Solving the two-electron integral problem in Hartree-Fock theory

Matt Challacombe and Eric Schwegler
Department of Chemistry and the Minnesota Supercomputer Institute,
1200 Washington Ave. South
Minneapolis, Minnesota 55415
(June 28, 1996)

First principles electronic structure methods are proving increasingly valuable for the prediction and interpretation of molecular properties. This is because they provide a framework for the systematic improvement of approximations to well defined models. A central model in first principles quantum chemistry is the Hartree-Fock (HF) theory. Conventional implementations of HF theory are limited by computation of the two-electron integrals required to build a Fock matrix, a step which requires CPU time typically scaling as the third power of system size. Here, we introduce and establish a linear scaling method for computation of the Fock matrix that is valid for insulating systems, circumventing the two-electron integral problem. The rewards of a first principles approach to biochemical problems are outlined and an initial application is made to several proteins of current interest.

In molecular orbital theory, the HF approximation provides a reference for methods that include the effects of dynamical electron correlation [1]. Within the density functional theory, HF exchange is a nessesary ingredient in the most accurate functionals [2,3]. Unmodified, HF theory describes the essential physics in systems where the effects of dynamical correlation are small, and generally yields qualitatively correct results.

The expensive component of a HF calculation is construction of the exchange and Coulomb contributions to the Fock matrix. Computation of these matrices entails the calculation of two-electron integrals and their contraction with the density matrix. Because the two-electron integrals are four-index quantities, computation of the Fock matrix is often attributed an $\mathcal{O}(N^4)$ computational complexity, where N is the number of basis functions. In practice, the use of Gaussian basis sets and the direct SCF methods employed by standard quantum chemistry programs yield computational complexities that are $\approx N^3$, and which approach a quadratic limit as N becomes large [4]. This sub-N^4 scaling is achieved by the use of density weighted integral screening, which avoids the computation and handling of insignificant two-electron integrals.

Here, we report linear scaling computation of the Fock matrix for complex, three-dimensional systems. Linear scaling is obtained with highly specialized methods that exploit the unique characteristics of the exchange and Coulomb interactions. Linear scaling computation of the exchange matrix is accomplished with recently introduced [5] thresholding criteria that enforce the locality of exchange interactions. These criteria result from an asymptotic form of the density matrix that is valid for insulating systems, and have been implemented in subroutine ONX. Computation of the exchange matrix with ONX requires thresholding parameters that may easily be determined by small representative calculations.

Recently, a variety of fast hierarchical multipole methods have been reported, which in principle are able to compute the Coulomb matrix in $\mathcal{O}(N)$ cpu time. These methods employ clustering technologies that rely on rapidly convergent series expansions to reduce complexity, and can be classified as either fast multipole methods (FMM) or tree-codes (see [6,7] for a discussion). In addition to hierarchical multipole methods, an exciting Ewald-like partitioning scheme (KWIK) [8] has recently been introduced that is also potentially $\mathcal{O}(N)$. While the fast quantum Coulomb methods that have appeared so far are very promising, until now they have not demonstrated linear scaling for three-dimensional systems or control over errors in converged SCF energies. Here, we report results obtained with an improved version of our quantum chemical tree-code (QCTC) that satisfy these criteria.

The improvements made to QCTC include the use of high-order multipole expansions, multiresolution of the density, linked lists, and an improved near-field algorithm: Large (15'th order) multipole expansions based upon real harmonics are used in the translation and contraction of tensor elements. These expansion methods take advantage of the recursion relations introduced by White and Head-Gordon [9], and employ real arithmetic as discussed by Pérez-Jordá and Yang [10]. As before [6,7], the density is decomposed into a hierarchy of moments corresponding to different levels of spatial discretization, forming a three-dimensional tree. As we previously suggested [6,7], multiresolution of the density is used to overcome penetration effects at fine levels of spatial refinement. A linked list is used to efficiently traverse the tree, which typically involves 12 to 14 tiers with terminating nodes containing less than 50 distributions. Multipole and penetration admissibility criteria are used that depend dynamically on the significance of interacting charge distributions and their separation. Distributions in terminating nodes are sorted, allowing the abortion of near-field integral loops when insignificant distributions are encountered.

Subroutines ONX and QCTC have been interfaced

TABLE I. Results of single point MONDO RHF/3-21G calculations on selected proteins. Structures were obtained from the protein data bank and capped where appropriate to obtain a closed shell system.

Protein	Energy (hartree)	Atoms	Basis Functions	Time/Cycle (hours) Exchange	Time/Cycle (hours) Coulomb
EDP	-8154.72370	255	1461	0.8	0.3
	-8154.72323[a]			0.6[b]	
CRD	-16737.90643	572	3237	2.7	1.1
P53	-17115.30474	698	3836	2.1	1.3

[a]Energy obtained with GAUSSIAN 94 default settings
[b]CPU Time/GAUSSIAN 94 SCF cycle on a Cray C-90

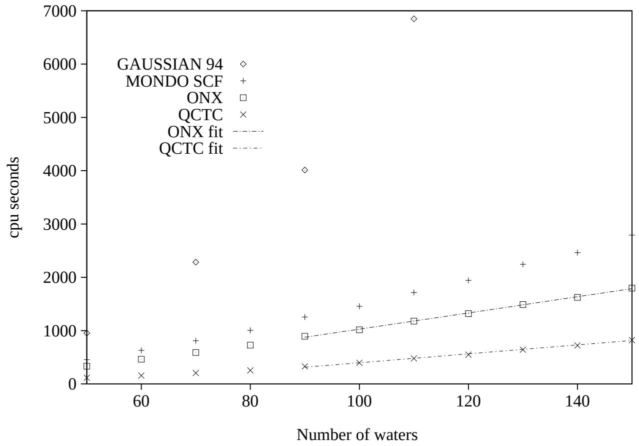

FIG. 1. Average CPU timings for construction of the Fock matrix two-electron components corresponding to a sequence of water clusters. Average CPU times for MONDO and GAUSSIAN 94 SCF cycles are also shown. Fits of $a + bN$ to the Fock matrix two-electron component times are shown as an aid to gauging linearity.

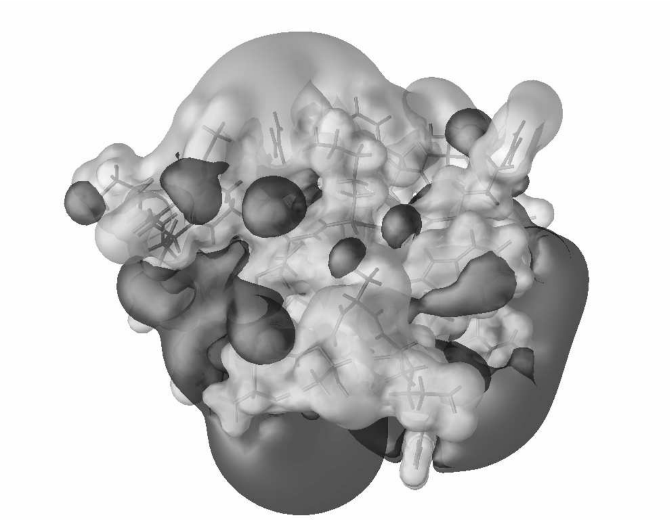

FIG. 2. Iso-surfaces of the CRD electrostatic potential at the RHF/3-21G level of theory. CRD belongs to a family of scorpion toxins that inhibit voltage–gated and Ca^{++} activated K^{+} channels. The light grey surface corresponds to a value of 0.035 au, and the dark to -0.035 au. The bulbous positive surface at the top is associated with the amino acid residues of CRD that mediate the toxin–receptor interaction; from upper left to right they are M29, R34, and K27. Note the localization of negative potential corresponding to lone pair densities, and to the middle right, the negative surface above the ring of W14.

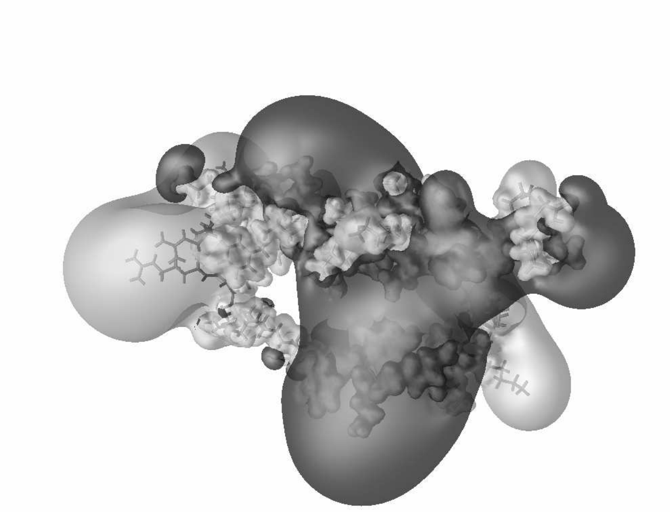

FIG. 3. Iso-surfaces of the P53 tetramerization monomer electrostatic potential at the RHF/3-21G level of theory. Mutations in the P53 tumor suppressor are the most frequently observed genetic alterations in human cancer. The light grey surface corresponds to a value of 0.06 au, and the dark to -0.06 au. The midsection is spanned by a large negative potential lobe transverse to the helix axis. Both the N terminus (upper right) and the bend (lower left) sport peanut shaped lobes of positive potential.

with a modified version of the *ab initio* quantum chemistry code HONDO 95.3 [11] (hereafter referred to as MONDO). We have performed MONDO SCF calculations at the RHF/3-21G level of theory on a series of three-dimensional water clusters taken from an equilibrated molecular dynamics simulation with a density corresponding to standard temperature and pressure. Timings for computation of the exchange and Coulomb matrices are shown in Figure 1, together with average cpu times for one MONDO SCF cycle and one GAUSSIAN 94 [12] SCF cycle. All timings were obtained on an IBM RS6000 model 590. The MONDO timings correspond to diagonalization plus assembly of the two–electron matrices with the ONX(-5) thresholding parameters appropriate for water [5] and the QCTC thresholding parameter `thresh`$=10^{-6}$ [7]. Average MONDO timings were computed by averaging over the last 8 SCF cycles. The average GAUSSIAN 94 SCF timings were obtained with default options and by dividing the total cpu time spent in LINK 502 by the total number of SCF cycles. For both methods, a 10^{-5} convergence criteria in the total energy was used. Errors in the MONDO energies were found to be 0.4 μhartree per atom by comparison to GAUSSIAN 94 calculations employing the tight option. For reference, GAUSSIAN 94 was found to achieve a precision of 0.6 μhartree per atom with default parameters.

Although tree–codes are often attributed an $\mathcal{O}(N \log N)$ complexity, Esselink [13,14] has argued that they are $\mathcal{O}(N)$, while others [15,16] have reasoned that the FMM is in fact $\mathcal{O}(N \log N)$. In practice, a strict adherence to these theoretical complexities can only be observed with ideal (classical, homogeneous, flat or square) systems. More important than postulated asymptotes is the observed behavior with application to real (irregular, quantum, three-dimensional) systems *and* control over errors in the corresponding converged total energies. To examine the adherence of our methods to a linear regime, the 7 largest water cluster timings shown in Figure 1 were fit to $a+bN$. These fits, shown in Figure 1, yield correlation coefficients equal to 0.999, substantiating our claim of linear scaling.

Ab initio quantum chemical methods provide a rigorous framework for the prediction and interpretation of key biomolecular properties such as NMR shielding tensors [17–19], vibrational frequencies [20], etc. While applied to small representative models with increasing frequency and success, the potential of first principles quantum chemistry remains largely untapped due to the unfavorable scaling of conventional methods. To emphasize that this scaling is not an intrinsic feature of the HF theory of insulators, we list breakthrough applications to proteins of current interest in Table 1, including endothelin (EDP) [21], charybdotoxin (CRD) [22], and the tetramerization monomer of P53 [23]. These MONDO SCF calculations were carried out on an IBM RS6000 model 590, using the ONX(-5) thresholding parameters appropriate for peptides [5], the QCTC thresholding parameter `thresh`$=10^{-6}$, and a 10^{-5} total energy convergence criteria. Note that the CPU time required to compute the exchange matrix for P53 is less than that for charybdotoxin, which has fewer atoms. This is because charybdotoxin is a compact globular protein, with a core consisting entirely of six cysteine residues involved in three disulfide bonds.

An important molecular property readily obtained from HF theory is the electrostatic potential, which is widely implicated in molecular recognition, binding, and the enhanced diffusion of charged substrates [24]. Isosurfaces of the electrostatic potential of CRD are shown in Figure 2, and those of P53 in Figure 3. A modified QCTC was used to compute the electrostatic potential, and IBM Visualization Data Explorer [25] was used to compute and render the surfaces.

We have established that Gaussian-based HF theory can achieve an $\mathcal{O}(N)$ computational complexity without loss of precision, and that these linear scaling methods are applicable to large systems of biological relevance with commonly available workstation technology. Although the methods employed here are highly specialized, they are by no means fully developed; significant additional economies are anticipated to result from both code optimization and further theoretical development. Linear scaling methods for computation of the exchange–correlation matrix are also possible [26,27], and in combination [2,3] with the methods presented here, will enable large *ab initio* biomolecular studies to within chemical accuracy.

ACKNOWLEDGMENTS

M.C. recognizes a CISE postdoctoral fellowship from the NSF with matching support from the Minnesota Supercomputer Institute. Computational resources were provided by the University of Minnesota-IBM Shared Research Project.

[1] Mok, D. K. W., Neumann, R. & Handy, N. C. *J. Phys. Chem.* **100**(15), 6225–30 (1996).

[2] Becke, A. D. *J. Chem. Phys.* **98**(7), 5648–52 (1993).

[3] Becke, A. D. *J. Chem. Phys.* **104**(3), 1040–6 (1996).

[4] Strout, D. L. & Scuseria, G. E. *J. Chem. Phys.* **102**(21), 8448–8452 (1995).

[5] Schwegler, E. & Challacombe, M. *J. Chem. Phys.* **105**(7) (1996). in press.

[6] Challacombe, M., Schwegler, E. & Almlöf, J. *Computational Chemistry: Review of Current Trends*, 53–107. World Scientific, Singapore (1996).

[7] Challacombe, M., Schwegler, E. & Almlöf, J. *J. Chem. Phys.* **104**(11), 4685–98 (1996).

[8] Dombroski, J. P., Taylor, S. W. & Gill, P. M. W. *J. Phys. Chem.* **100**(15), 6272–76 (1996).
[9] White, C. A. & Head-Gordon, M. *J. Chem. Phys.* **101**(8), 6593–605 (1994).
[10] Pérez-Jordá, J. M. & Yang, W. *J. Chem. Phys.* **104**(20), 8003–6 (1996).
[11] Dupuis, M., Marquez, A. & Davidson, E. R. *HONDO 95.3 from CHEM-Station*. IBM Corporation, Neighborhood Road, Kingston, NY. 12401, (1995).
[12] Frisch, M. J. *et al. Gaussian 94, Revision D.1.* Gaussian Inc., Pittsburgh PA, (1995).
[13] Esselink, K. *Inf. Proc. Lett.* **41**, 141–7 (1992).
[14] Esselink, K. *Comput. Phys. Comm.* **87**, 375–95 (1995).
[15] Aluru, S., Prabhu, G. M. & Gustafson, J. in *Proceedings SUPERCOMPUTING '94*, 420–6 (IEEE Computer Society Press, Los Alamitos, CA, 1995).
[16] Aluru, S. *J. Sci. Comp.* **17**(3), 773–6 (1996).
[17] Rauhut, G., Puyear, S., Wolinski, K. & Pulay, P. *J. Phys. Chem.* **100**(15), 6310–6 (1996).
[18] Szilagyi, L. *Prog. NMR Spect.* **27**, 325–443 (1995).
[19] De Dios, A. C. & Oldfield, E. *J. Am. Chem. Soc.* **116**(12), 5307–14 (1994).
[20] Kozlowski, P. M., Jarzecki, A. A. & Pulay, P. *J. Phys. Chem.* **100**(17), 7007–13 (1996).
[21] Andersen, N. H., Chen, C., Marshner, T. M., Krystek, S. R. & Bassolino, D. A. *Biochemistry* **31**, 1280 (1992).
[22] Bontems, F., Gilquin, B., Roumestand, C., Menez, A. & Toma, F. *Biochemistry* **31**, 7756 (1992).
[23] Clore, G. M. *et al. Nat. Struct. Biol.* **2**, 321 (1995).
[24] Honig, B. & Nicholls, A. *Science* **268**, 1144–9 (1995).
[25] IBM Corporation, NY 10577. *IBM Visualization Data Explorer Version 3.1*, (1995). DX chemistry extensions were used and are freely available from the Cornell Theory Center DX repository.
[26] Johnson, B. PhD thesis, Carnegie Mellon University, December (1993). Chapter 6.
[27] Pérez-Jordá, J. M. & Yang, W. *Chem. Phys. Lett.* **241**, 469–76 (1995).
[28] Goldstein, S. A., Pheasant, D. J. & Miller, C. *Neuron* **12**, 1377–88 (1994).