and this introduces a 50% error in the screen energy. However, for solvent with $\epsilon = 1.0$, the solvation energy itself is quite small; hence the absolute error remains irrelevant. Therefore, even in low dielectric solvent, the COSMO dielectric continuum model should still work reasonably well.

We have described the main features of the DAC method and its implementation into a semiempirical quantum chemistry package—MOPAC/DAC. We have used MOPAC/DAC to perform simulations of a large-scale system (1330 atoms) including much of the protein environment surrounding an active site of CDA. The results of these simulations will be discussed in the next section (Section III).

III. APPLICATIONS TO CYTIDINE DEAMINASE

A. *Calculational Details*

As shown in Scheme I, the CDA active site includes a Zn *atom* bound by the thiolate sulfur atoms of the Cys-129 and Cys-132 side chains and by a nitrogen atom of the His-102 side chain. Bound to this Zn-complex is a $OH^{(-)}$ or water molecule which displaces the NH_2 group, located at the C^4 position on the cytidine ligand, via nucleophilic hydration of the N^3–C^4 double bond. The product of this displacement is apparently the conventional keto tautomer of uridine. The carboxylate group of Glu-104 near the NH_2 group on the cytidine ligand assists in the reaction both by stabilizing the tetrahedral transition-state and by acting as a proton shuttle [11–13]. Several crystallographically determined structures have been obtained for different complexes of CDA; each complex represents a point along



SCHEME I

the reaction pathway. Recently, we performed several calculations focusing on the nature of the active site and how it differs for each complex.

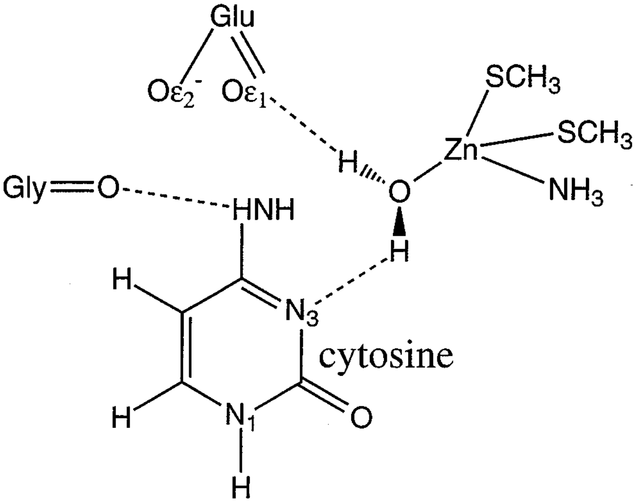
For these calculations, we created a system which represents roughly 30% of a monomer of the enzyme. This 1330-atom system, which surrounds the active site, was created by including any residue within 8.0 Å of the ligand (including contributions from the other monomer). Chain ends in the system were terminated by addition of acetyl or N-methyl groups. Our goal in choosing such a large system was for the purposes of including much of the protein environment in our calculations. Of course, a common approach for including a valid description of the protein environment surrounding the active site has been to combine QM methods with molecular mechanical (MM) force fields [65–78]. While QM/MM methods are promising, there remain difficulties in treating the boundaries between the QM and MM subsystems; these difficulties should be overcome in due time. As an alternative, we have taken advantage of the MOPAC/DAC implementations which allow larger QM simulations to be performed efficiently using purely QM methods. Performing larger QM simulations to account for a significant portion of the protein environment surrounding the active site has long been a goal in quantum chemistry.

In all of our semiempirical calculations, we use the PM3 Hamiltonian [52]. The PM3 parameterization is shown to work well for Zn complexes [79], and recently for simulations of the enzyme carbonic anhydrase which has a very similar hydrolytic mechanism and Zn tetrahedral coordination as found in CDA, although typically semiempirical methods give bond lengths on the order of 5% too large [77]. In addition, all of the optimized geometries of the largest systems were obtained from gas-phase calculations. Given that the active site of CDA is not solvent-accessible [8], the gas-phase approximation is a physically accurate description. *Ab initio* work on Zn complexes indicates that a reaction in an enzyme-active cavity may actually be better approximated by a gas-phase model rather than by a model reaction in solution [80]. This point is definitely more valid for CDA than in other zinc enzymes where the active site, such as in carbonic anhydrase, is partly solvent-accessible. Since our calculations include much of the protein environment surrounding the active site, our results are more physically reasonable than gas-phase calculations of a small system where only a few primary components of the active site are included.

A smaller system consisting of 154 atoms was created and calculations using both the PM3 semi-empirical Hamiltonian within MOPAC and DFT techniques within the DMol package [81] were performed as a comparison against the larger 1330-atom systems. This smaller system includes all the amino acids shown in Scheme I, such that residues 127–132 and residues 102–104 were included in the simulations. As in the larger 1330 systems, blocking end groups of acetyl or N-methyl were used to terminate chain ends. Residues not directly exposed to the active site were replaced by Glycine side chains (i.e., residues 128, 130, and 103).

Similar calculations using PM3 and DFT were also performed on an even smaller system containing only 77 atoms, as shown in Scheme II. This smallest system consists of modifying the system shown in Scheme I, such that Threonine-127 is replaced by Glycine, Cysteine-129 and Cysteine-132 are reduced to SCH_3, and Histidine-102 is reduced to NH_3. In addition, the ribose ring of the ligand is not included.

In the calculations for both the 154 and 77 atom systems a variety of dielectric constants were used within a dielectric continuum model. The DFT calculations were made using both the local and non-local approximations to the exchange-correlation interactions. Both minimal numerical basis sets and double numerical basis sets were used in the DFT



SCHEME II

calculations for a comparison. These results are intended to give a preliminary comparison of smaller sized systems to our larger system and do not represent a systematic study of convergence nor validation of the chosen theoretical (semiempirical) level.

For all of these (77, 154, and 1330 atom) systems, geometries of different starting arrangements were optimized, keeping backbone atoms fixed, with a 0.2 kcal/(mole Å) rms and 2.4 kcal/(mole Å) max tolerance in the gradients. Keeping the backbone atoms fixed is justified for these calculations by the fact that backbone atoms in several refined CDA complexes coincide with rms deviations of 0.15 Å for backbone atoms, and the only significant changes occur near the active site [13]. The initial starting structure for all of these systems was based on the x-ray crystallographic structure for the ground-state analog complex: 3-deazacytidine [12]. However, our initial structures where modified slightly to resemble the different protonation states of Glu-104 and Zn-H_2O which were minimized in our calculations. In addition, we replaced the 3-deaqzacytidine ligand with the cytidine ligand which correctly represents the ground-state complex of the enzyme.

B. *The Active Species of the Ground-State Complex*

Recently we performed calculations to specify the structure of the active species at the initiation of the reaction pathway [82]. In those results, which are summarized here, we addressed two important issues: (1) whether the active species consists of a zinc-coordinated hydroxide ion (Zn-$OH^{(-)}$) or a zinc-coordinated water molecule (Zn-H_2O); (2) which of the two carboxylate oxygen atoms of Glu-104 is protonated in the active species.

Initially, we optimized the geometry for a 1330-atom system which represents the active site, as shown in Scheme I, and the surrounding protein environment. This structure contains Zn-H_2O in the active site with the carboxylate group of Glu-104 unprotonated, and it represents the active site of the complex just before the active species is created by deprotonation of Zn-H_2O. Results of geometry optimization, shown in Fig. 1a, for this system indicate that the substrate water is stabilized by the carboxylate group of Glu-104 via hydrogen-bonding.

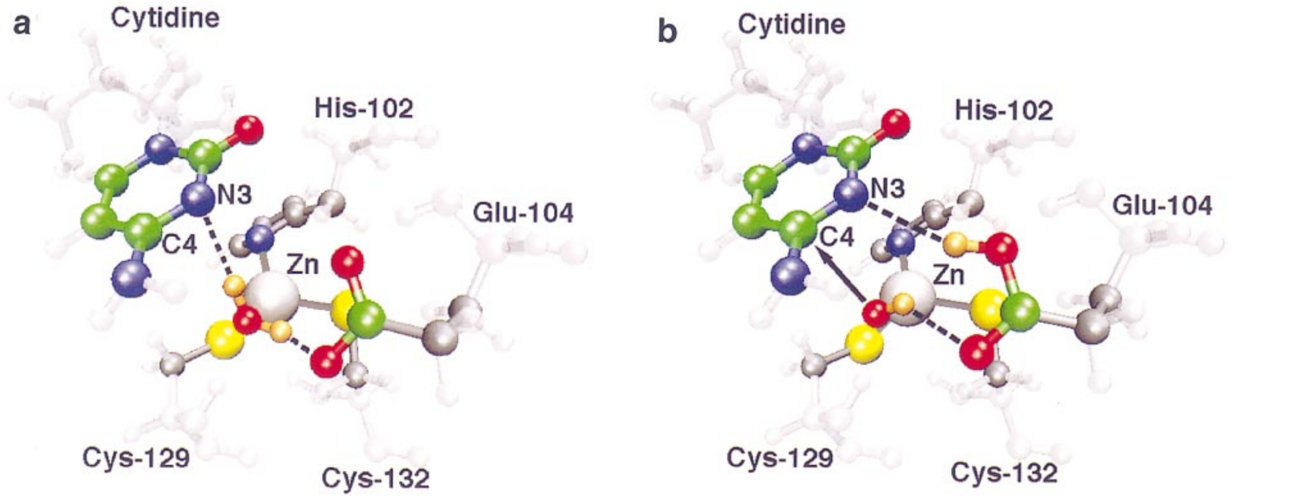


FIG. 1. Two structures for the active site of cytidine deaminase, whose geometries were optimized via divide-and-conquer within a semiempirical (PM3) framework. (a) Ground-state substrate complex just before the active species is created by deprotonation of Zn-H_2O. (b) This structure has the lowest energy of the structures calculated, and, as verified by experimental data, is the active species. A hydrogen-boning network is formed as a result of deprotonating the substrate H_2O and protonation of the carboxyl oxygen $O^{\epsilon 2}$ of Glu-104.

Activation of CDA above about pH 5 suggests that Zn-H_2O transfers a proton to the carboxylate group of Glu-104; however, two different carboxylate oxygen sites (O^{ϵ_1} and O^{ϵ_2}) could accept a proton and two possible protons could be donated by the substrate water. Four additional systems were created representing the configurations of these possible protonation arrangements and their geometries were optimized. The net charge of all of these 1330-atom systems remains the same since the proton from Zn-H_2O is not removed from the system but only displaced to a nearby location on the Glu-104 side chain.

We conclude that one structure, shown in Fig. 1b, is lower in energy and is prominently the lowest energy structure than the other four structures by at least 46.0 kcal/mol. In this structure, a hydrogen-bonding network is formed between the hydroxide proton of Zn-$OH^{(-)}$, the protonated O^{ϵ_2} of Glu-104, and the N^3 position of the cytidine (CTD) substrate. This hydrogen-bonding network gives credence to the conclusion that this structure is lower in energy, since the higher energy structures do not form such a nicely structured hydrogen-bonding network.

Considering all calculated structures, not only is the low energy structure, shown in Fig. 1b, lower in energy compared to the other two-proton structures, but it also correlates quite well in detail with the geometry observed at the active site of the enzyme in the x-ray crystal structure [12]. Both the crystal structure and our calculations for the low energy structure indicate that the oxygen of the hydroxide and C^4 of the cytidine ligand are significantly closer than the van der Waals limit. Moreover, this short OH-C^4 distance implies, in turn, that the nucleophilic attack is being initiated as the ground-state Zn-$OH^{(-)}$ structure is formed. Both evidence from the energetics and from comparison to crystallographic data supports the fact that the low energy structure closely resembles the structure of the active species.

Table I shows results for the 154- and 77-atom systems using several DFT models and the PM3 semiempirical under different solvation conditions with varying dielectric constants. The energies (kcal/mol) listed are the differences between the structure containing Zn-H_2O in the active site and the structure containing Zn-OH in the active site, where a negative number implies that the latter is lower in energy. Note that the net charge on the Zn-H_2O and Zn-OH structures is identical. The proton abstracted from Zn-H_2O is not removed from the Zn-OH system, but rather only displaced to the nearby Glu-104 side chain. The calculations for the 154-atom system show that the Zn-OH structure is consistently lower in energy compared to the Zn-H_2O structure which is the same conclusion that was obtained for the much larger 1330-atom system. However, the energy difference between the Zn-OH and Zn-H_2O structures is much greater for the 1330-atom system which might be attributable to the fact that this large system contains a significant portion of the surrounding protein environment.

The results for the 154-atom system, listed in Table I, support two points which were made regarding the larger 1330-atom system. The first point is that our choice to calculate our 1330 system within the gas-phase was judicious, and this was initially justified since the active site is solvent inaccessible. The results for the 154-atom system lend further support to this initial justification since the energy differences remain relatively unchanged with respect to differences in the dielectric constant (indicating that the active site for this model system is for the most part solvent inaccessible). Note that the energies for the 77-atom system are widely dependent on the choice of dielectric constant indicating that the active site for this smaller model system is more solvent accessible. The second point is that the results for the 154-atom system indicate that the energies remain relatively indifferent with respect to the DFT method used. This is not the case for the 77-atom system and it appears

TABLE I
Energy Differences for Different Models under Different Solvation Effects for Two Different Systems

Method	ϵ	154 atoms	77 atoms
PM3	1.0	−22.64	−2.17
	4.0	−23.67	0.75
	10.0	−23.66	2.22
	78.5	−23.52	3.41
LDA/Min	1.0	−14.76	29.29
	4.0	−14.78	
	10.0	−14.87	13.58
LDA/DN	1.0	−16.03	5.16
	10.0	−12.79	
	78.5		16.89
BPW/DNP	1.0	−14.09	−2.10
	78.5	−12.32	−10.04

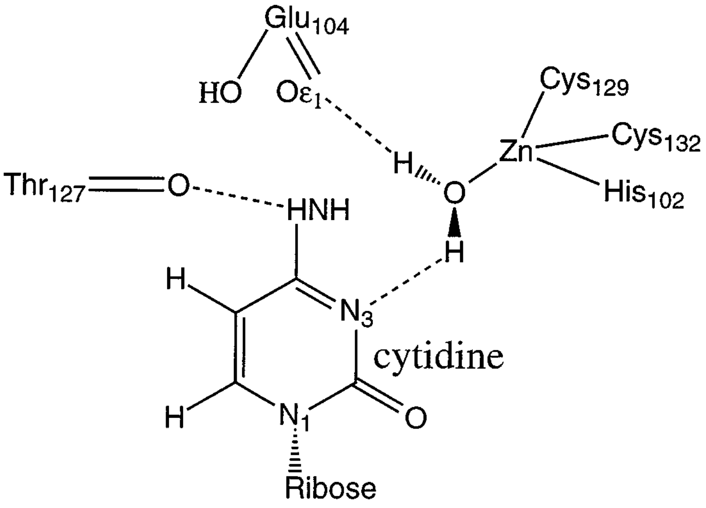
Note. The solvation model COSMO was used [47, 61], and the second column gives ϵ which signifies the dielectric constant used for that calculation (i.e., 1.0 = gas phase, and 78.5 = water). The last two columns are the energy differences (kcal/mol) between the structure containing Zn-H_2O in the active site and the structure containing Zn-OH in the active site. A negative number implies that the latter structure (Zn-OH) is lower in energy. The 154- and 77-atom systems are illustrated by Schemes I and II, respectively. Legend. PM3, semiempirical Hamiltonian within the MOPAC routine [57]; LDA/Min, local density approximation with minimal numerical basis set; LDA/DN, local density approximation with double numerical basis set; BPW/DNP, Becke ’88 3 parameterization formula for the exchange energy [87], Perdew and Wang ’91 gradient corrected formula for the correlation energy [88, 89], double numerical plus polarization basis set.

that larger systems may be less sensitive to the theoretical method employed. This suggests that for our large 1330-atom system use of the PM3 semiempirical method should give quite reasonable results with respect to the much more computationally intensive use of DFT.

Systems with protonating arrangements corresponding to low- and high-pH states were also considered [82]. The low-pH structure consists of a three-proton arrangement containing Zn-H_2O with the $O^{\epsilon 2}$ of Glu-104 protonated as shown in Scheme III. We predicted that at low pH the Zn-O bond length should be quite large, as if the substrate H_2O is very weakly bound to the Zn atom. Experimental x-ray crystallography data have been collected for the ligand-free CDA structure at low pH. Preliminary comparisons of the low-pH structure to the ligand-free CDA structure under normal conditions indicate that indeed the substrate H_2O is displaced away from the Zn atom under the low-pH conditions [83].

C. *The Valence Buffer Effect in Cytidine Deaminase*

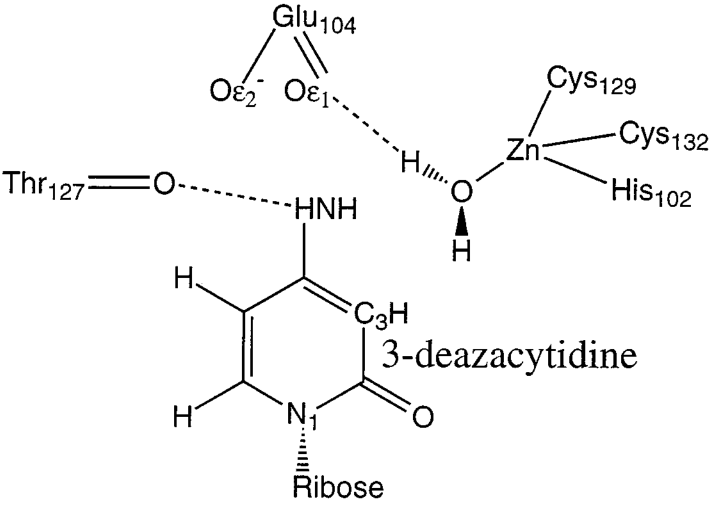
Several crystallographic complexed structures, which represent points along the reaction pathway, have been experimentally determined. We performed calculations on three of these structures [84], and in this section a summary of the results are provided. The first structure, as shown in Scheme IV, is a complex of CDA with 3-deazacytidine, whose structure resembles the ground-state complex with its cytidine substrate.



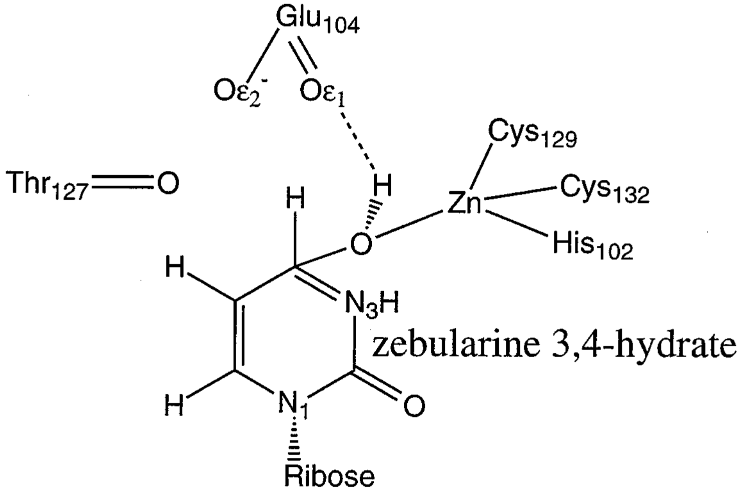
SCHEME III

The second structure, as shown in Scheme V, is a complex of zebularine 3-4 hydrate which is an analog structure resembling the transition-state complex. The final and third structure, as shown in Scheme VI, is a complex of uridine which is the product-state. For each of these structures a 1330-atom subsystem was created and the geometry optimized as described for the ground-state complex in Subsection III.A.

One goal for calculating these three structures is to investigate the bond distance between the Zn atom and the two S_γ atoms of the Cys-129 and the Cys-132 side chains, $d_{Zn-S_\gamma 129}$ and $d_{Zn-S_\gamma 132}$. Xiang *et al.* report that each distance increases from the deazacytidine complex to the zebularine 3–4 hydrate complex [12]. In their work they find that more significant increases are found for the Zn-S_γ132 bond distance since in the transition-state analog structure there is a shortening of the hydrogen bond formed between Glu-104 and the 4-OH oxygen of the ligand. Xiang *et al.* further report that the two bond distances, $d_{Zn-S_\gamma 129}$ and



SCHEME IV

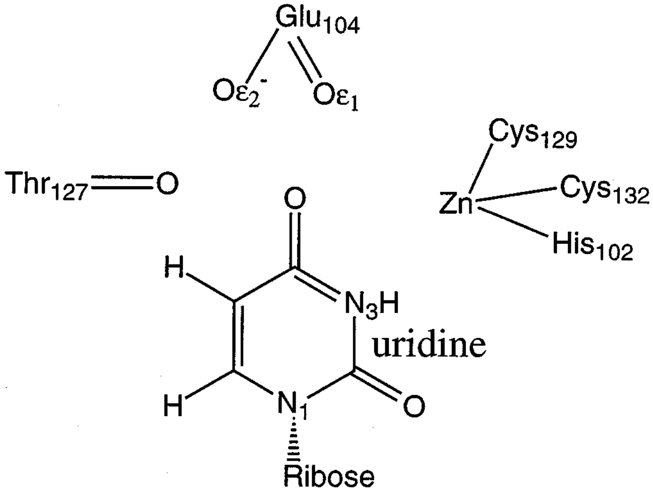


SCHEME V

$d_{Zn-S_\gamma 132}$, again decrease from the zebularine 3-4 hydrate complex to the uridine complex, with the more significant decrease for the Zn-S_γ132 bond distance.

This increase/decrease of the Zn-S_γ129 and Zn-S_γ132 bond distances represents what is defined as the "valence-buffer" mechanism [12, 85, 86]. The bond valence determines the strength of the bond to a metal which varies inversely with the distance. Therefore, the experimental results show that these two bonds weaken as the enzyme progresses from the ground-state to the transition-state, and strengthen as the enzyme progresses from the transition-state to the product state. More significant increases/decreases of the Zn-S_γ 132 bond distance indicate that this bond is weakened/strengthened more significantly.

Table II shows the results for the bond distances, $d_{Zn-S_\gamma 129}$ and $d_{Zn-S_\gamma 132}$, obtained theoretically and experimentally for different CDA complexes. Our results do not compare well to the experimental bond distances for the deazacytidine complex which show that



SCHEME VI

TABLE II
The Calculated Distances (in Å) between the Zn Atom and the Two S_γ Atoms of the Cys-129 and Cys-132 Residues

CDA complex	Theoretical $d_{Zn-S_\gamma 129}$	Theoretical $d_{Zn-S_\gamma 132}$	Experimental $d_{Zn-S_\gamma 129}$	Experimental $d_{Zn-S_\gamma 132}$
Cytidine	2.433	2.428		
Deazacytidine	2.413	2.345	2.301	2.087
Zebularine 3-4 hydrate	2.428	2.382	2.407	2.346
Uridine	2.407	2.332	2.310	2.265

Note. These distances are taken from the optimized geometries of different CDA complex structures. Each optimized structure was represented by a 1330-atom subsystem surrounding the active site.

$d_{Zn-S_\gamma 129}$ is significantly larger than $d_{Zn-S_\gamma 132}$, although typically semiempirical methods give bond lengths on the order of 5% too large [77]. However, our results do verify that the these two bond distances increase as the structure changes from the ground-state analog complex to the transition state analog complex and vice versa as the structure changes from the transition-state analog complex to the product complex. Our results further verify that the Zn-S_γ132 is weakened in the transition-state analog complex as there is a shortening of the hydrogen-bond formed between Glu-104 and the 4-OH oxygen of the ligand.

IV. SUMMARY

The divide-and-conquer methodology has made possible QM calculations for large biological systems. Quantum mechanical modeling of enzymes can offer a great deal of information and understanding to complement experimental study. Even at the semiempirical level, we are able to model the structure and energetics of enzyme systems, as demonstrated here for CDA. We have accurately determined the nature of the active species structure for CDA, as well as predicted the low-pH structure. Experimental crystallographic data suggest that our prediction of the low-pH structure is encouraging. Finally, we verify the valence-buffer effect, whereby the bond distance for Zn-S_γ132 is weakened as the enzyme passes through the transition state. Rapid progress in the development of linear-scaling techniques, combined with molecular dynamics, will make great progress in making molecular modeling an equal partner of experimental research in biochemistry and biophysics.

REFERENCES

1. R. M. Cohen and R. Wolfenden, Cytidine deaminase from *Escherichia coli:* Purification, properties, and inhibition by the potential transition state analog 3,4,5,6-tetrahydrouridine, *J. Biol. Chem.* **246**, 7561 (1971).
2. R. M. Cohen and R. Wolfenden, The equilibrium of hydrolytic deamination of cytidine and N^4-methylcytidine, *J. Biol. Chem.* **246**, 7566 (1971).
3. V. E. Marquez, P. S. Liu, J. A. Kelley, J. S. Driscoll, and J. J. McCormick, Synthesis of 1,3-diazepin-2-one nucleosides as transition state inhibitors of cytidine deaminase, *J. Med. Chem.* **23**, 713 (1980).
4. P. S. Liu, V. E. Marquez, J. S. Driscoll, R. W. Fuller, and J. J. McCormick, Cyclic urea nucleosids. Cytidine deaminase activity as a function of aglycon ring size, *J. Med. Chem.* **24**, 662 (1981).

5. G. W. Ashley and P. A. Bartlett, Inhibition of *Escherichia coli* cytidine deaminase by a phospha-pyrimidine nucleoside, *J. Biol. Chem.* **259**, 13,621 (1984).
6. C. H. Kim, V. E. Marquez, D. T. Mao, D. R. Haines, and J. J. McCormick, Synthesis of pyrimidine-2-one nucleosides as acid-stable inhibitors of cytidine deaminase, *J. Med. Chem.* **29**, 1374 (1986).
7. D. K. Wilson, F. B. Rudolph, and F. A. Quiocho, Atomic structure of adenosine deaminase complexed with a transition state analog: Understanding catalysis and immunodeficiency mutations, *Biochemistry* **252**, 1278 (1991).
8. L. Betts, S. Xiang, S. A. Short, R. Wolfenden, and C. W. Carter, Jr., Cytidine deaminase: The 2.3 Å crystal structure of an enzyme: Transition-state analog complex, *J. Mol. Biol.* **235**, 635 (1994).
9. D. C. Carlow, A. A. Smith, C. C. Yang, S. A. Short, and R. Wolfenden, Major contribution of a carboxymethyl group to transition-state stabilization by cytidine deaminase: Mutation and rescue, *Biochemistry* **34**, 4220 (1995).
10. S. Xiang, S. A. Short, R. Wolfenden, and C. W. Carter, Jr., Transition-state selectivity for a single hydroxyl group during catalysis by cytidine deaminase, *Biochemistry* **34**, 4516 (1995).
11. D. C. Carlow, S. A. Short, and R. Wolfenden, Role of glutamate-104 in generating a transition state analogue inhibitor at the active site of cytidine deaminase, *Biochemistry* **35**, 948 (1996).
12. S. Xiang, S. A. Short, R. Wolfenden, and C. W. Carter, Jr., Cytidine deaminase complexed to 3-deazacytidine: A "valence buffer" in zinc enzyme catalysis, *Biochemistry* **35**, 1335 (1996).
13. S. Xiang, S. A. Short, R. Wolfenden, and C. W. Carter, Jr., The structure of the cytidine deaminase-product complex provides evidence for efficient proton transfer and ground-state destabilization, *Biochemistry* **36**, 4768 (1997).
14. L. Frick, J. P. M. Neela, and R. Wolfenden, Transition state stabilization by deaminases: Rates of nonenzymatic hydrolysis of adenosine and cytidine, *Bioorg. Chem.* **15**, 100 (1987).
15. V. E. Marquez, in *Developments in Cancer Chemotherapy*, edited by R. I. Glazer (CRC Press, Boca Raton, FL, 1984), p. 91.
16. B. Chandrasekaren, R. L. Capizzi, T. E. Kute, T. Morgan, and J. Dimling, Modulation of the metabolism and pharacokinetics of 1-β-D-arabinofuranosyluracil in leukemic mice, *Cancer Res.* **49**, 3259 (1989).
17. W. Kohn and L. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev. A* **140**, 1133 (1965).
18. R. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford Univ. Press, New York, 1989).
19. W. Yang, Direct calculation of electron density in density-functional theory, *Phys. Rev. Lett.* **66**, 1438 (1991).
20. W. Yang and T.-S. Lee, A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules, *J. Chem. Phys.* **163**, 5674 (1995).
21. S. Baroni and P. Giannozzi, Towards very large scale electronic structure calculations, *Europhys. Lett.* **17**, 547 (1992).
22. G. Galli and M. Parrinello, Large scale electronic structure calculations, *Phys. Rev. Lett.* **69**, 3547 (1992).
23. F. Mauri, G. Galli, and R. Car, Orbital formulation for electronic-structure calculations with linear system-size scaling, *Phys. Rev. B* **47**, 9973 (1993).
24. P. Ordejón, D. Drabold, M. Grumbach, and R. M. Martin, Unconstrained minimization approach for electronic computations that scales linearly with system size, *Phys. Rev. B* **48**, 14,646 (1993).
25. P. Ordejón, E. Artacho, and J. M. Soler, Self-consistent order-N density functional calculations for very large systems, *Phys. Rev. B* **53**, R10,441 (1996).
26. J. Kim, F. Mauri, and G. Galli, Total energy global optimizations using nonorthogonal localized orbitals, *Phys. Rev. B* **52**, 1640 (1995).
27. X.-P. Li, R. W. Nunes, and D. Vanderbilt, Density-matrix electronic-structure method with linear system-size scaling, *Phys. Rev. B* **47**, 10,891 (1993).
28. M. S. Daw, Model for energetics of solids based on the density matrix, *Phys. Rev. B* **47**, 10,895 (1993).
29. D. A. Drabold and O. F. Sankey, Maximum entropy approach for linear scaling in the electronic structure problem, *Phys. Rev. Lett.* **70**, 3631 (1993).

30. W. Zhong, D. Tománek, and G. F. Bertsch, Total energy calculations for extremely large clusters: The recursive method, *Solid State Comm.* **86**, 607 (1993).

31. M. Aoki, Rapidly convergent bond order expansion for atomistic simulations, *Phys. Rev. Lett.* **71**, 3842 (1993).

32. A. P. Horsfield, A. M. Bratkovsky, M. Fearn, D. G. Pettifor, and M. Aoki, Bond-order potentials: Theory and implementation, *Phys. Rev. B* **53**, 12,694 (1996).

33. S. Goedecker and L. Colombo, Efficient linear scaling algorithm for tight-binding molecular dynamics, *Phys. Rev. Lett.* **73**, 122 (1994).

34. E. B. Stechel, A. R. Williams, and P. J. Feibelman, N-scaling algorithms for density-functional calculations of metals and insulators, *Phys. Rev. B* **49**, 10,088 (1994).

35. W. Hierse and E. B. Stechel, Order-N methods in self-consistent density-functional calculations, *Phys. Rev. B* **50**, 17,811 (1994).

36. E. Hernández and M. J. Gillan, Self-consistent first-principles techniques with linear scaling, *Phys. Rev. B* **51**, 10,157 (1994).

37. E. Hernández, M. J. Gillan, and C. M. Goringe, Linear-scaling density-functional-theory technique: The density-matrix approach, *Phys. Rev. B* **53**, 7147 (1996).

38. W. Yang and J. M. Pérez-Jordá, in *Encyclopedia of Computational Chemistry*, edited by P. Schleyer (Wiley, New York), in press.

39. W. Yang, Direct calculation of electron density in density-functional theory: Implementation for benzene and a tetrapeptide, *Phys. Rev. A* **44**, 7823 (1991).

40. S. L. Dixon and K. M. Merz, Jr., Semiempirical molecular orbital calculations with linear system size scaling, *J. Chem. Phys.* **104**, 6643 (1996).

41. S. L. Dixon and K. M. Merz, Jr., Fast, accurate semiempirical molecular orbital calculations for macromolecules, *J. Chem. Phys.* **107**, 879 (1997).

42. R. T. Gallant and A. St-Amant, Linear scaling for the charge density fitting procedure of the linear combination of Gaussian-type orbitals density functional method, *Chem. Phys. Lett.* **256**, 569 (1996).

43. S. K. Goh and A. St-Amant, Using a fitted electronic density to improve the efficiency of a linear combination of Gaussian-type orbitals calculation, *Chem. Phys. Lett.* **264**, 9 (1997).

44. W. Thiel, Semiempirical methods: Current status and perspectives, *Tetrahedron* **44**, 7393 (1988).

45. J. J. P. Stewart, *Review in Computational Chemistry* (VCH, New York, 1990), Vol. 1, p. 45.

46. J. J. P. Stewart, MOPAC: A semiempirical molecular orbital program, *J. Comput. Aided Mol. Design* **4**, 1 (1990).

47. T.-S. Lee, D. York, and W. Yang, Linear-scaling semiempirical quantum calculations for macromolecules, *J. Chem. Phys.* **105**, 2744 (1996).

48. J. J. P. Stewart, Application of localized molecular orbitals to the solution of the semiempirical self-consistent field equations, *Int. J. Quantum Chem.* **58**, 133 (1996).

49. A. D. Daniels, J. M. Willam, and G. E. Scuseria, Semiempirical methods with conjugate gradient density matrix search to replace diagonalization for molecular systems containing thousands of atoms, *J. Chem. Phys.* **107**, 425 (1997).

50. M. J. S. Dewar and W. Thiel, Ground states of molecules, 38, the MNDO method: Approximations and parameters, *J. Am. Chem. Soc.* **99**, 4899 (1977).

51. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, AM1: A new general purpose quantum mechanical model, *J. Am. Chem. Soc.* **107**, 3902 (1985).

52. J. J. P. Stewart, Optimization of parameters for semiempirical methods. I. Method, *J. Comp. Chem.* **10**, 209 (1989).

53. J. J. P. Stewart, Optimization of parameters for semiempirical methods. I. Applications, *J. Comp. Chem.* **10**, 221 (1989).

54. M. J. S. Dewar and W. Thiel, A. semiempirical model for the two-center repulsion integrals in the NDDO approximation, *Theor. Chem. Acta* **46**, 89 (1977).

55. R. S. Mulliken, Criteria for the construction of good self-consistent-field molecular orbital wave functions, and the significance of LCAO-MO population analysis, *J. Chem. Phys.* **36**, 3428 (1962).

56. Q. Zhao and W. Yang, Analytical energy gradients and geometry optimization in the divide-and-conquer method for large molecules, *J. Chem. Phys.* **102**, 9598 (1995).

57. J. J. P. Stewart, *MOPAC7 Version 2 Manual* (QCPE, Bloomington, 1993).

58. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes* (Cambridge Univ. Press, Cambridge, UK, 1992).

59. M. J. S. Dewar and Y. Yamaguchi, Analytical first derivatives of the energy in MNDO, *Comput. Chem.* **2**, 25 (1978).

60. C. Lee and W. Yang, The divide-and-conquer density-functional approach: Molecular internal rotation and density of states, *J. Chem. Phys.* **96**, 2408 (1992).

61. D. York, T.-S. Lee, and W. Yang, Parameterization and efficient implementation of a solvent model for linear-scaling semiempirical quantum mechanical calculations of biological macromolecules, *Chem. Phys. Lett.* **263**, 297 (1996).

62. D. M. York, T.-S. Lee, and W. Yang, Quantum mechanical study of aqueous polarization effects on biological macromolecules, *J. Am. Chem. Soc. Comm.* **118**, 10,940 (1996).

63. D. M. York, T.-S. Lee, and W. Yang, Quantum mechanical treatment of biological macromolecules in solution using linear-scaling electronic structure methods, *Phys. Rev. Lett.* **118**, 5011 (1998).

64. A. Klamt and G. Shuurmann, COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient, *Perkins Trans.* **2**, 799 (1993).

65. A. Warshel and M. Levitt, Theoretic studies of enzymic reactions: Dielectric electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme, *J. Mol. Biol.* **103**, 227 (1976).

66. U. C. Singh and P. A. Kollman, A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $CH_3CI + CI^-$ exchange reaction and gas phase protonation of polyethers, *J. Comp. Chem.* **7**, 718 (1986).

67. M. J. Field, P. A. Bash, and M. Karplus, A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations, *J. Comp. Chem.* **11**, 700 (1990).

68. J. Gao and X. Xia, A priori evaluation of aqueous polarization effects through Monte Carlo QM-MM simulations, *Science* **258**, 631 (1992).

69. A. Warshel, Computer simulations of enzymatic reactions, *Curr. Opin. Struct. Bio.* **2**, 230 (1992).

70. R. V. Stanton, D. S. Hartsough, and K. M. Merz, Jr., Calculation of solvation free energies using a density functional/molecular dynamics coupled potential, *J. Phys. Chem.* **97**, 11,868 (1993).

71. D. Wei and D. R. Salahub, A combined density-functional and molecular dynamics simulation of a quantum water molecule in aqueous solution, *Chem. Phys. Lett.* **224**, 291 (1994).

72. V. Thery, D. Rinaldi, and J. L. Rivail, Quantum mechanical computations on very large molecular systems: The local self-consistent field method, *J. Comp. Chem.* **15**, 269 (1994).

73. J. Gao, A combined QM/MM simulation study of the claisen rearrangement of allyl vinyl ether in aqueous solution, *J. Am. Chem. Soc.* **116**, 1563 (1994).

74. F. Maseras and K. Morokuma, IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states, *J. Comp. Chem.* **16**, 1170 (1995).

75. S. Humbel, S. Sieber, and K. Morokuma, The IMOMO method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and SN2 reaction: $RCI + CI-$, *J. Chem. Phys.* **105**, 1959 (1996).

76. X. Assfeld and J. L. Rivail, Quantum chemical computations on parts of large molecules: The *ab initio* local self-consistent field approach, *Chem. Phys. Lett.* **263**, 100 (1996).

77. K. M. Merz, Jr., and L. Banci, Binding of bicarbonate to human carbonic anhydrase II: A continuum of binding states, *J. Am. Chem. Soc.* **119**, 863 (1997).

78. M. Peräkyla and P. A. Kollman, A simulation of the catalytic mechanism of aspartyl-glucosaminidase using *ab initio* quantum mechanics and molecular dynamics, *J. Am. Chem. Soc.* **119**, 1189 (1997).

79. J. J. P. Stewart, Optimization of parameters for semi-empirical methods. III. Extension of PM3 to Be, Mg, Zn, Ga, Ge, As, Se, Cd, In, Sn, Sb, Te, Hg, TI, Pb, and Bi, *J. Comp. Chem.* **12**, 320 (1991).

80. I. Bertini, C. Luchinat, M. Rosi, A. Sgamelloti, and F. Tarantelli, pK_a of zinc-bound water and nucleophilicity of hydroxo- containing species: *Ab initio* calculations on models for zinc enzymes, *Inorg. Chem.* **29**, 1460 (1990).

81. DMoI, product of Biosym Technologies, Inc., San Diego, CA, 92121.

82. J. P. Lewis, C. W. Carter, Jr., J. Hermans, W. Pan, T. S. Lee, and W. Yang, Active species for the ground-state complex of cytidine deaminase: A linear-scaling quantum mechanical investigation, *J. Am. Chem. Soc.* **120**, 5407 (1998).

83. J. Erkmann, J. C. W. Carter, J. P. Lewis, J. Hermans, and W. Yang, Low-pH studies of cytidine deaminase, in preparation.

84. J. P. Lewis, S. Liu, C. W. Carter, Jr., J. Hermans, T. S. Lee, and W. Yang, A linear-scaling quantum mechanical investigation of the “valence-buffer” effect in cytidine deaminase, in preparation.

85. I. D. Brown and R. D. Shannon, Empirical bond-strength-bond-length curves for oxides, *Acta Crystallogr. A* **29**, 266 (1973).

86. I. D. Brown, Chemical and steric constraints in inorganic solids, *Acta Crystallogr. B* **48**, 553 (1992).

87. A. D. Becke, A multicenter numerical integration scheme for polyatomic molecules, *J. Chem. Phys.* **88**, 2547 (1988).

88. J. P. Perdew and Y. Wang, Accurate and simple analytic representation of the electron-gas correlation energy, *Phys. Rev. B* **45**, 13,244 (1992).

89. A. D. Becke, Density-functional thermochemistry. II. The effect of the Perdew–Wang generalized-gradient correlation correction, *J. Chem. Phys.* **97**, 9173 (1992).