Geometry Optimization of Kringle 1 of Plasminogen Using the PM3 Semiempirical Method

ANDREW D. DANIELS,[1] GUSTAVO E. SCUSERIA,[1] ÖDÖN FARKAS,[2,*] H. BERNHARD SCHLEGEL[2]

[1]*Department of Chemistry and Center for Nanoscale Science and Technology, Rice University, P.O. Box 1892, Houston, Texas 77251-1892*
[2]*Department of Chemistry, Wayne State University, Detroit, Michigan 48202*

Received 1 July 1999; accepted 7 July 1999

ABSTRACT: The results of a geometry optimization on the 1226 atom Kringle 1 of plasminogen are presented. The energy and gradients were calculated using a linear-scaling PM3 semiempirical method with a conjugate gradient density matrix search replacing the diagonalization step. The geometry was optimized with the rational function optimization technique combined with a modified version of the direct inversion in the iterative subspace procedure. The optimization required 362 geometry update steps to reach a local minimum. An analysis is given of the optimization and timing results using a single processor on the SGI Origin2000. © 2000 John Wiley & Sons, Inc. Int J Quant Chem 77: 82–89, 2000

Correspondence to: G. E. Scuseria.
**Permanent address:* Department of Organic Chemistry, Eötvös Loránd University, P.O. Box 32, Budapest 112., H-1518, Hungary.
Contract grant sponsor: National Science Foundation.
Contract grant number: CHE-9618323.
Contract grant sponsor: Gaussian, Inc.
Contract grant sponsor: Keck Center for Computational Biology.
Contract grant number: LM07093.

© 2000 John Wiley & Sons, Inc.
CCC 0020-7608 / 00 / 010082-08

Introduction

Until recently, large biomolecules such as proteins were outside the computational reaches of quantum-mechanical calculations. However, throughout the 1990s, much effort has been spent developing linear-scaling semiempirical algorithms, which make it possible to compute energies and gradients for molecules containing thousands of atoms. Using these methods, it is now possible to perform semiempirical geometry optimizations on large biological molecules, such a proteins. Most of the semiempirical geometry optimizations performed on proteins thus far have been carried out using the divide and conquer (DAC) approach to reduce the time required for the diagonalization step to linear scaling with the number of atoms. Lewis and coworkers performed a geometry optimization on a 1330 atom model for the cytidine deaminase active site [1]. They utilized the semiempirical PM3 Hamiltonian [2] with DAC to obtain linear scaling. For increased speed, the protein backbone was held fixed during the calculation. Later, Vincent et al. also used a linear-scaling PM3 semiempirical program with DAC to carry out a full optimization on the geometry of the 1960 atom hen egg white lysozyme [3]. One difficulty with DAC methods is that the errors introduced by replacing diagonalization with DAC are difficult to control. Even though the DAC approximation is exact when the subsystem sizes reach the size of the entire system, the error is not a simple function of the subsystem size. Thus, it is difficult to predict the accuracy of a DAC calculation without also performing a calculation using diagonalization.

Therefore, if other linear-scaling methods for replacing diagonalization could be found which only contain parameters in which the error as compared with the diagonalization result is predictable, they might be more appropriate candidates for performing geometry optimizations on large biological molecules. In a previous article, we presented a performance comparison of several linear-scaling replacements for diagonalization that meet this criterion [4]. One is conjugate gradient density matrix search (CGDMS) which minimizes an energy functional with respect to the density matrix [5]. Another is pseudodiagonalization (PD) which works by using Jacobi notations on a guess set of orbitals for the system [6]. A third candidate is purification of the density matrix which performs transformations on a guess density matrix to drive it toward idempotency [7]. The last method, the Chebyshev expansion method (CEM), forms the density matrix as a polynomial expansion of the Hamiltonian [8]. In these linear-scaling diagonalization replacements, the introduced errors are easily controlled by changing simple thresholds. The errors depend on these thresholds in a way that is, for the most part, system-independent or is otherwise predictable. As the thresholds approach zero, the methods become exact. Thus, these methods might be more desirable than is DAC for use in large-molecule optimizations. The first such semiempirical geometry optimization on a protein was carried out by Stewart who optimized the geometry of a 740 atom crambin molecule using PM3 with PD [9].

In this article, we present a gas-phase geometry optimization on the 1226 atom kringle 1 of plasminogen using PM3 implemented with CGDMS. Plasminogen is a protein present in human blood plasma and is a key component in the fibrolytic mechanism. It is also thought to play a role in tissue repair, malignant transformation, macrophage function, ovulation, and embryo implantation. This is, to the best of our knowledge, the largest protein to be optimized semiempirically using a method other than DAC to achieve linear scaling of the computational time with system size.

Methods

The geometry optimization was carried out using the linear-scaling PM3 code within a developmental version of the *Gaussian* suite of programs [10]. The diagonalization step of the SCF calculation is replaced by CGDMS. CGDMS searches for the density matrix directly by using the method of conjugate gradients to minimize a functional with respect to the density matrix. To obtain linear scaling with CGDMS, the zero elements of the density and Fock matrices must be neglected. In this calculation, all matrix elements below the neglect threshold of 1×10^{-5} au are discarded. Also, a distance cutoff of 15 Å is used to determine the form of the Fock matrix. These parameters are set

such that an accuracy in energy as compared to the diagonalization result of about 0.05 kcal/mol is ensured. For more information about the implementation of CGDMS and its parameters, see [4, 5].

The geometry was updated via an $\mathcal{O}(N^2)$ scaling rational function optimization (RFO) technique (developed from the regular RFO method by Farkas and Schlegel [11]) combined with a modified version of the direct inversion in the iterative subspace procedure (GDIIS) (introduced by Császár and Pulay [12] and modified by Farkas and Schlegel [13]) using redundant internal coordinates. This optimization technique has been found to converge the geometries of large systems with about the same number of steps as that of the regular quasi-Newton based methods, but diagonalization of the Hessian matrix via an $\mathcal{O}(N^3)$ scaling operation is not required. The GDIIS procedure uses the information from the previous points; therefore, it can quickly recover after steps resulting in higher energy and large forces. The coordinate transformations of the gradients and geometry updates between redundant interval and Cartesian coordinates were performed via a very fast $\mathcal{O}(N^2)$ scaling algorithm introduced by Farkas and Schlegel [14]. Even though the geometry updates step scales as $\mathcal{O}(N^2)$ with the system size, the CPU time that it requires in the plasminogen calculation is small compared to the energy and gradient update CPU time as discussed below.

Starting coordinates for kringle 1 of plasminogen were obtained from the Protein Data Base [15]. Hydrogen atoms were added to the structure using *GaussView* [16], with the resulting geometry relaxed using the MM3 force field. This structure was used for input for the PM3 optimization using CGDMS as a replacement for diagonalization.

Results and Discussion

All calculations were carried out on a single MIPS R10K/195 MHz processor of an SGI Origin2000 computer. When optimizing large molecules such as proteins, it is not entirely clear at which point to stop the geometry optimization. The defaults in *Gaussian* for geometry-optimization convergence criteria are the following: RMS gradient = 0.36 kcal/mol/Å, maximum gradient = 0.53 kcal/mol/Å, RMS displacement = 0.0012 au, and maximum displacement = 0.0018 au, where the displacement is the Cartesian displacement measured in bohrs. However, proteins contain large, floppy chains, causing the Cartesian displacement to be quite large with even small changes in angles and dihedrals. We decided that a better convergence criterion for this optimization would be to use the change in internal coordinates for the system instead of the Cartesian displacement. Also, for such a large system, very small changes in bond lengths and angles should not be as important as in smaller molecules. Thus, it is not necessary to determine the geometry of the structure to such a high accuracy. We decided to loosen the convergence thresholds for this optimization to an RMS gradient = 3.6 kcal/mol/Å, maximum gradient = 5.3 kcal/mol/Å, and RMS internal coordinate displacement = 0.012 au (bohrs/radians).

Using these criteria, and starting from the MM3 geometry, the PM3 optimization converged in 362 geometry updates, much faster than the number of steps reported by Stewart [9] for the 740 atom crambin molecule (over 2000 cycles) and Vincent et al. [3] for the 1960 atom hen egg white lysozyme (1023 cycles). However, these numbers of cycles are not directly comparable because the convergence criteria differ between calculations. Also, different proteins are used in each case, so the starting structures of the various proteins are of different degrees of accuracy and the geometries themselves have different convergence properties. Our calculation confirms the findings of Stewart and Vincent that a large number of steps is required to optimize the geometry of proteins. Nevertheless, the number of steps in our calculation is about one-third the number of atoms. The average time for the energy and gradient update is 64.8 CPU min, and for the geometry update, 9.8 CPU min. Thus, the entire calculation required about 18.8 CPU days to complete.

The RMS gradient, as shown in Figures 1 and 2, oscillates quite wildly as the optimization progresses. This trend was also found in previous calculations [3, 9]. However, these fluctuations become smaller in size as the optimization proceeds, so that near the end of the calculation (in the last 20 steps) the maximum oscillations in the RMS gradient are no larger than 1 kcal/mol/Å.

Figures 3 and 4 show the RMS internal coordinate displacement as a function of the geometry

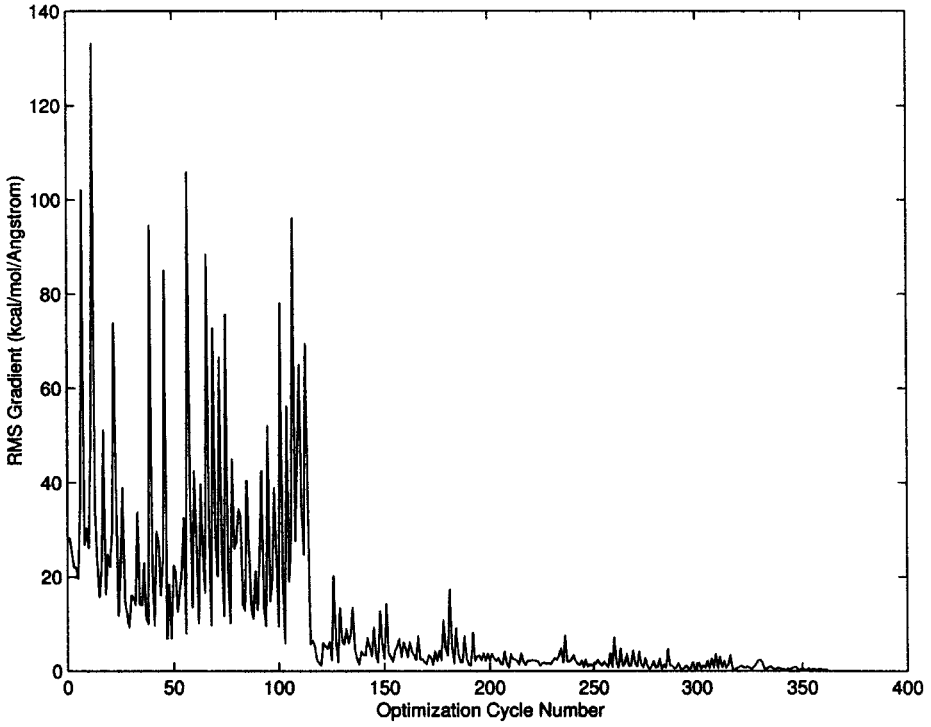

FIGURE 1. RMS gradient (kcal / mol / Å) is shown as a function of the geometry optimization step.

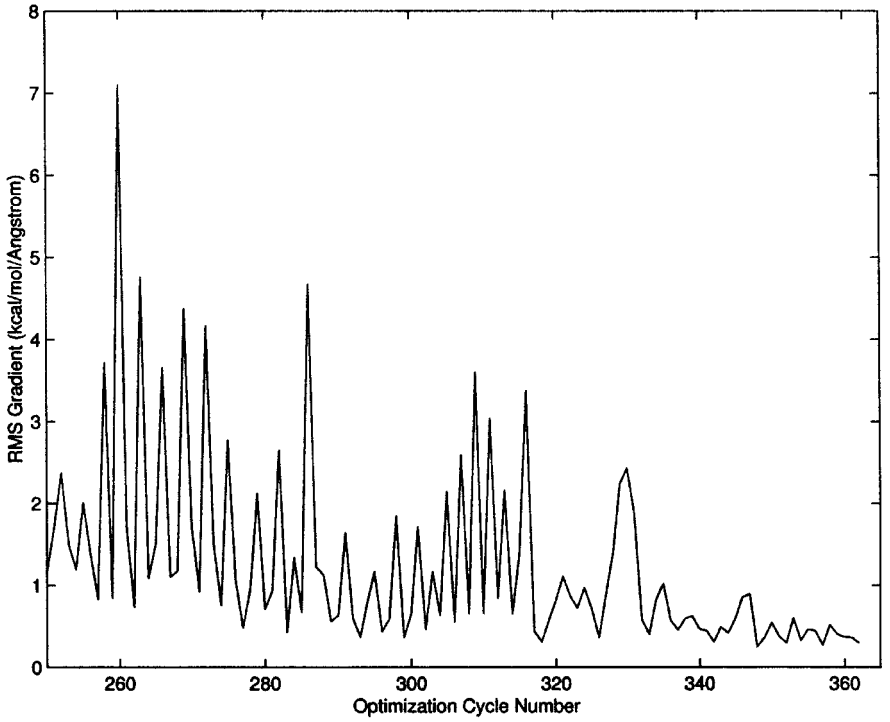

FIGURE 2. RMS gradient (kcal / mol / Å) is shown as a function of the geometry optimization step for the final portion of the optimization.

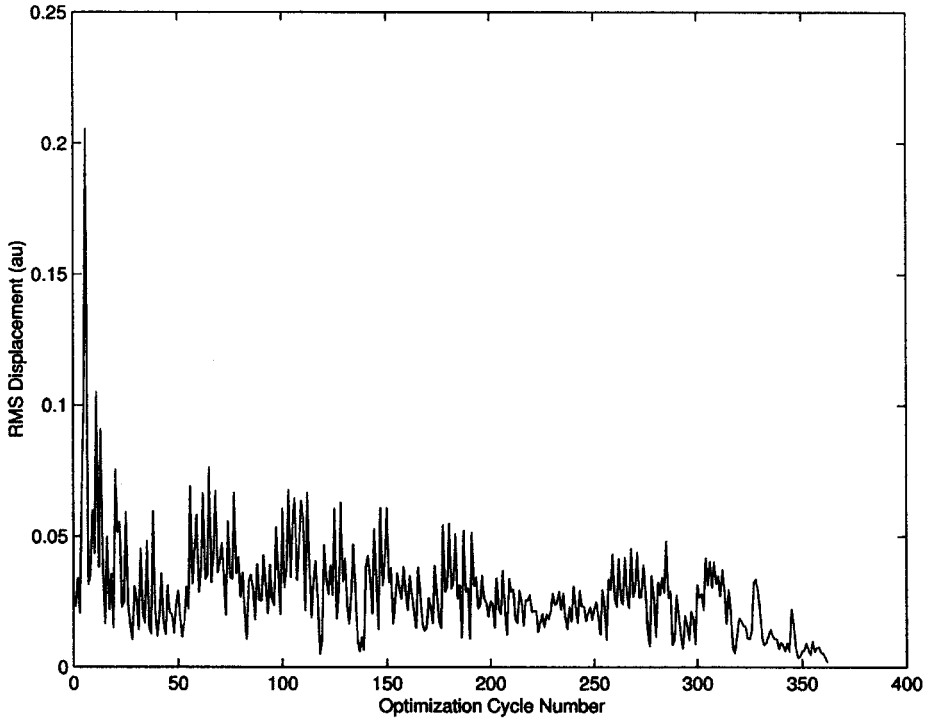

FIGURE 3. RMS internal coordinate displacement (bond lengths in bohrs and angles in radians) is shown as a function of the geometry-optimization step.

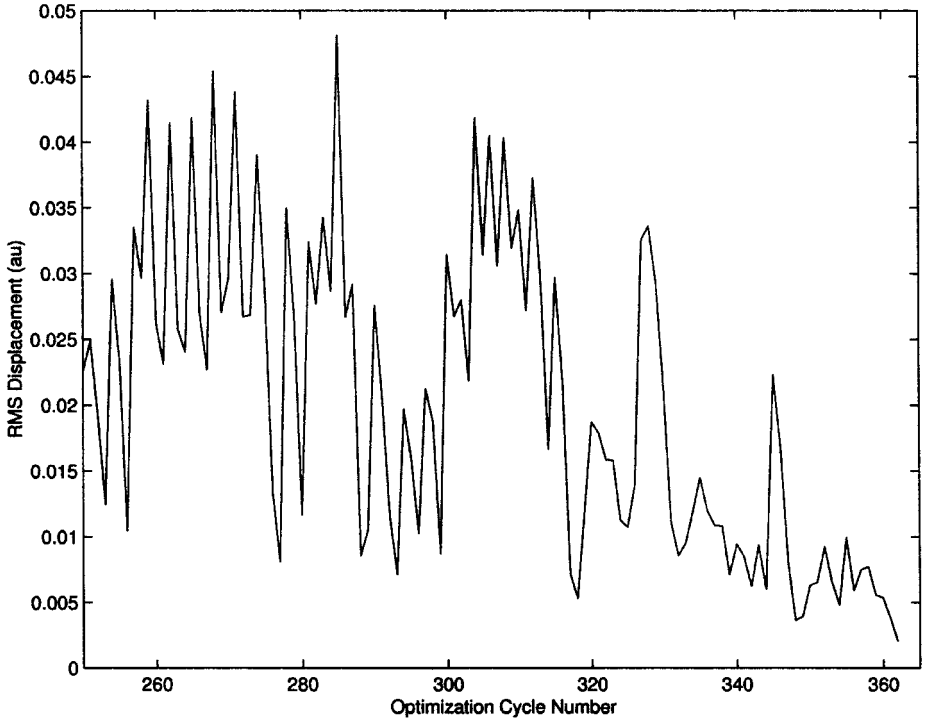

FIGURE 4. RMS internal coordinate displacement (bond lengths in bohrs and angles in radians) is shown as a function of the geometry-optimization step for the final portion of the optimization.

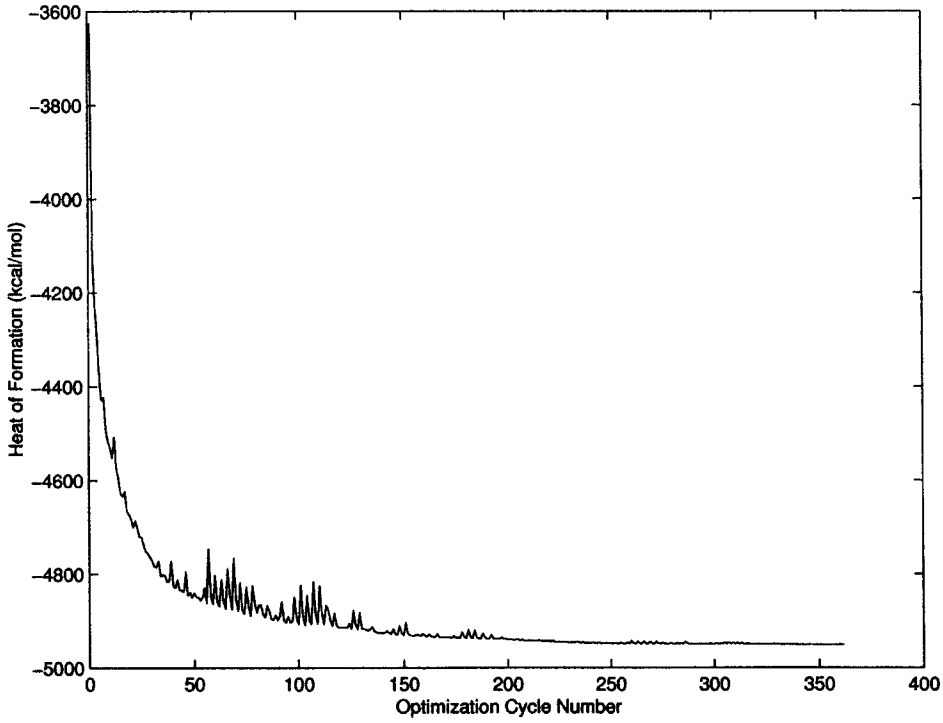

FIGURE 5. Heat of formation (kcal / mol) is shown as a function of the geometry-optimization step.

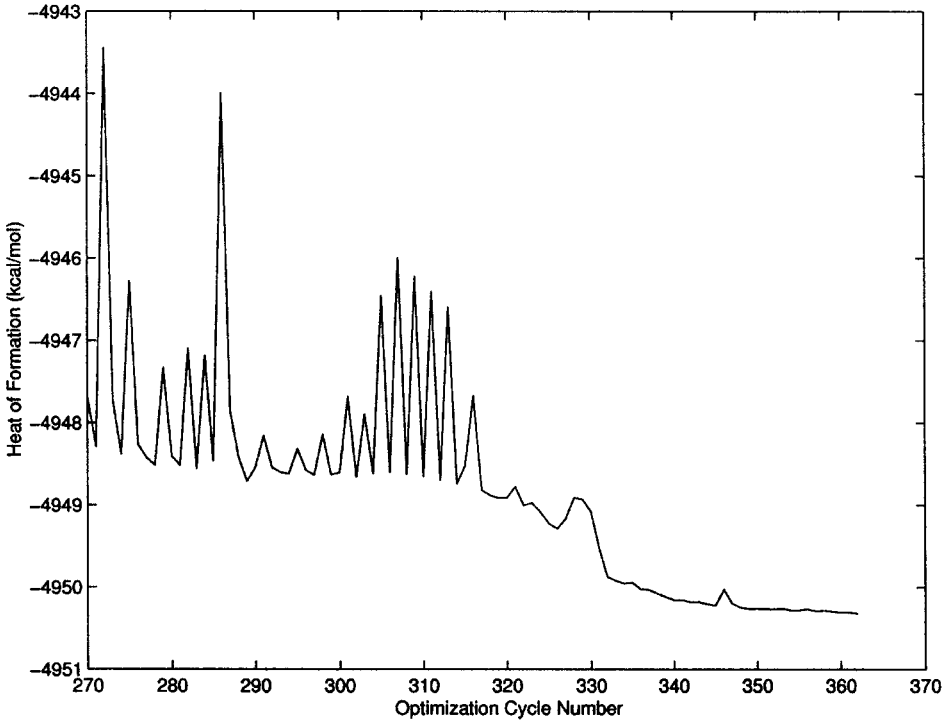

FIGURE 6. Heat of formation (kcal / mol) is shown as a function of the geometry-optimization step for the final portion of the optimization.

update step. As with the forces, the RMS displacements fluctuate wildly with the optimization step. It can also be noted that the displacements decrease very slowly as the optimization progresses. This is explainable by the large number of degrees of freedom in the system. Furthermore, the displacements are quite large throughout the calculation because the potential-energy surface is very flat. Thus, large displacements cause very small changes in energy.

Figures 5 and 6 show the heat of formation as a function of the geometry-update steps. As in [9], the heat of formation quickly decreases during the first few geometry updates. Then, it gradually levels off and decreases slowly in the final iterations. However, the energy does not decrease monotonically, but rather contains spikes as it decreases. This is probably an effect of using a different optimizer than the one utilized in [9]. These spikes do not adversely affect the optimization, for the general trend in the energy is to decrease monotonically if the spikes are discarded. In the last several steps of the optimization, the spikes in energy disappear. Thus, over the last 15 geometry updates, the change in the heat of formation is only -0.07 kcal/mol.

The final, optimized structure has an RMS deviation from the crystal structure of 2.4 Å as measured by *Quanta* [17]. Perhaps the most significant reason for this large deviation in structure is that the PM3 calculation was carried out in the gas phase, whereas the X-ray structure was determined from protein molecules along with several water molecules in a crystalline form. The exclusion of the effects of solvation on the surface of the protein may tend to cause the molecule to expand somewhat.

Several calculations on plasminogen were carried out to optimize the parameters in the geometry-optimization step (such as the maximum step size). These calculations all used the same starting geometry (with an energy 1300 kcal/mol above the energy at the optimized geometry), but contained different optimization parameters. The final geometries of each of these calculations differed from one another slightly. The RMS deviations between their geometries were at most 0.9 Å, and their energies differed by less than 6 kcal/mol. This occurred because the molecule is very floppy, causing the potential-energy surface to be very flat. Thus, each optimization obtained different local minima on the potential-energy surface.

Conclusions

CGDMS in conjunction with the GDIIS/RFO optimizer is an effective tool for semiempirical geometry optimizations of proteins. We found that the number of geometry update steps required by the GDIIS/RFO optimizer is roughly equal to one-third the number of atoms in the system. Thus, with linear-scaling energy and gradient updates, such as PM3/CGDMS, calculations on proteins containing over 1000 atoms are well within the reach of current computational resources. We expect that protein geometry optimizations using semiempirical methods will soon become routine. Also, since proteins are naturally in a solvated environment, it is important to add solvation effects to the calculations. Work along these lines is in progress.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (CHE-9618323) and Gaussian, Inc. One of the authors (A.D.D.) was supported by a training fellowship from the Keck Center for Computational Biology (NLM Grant No. LM07093). We thank Raul Cachan for useful discussions.

References

1. Lewis, J. P.; Carter, C. W., Jr.; Hermans, J.; Pan, W.; Lee, T.-S.; Yang, W. J Am Chem Soc 1998, 120, 5407.
2. Stewart, J. J. P. Int J Comput Chem 1989, 10, 209. Stewart, J. J. P. Int J Comput Chem 1989, 10, 221.
3. Vincent, J. J.; Dixon, S. L.; Merz, K. M., Jr. Theor Chem Acc 1998, 99, 220.
4. Daniels, A. D.; Scuseria, G. E. J Chem Phys 1999, 110, 1321.
5. Daniels, A. D.; Millam, J. M.; Scuseria, G. E. J Chem Phys 1997, 107, 425.
6. Stewart, J. J. P. Int J Quantum Chem 1996, 58, 133.
7. Palser, A. H. R.; Manolopoulos, D. E. Phys Rev B 1998, 58, 12704.
8. Goedecker, S.; Teter, M. Phys Rev B 1995, 51, 9455.
9. Stewart, J. J. P. Theochem 1997, 401, 195.
10. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochter-

ski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 99, Development Version (Revision 0.2); Gaussian, Inc,: Pittsburgh, PA, 1998.

11. Farkas, Ö.; Schlegel, H. B. J Chem Phys, in press.
12. Császár, P.; Pulay, P. Theochem 1984, 114, 31.
13. Farkas, Ö.; Schlegel, H. B. J Chem Phys, in press.
14. Farkas, Ö.; Schlegel, H. B. J Chem Phys 1998, 109, 7100.
15. Hoover, G. J.; Menhart, N.; Martin, A.; Warder, S.; Casellino, F. J. Biochemistry 1993, 32, 10936.
16. GaussView, beta-2.0; Gaussian, Inc.: Pittsburgh, PA, 1998.
17. Quanta97; Molecular Simulations, Inc.: San Diego, CA, 1998.