JCTC Journal of Chemical Theory and Computation

# Validation of Linear Scaling Semiempirical LocalSCF Method

Victor M. Anisimov,*[,†] Vladislav L. Bugaenko,[‡] and Vladimir V. Bobrikov[‡]

*Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland,
20 Penn Street, Baltimore, Maryland 21201, and Quantum Biochemistry Group,
Konstantina Fedina-3/24, 105215 Moscow, Russian Federation*

Received July 5, 2006

**Abstract:** The numerical accuracy of linear scaling semiempirical methods LocalSCF and MOZYME is analyzed in comparison to conventional matrix diagonalization with respect to a variety of molecular properties including conformational energy, dipole moment, atomic charges, and bond orders. Major semiempirical MNDO, AM1, PM3, and PM5 Hamiltonians were considered in the study. As the numerical tests demonstrate, both LocalSCF and MOZYME reasonably reproduce matrix diagonalization results with the deviations being below the accuracy of semiempirical methods. However, the economical LocalSCF memory consumption and faster calculations are more beneficial for the quantum-mechanical modeling of large biological systems. The computational performance of the LocalSCF method is tested on the conformational energy calculation of a series of molecular dynamics snapshots of insulin in a large box of water.

## 1. Introduction

The linear scaling LocalSCF method was proposed recently as a computationally inexpensive alternative to matrix diagonalization for application to large biomolecular systems.[1] The purpose of the method is to bypass the prohibitive quadratic scaling of computer memory and cubic scaling of computer time by the size of the molecular system of the conventional diagonalization procedure, thereby allowing application to real-size biological systems at the quantum-mechanical level. Currently implemented in a semiempirical framework,[2] the LocalSCF method allows conducting molecular orbital computer simulations of hundreds of thousands of atoms on a personal computer. In LocalSCF, the linear scaling regimen is obtained with the help of the variational finite localized molecular orbital (VFL) approximation.[1] The VFL seeks a solution of the self-consistent field (SCF) task in the reduced space of compact or localized molecular orbitals (LMOs). For any LMO expansion fixed in the beginning of the SCF calculation, the resulting density matrix and the total energy are the closest possible solution to the matrix diagonalization because of the variational principle under the constraint of the reduced LMO expansion.

Solving the SCF task using the VFL approximation does not guarantee a high-quality wave function because the initially selected fixed LMO expansion may not be optimal for the particular system of interest. Therefore, some kind of orbital expansion procedure is needed in order to account for the individual aspects of the molecular structure. The combination of the VFL method with a LMO expansion procedure creates the linear scaling LocalSCF method. LocalSCF calculations are conducted in the following way. After the SCF convergence is achieved for the initial-guess LMOs, an orbital expansion is undertaken on the basis of an expansion algorithm. This algorithm checks the atomic centers adjacent to each LMO and estimates the energy gain upon expanding the LMO on the particular center. If the expected energy gain is greater than a threshold value, the expansion is made; otherwise, the center is omitted, and another atom in the neighborhood is tested in a similar way. After the expansion step is made, the SCF refinement of linear coefficients of LMOs begins. The cycles of SCF and orbital expansion are repeated until the total energy is converged.

Another LMO-based linear scaling semiempirical method we consider in our study is MOZYME, which was developed

[*] Corresponding author e-mail: victor@outerbanks.umaryland.edu.
[†] University of Maryland.
[‡] Quantum Biochemistry Group, Konstantina Fedina-3/24.

10.1021/ct600222t CCC: $33.50
Published on Web 10/05/2006

by Stewart.[3] Both LocalSCF and MOZYME operate with LMOs, as their common ground. Despite this similarity, these methods are based on different principles. The atomic expansion of MOZYME LMOs is the integral part of the SCF procedure. MOZYME SCF performs occupied-virtual-orbital Jacobi rotations to achieve block diagonalization of the Fock matrix. The rotation has the purpose of orthogonalization of the occupied set of LMOs to the corresponding virtual set. Each orbital rotation combines expansion sets of occupied and virtual LMOs involved in the rotation, and as a result, the LMO size uncontrollably grows during the SCF iterations until the LMOs reach about 150 atomic centers on average. After this limit is achieved, adding new atomic centers to the LMO expansion does not change the density matrix, and such atomic centers with negligibly small linear coefficients can be effectively removed from the LMOs. For molecules consisting of several hundreds of atoms and larger, the MOZYME LMOs are considerably shorter than the canonical molecular orbitals, thereby providing a significant savings in computer resources. In contrast to MOZYME, the VFL-based SCF calculation in LocalSCF is conducted on fixed-size LMOs and the LMO expansion procedure is independent from SCF. This provides a free hand to control the LMO size, and the user may choose shorter or longer LMOs depending on the speed and desired degree of agreement with the matrix diagonalization result. A distant analogy can be portrayed between VFL and the Roothan−Hall approximation.[4,5] If the Roothaan−Hall method seeks a wave function in the finite space of atomic orbitals, the VFL approximation utilizes additional degrees of freedom by constructing a wave function in the reduced space of molecular orbital expansion. A corresponding similarity may be observed between the selection of a particular basis set and the selection of the LMO size. In both cases, there is room for a rational human choice, and both solutions approach variationally the target function under elimination of the corresponding constraints. Accordingly, the VFL solution of SCF equations approaches the conventional matrix diagonalization limit when localized molecular orbitals are expanded to the size of conventional MOs. In practice, the total energy and molecular wave function converge well before the LMO expansion reaches the conventional MO limit. In most cases, the LMO expansion reaches 30 atomic centers on average. This helps to save significant computational resources when dealing with large biological molecules. Note that the wave function constructed from the short LMO expansion is a variational approximation to the wave function constructed from fully delocalized canonical molecular orbitals, and as such, the LocalSCF wave function may be reasonably good even if a very short LMO expansion is employed. The practical outcomes of the LocalSCF approximation will be studied in this work in comparison to the matrix diagonalization and the linear scaling method MOZYME. The analysis will be limited to the neglect of diatomic differential overlap (NDDO)-based semiempirical methods LocalSCF and MOZYME operating with LMOs. For the description of other linear scaling methods, readers are referred to the original articles.[6−10]

## 2. Methods

A small protein pro-insulin (PDB accession code 1EFE) containing 60 amino acids was downloaded from the Protein Data Bank.[11] Amino acid protonation was assigned on the basis of the assumption of physiological pH, and the total charge of +1 was neutralized by adding a chloride counterion. The cleaned structure was placed in an orthorhombic box of explicit TIP3P water with the solvent extending at least 10 Å from the solute molecule in each direction. The structure was equilibrated by a CHARMM force field[12] under a periodic boundary condition. A 200 ps molecular dynamics (MD) NPT simulation was undertaken at 300 K with the Leapfrog integrator and with a 2 fs time step. Additionally, the Langevin piston pressure control[13] was applied; covalent bonds to hydrogen atoms were restrained by the SHAKE algorithm,[14] and the particle mesh Ewald algorithm was used to treat the long-range electrostatics.[15] The first 10 ps of the simulation time were considered as an equilibration, and the other 190 ps were treated as a productive run out of which 20 snapshots were extracted with a 10 ps time interval for further analysis by quantum-mechanical methods. These 20 structures, each containing 20 058 atoms, were further processed by removal of the chloride counterion, leading to the molecular charge +1, and by the removal of water molecules, preserving only the 100 closest ones to the ionized amino acids. This resulted in 20 insulin conformations, each containing 1247 atoms. This size of the system was selected to allow performing matrix diagonalization calculations which fit to 1 GB of computer memory.

LocalSCF and MOZYME calculations were performed in regular and fast modes with the keyword PRECISE set on. Regular mode calculations were applied to the insulin snapshots containing 1247 atoms. The LocalSCF SCF procedure was performed by the steepest descent gradient optimization of linear coefficients of molecular orbitals, where the derivative of total energy versus the change in linear coefficients was utilized. In the regular mode LMO refining, SCF was converged to either 0.0005 eV for the gradient norm or to a value of 0.002 eV for the maximum component of the gradient for the linear coefficients of LMOs or to 0.0001 kcal/mol total energy change, whichever came first. The LMO expansion gradient threshold was set to 0.04 eV, and the expansions were performed until the total energy change was smaller than 0.5 kcal/mol between the expansions. The root-mean-square (RMS) value for LMO nonorthogonality was limited to 0.001. MOZYME regular mode calculations were conducted with the cutoff for NDDO approximation being set to 12 Å; SCF convergence criteria were set to 0.01 kcal/mol. Matrix diagonalization SCF was conducted until the total energy was converged with the 0.0001 kcal/mol threshold. Hereafter, we will use the name MOPAC as a synonym of matrix diagonalization.

LocalSCF and MOZYME in fast calculation modes were also applied to the large 20 058-atom box of insulin in water. Here, the MOZYME cutoff for NDDO approximation was reduced to 6 Å. In case of LocalSCF, the fast multipole method (FMM)[16,17] for the calculation of Coulomb integrals was turned on. The well-separation parameter was set to 2. The near field was set to 10 Å with a cube edge of 5 Å. The

***Table 1.*** AM1 Total Energy, Dipole Moment, and Conformational Energies of Insulin Conformations Containing 1247 Atoms.

| | | | | | | | conformational energy[c] | | |
|---|---|---|---|---|---|---|---|---|---|
| conf. | LSCFR[a] | MOZ12[a] | MOP[a] | LSCFR[b] | MOZ12[b] | MOP[b] | LSCFR | MOZ12 | MOP |
| 1 | −7859.73 | −7866.73 | −7866.30 | 253.33 | 253.34 | 253.28 | | | |
| 2 | −7808.01 | −7815.22 | −7814.45 | 259.15 | 259.14 | 259.10 | 51.72 | 51.52 | 51.86 |
| 3 | −7883.81 | −7891.53 | −7890.54 | 250.09 | 249.99 | 250.00 | −75.80 | −76.31 | −76.10 |
| 4 | −7838.84 | −7846.53 | −7845.51 | 224.07 | 224.02 | 224.02 | 44.97 | 44.99 | 45.04 |
| 5 | −7777.99 | −7786.06 | −7784.70 | 259.18 | 259.24 | 259.13 | 60.85 | 60.48 | 60.81 |
| 6 | −7842.35 | −7849.25 | −7849.19 | 211.81 | 211.76 | 211.75 | −64.36 | −63.20 | −64.49 |
| 7 | −7900.92 | −7908.24 | −7907.94 | 207.32 | 207.18 | 207.22 | −58.57 | −58.99 | −58.75 |
| 8 | −7871.47 | −7879.05 | −7878.29 | 247.39 | 247.44 | 247.32 | 29.45 | 29.20 | 29.65 |
| 9 | −7864.70 | −7873.22 | −7871.67 | 275.46 | 275.45 | 275.36 | 6.78 | 5.82 | 6.62 |
| 10 | −7889.75 | −7897.31 | −7896.69 | 225.83 | 225.80 | 225.74 | −25.06 | −24.08 | −25.02 |
| 11 | −7944.17 | −7953.49 | −7952.30 | 234.92 | 235.00 | 234.81 | −54.42 | −56.18 | −55.62 |
| 12 | −7878.29 | −7886.17 | −7885.26 | 202.73 | 202.75 | 202.67 | 65.89 | 67.32 | 67.05 |
| 13 | −7895.03 | −7902.26 | −7902.03 | 194.59 | 194.55 | 194.51 | −16.74 | −16.09 | −16.78 |
| 14 | −7860.07 | −7867.09 | −7867.11 | 155.62 | 155.62 | 155.58 | 34.97 | 35.17 | 34.93 |
| 15 | −7982.52 | −7988.87 | −7989.38 | 145.24 | 145.30 | 145.19 | −122.46 | −121.78 | −122.27 |
| 16 | −7970.14 | −7977.65 | −7976.99 | 151.29 | 151.29 | 151.23 | 12.39 | 11.22 | 12.39 |
| 17 | −7903.43 | −7911.07 | −7910.33 | 174.45 | 174.51 | 174.40 | 66.71 | 66.58 | 66.66 |
| 18 | −7848.29 | −7856.44 | −7855.14 | 161.91 | 161.99 | 161.83 | 55.14 | 54.63 | 55.20 |
| 19 | −8005.24 | −8013.12 | −8012.40 | 197.49 | 197.50 | 197.42 | −156.95 | −156.68 | −157.26 |
| 20 | −7960.22 | −7968.14 | −7967.33 | 138.38 | 138.49 | 138.34 | 45.03 | 44.97 | 45.07 |
| RMS | 6.94 | 0.85 | | 0.07 | 0.09 | | 0.41 | 0.60 | |

[a] Total energy, kcal/mol; LocalSCF (LSCFR), MOZYME (MOZ12), MOPAC (MOP); LocalSCF FMM is off; MOZYME cutoff for NDDO approximation is 12 Å. [b] Total dipole, Debye. The dipole moment error is the module of the error vector $|\bar{\mu}_{\mathrm{MOPAC}} - \bar{\mu}_{\mathrm{test}}|$ where the test represents LocalSCF or MOZYME dipole vectors. [c] Conformational energy, kcal/mol. $E_i^{\mathrm{conf}} = E_i^{\mathrm{tot}} - E_{i-1}^{\mathrm{tot}}$.

series for the electrostatic potential was truncated at the fourth term, and the *B* and *C* operators were truncated at the seventh and third terms, respectively.

MNDO[18], AM1[19], PM3[20], and PM5[21] semiempirical Hamiltonians were used in this work. MOZYME and matrix diagonalization calculations were performed by using the MOPAC2002 program package.[21] LocalSCF results were collected by using the LocalSCF computer program.[2] All calculations were performed on a single CPU personal computer equipped with an Intel Pentium-4 3.0 GHz processor with 1 GB of random access memory under the Microsoft Windows XP operating system.

## 3. Results and Discussion

**3.1. Regular Mode Calculation of Small Protein.** LocalSCF, MOZYME, and MOPAC calculations were performed on the 20 insulin conformations, each consisting of 1247 atoms. From the very beginning, matrix diagonalization calculations were unsuccessful because of the SCF convergence problem. The convergence was especially problematic on a dry protein. Preserving water molecules near ionized amino acids somewhat helped to improve the SCF convergence, but several conformations still remained problematic. A solution was found to feed the LocalSCF density matrix as an initial guess to the matrix diagonalization calculation. After this, MOPAC calculations were converging quickly in a few successive iterations, thereby confirming the good quality of the LocalSCF density matrix.

Details of the computational results for each of the 20 insulin conformations for energy and dipole moment are shown in Table 1 for the AM1 Hamiltonian. Synchronous energy changes on the path from the first to the last conformation indicate that the protein structure is in fact in the process of equilibration. This conclusion is also supported by a synchronous change in the dipole moment. The incomplete equilibration is acceptable in this case because the purpose of the classical MD simulation was to generate realistic protein conformations for the subsequent comparison of linear scaling methods of interest with matrix diagonalization. As the data in Table 1 show, the absolute RMS deviations of the conformational energy in reproducing matrix diagonalization results are 0.41 and 0.60 kcal/mol or 0.9% and 3.8% for LocalSCF and MOZYME, respectively. RMS deviations for the dipole moment are 0.07 and 0.09 D for LocalSCF and MOZYME, respectively; both are in good agreement with matrix diagonalization. Similar studies were performed using other semiempirical Hamiltonians, and the results are presented in Table 2. Additional data include partial atomic charges, bond orders, and geometry gradients. The RMS differences for scalar properties (energy, partial atomic charges, and bond orders) are obtained from eq 1, where index *i* runs over *N* values of a property and index *c* runs over *M* protein conformations. The RMS differences for vector values are calculated according to eq 2. Insulin conformation number 1 is shown in Figure 1, illustrating the water placement in the vicinity of ionized amino acids.

The data collected in Tables 1 and 2 demonstrate that

$$\mathrm{RMS} = \sqrt{\frac{\sum_{c}^{M}\sum_{i}^{N}(V_{ic}^{\mathrm{MOPAC}} - V_{ic}^{\mathrm{test}})^2}{MN}} \qquad (1)$$

***Table 2.*** LocalSCF (LSCFR)[a] and MOZYME (MOZ12)[b] RMS Differences for 20 Insulin Conformations Containing 1247 Atoms in Comparison with Matrix Diagonalization in Their Regular Mode Settings

| | MNDO | | AM1 | | PM3 | | PM5 | |
|---|---|---|---|---|---|---|---|---|
| | LSCFR | MOZ12 | LSCFR | MOZ12 | LSCFR | MOZ12 | LSCFR | MOZ12 |
| $E_{\mathrm{tot}}$[c] | 3.50 | 0.61 | 6.94 | 0.85 | 8.22 | 0.45 | 10.50 | 1.98 |
| $E_{\mathrm{conf}}$[c] | 0.34 | 0.62 | 0.41 | 0.60 | 0.29 | 0.57 | 0.20 | 0.56 |
| $\mu$[d] | 0.04 | 0.10 | 0.07 | 0.09 | 0.09 | 0.09 | 0.12 | 0.07 |
| $Q_{\mathrm{non-hydrogen}}$[e] | 0.0002 | 0.0005 | 0.0002 | 0.0005 | 0.0003 | 0.0006 | 0.0002 | 0.0007 |
| $Q_{\mathrm{hydrogen}}$[e] | 0.0001 | 0.0003 | 0.0001 | 0.0003 | 0.0001 | 0.0003 | 0.0001 | 0.0004 |
| $\mathrm{BO}_{AB}$[f] | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0000 | 0.0001 | 0.0001 |
| $\nabla_{\mathrm{non-hydrogen},\%}$[g] | 0.36 | 0.61 | 0.42 | 0.60 | 0.63 | 0.60 | 0.85 | 1.09 |
| $\nabla_{\mathrm{hydrogen},\ \%}$[g] | 0.91 | 0.87 | 1.08 | 0.89 | 1.47 | 0.96 | 1.29 | 1.06 |
| CPU time, sec[h] | 150 | 75 | 170 | 115 | 156 | 128 | 160 | 120 |

[a] FMM is off. [b] Cutoff for NDDO approximation is 12 Å. [c] Total and conformational energy differences, kcal/mol. [d] Dipole moment differences, Debye. [e] Partial atomic charge differences, electron units. [f] Bond orders; only values larger than $10^{-6}$ were taken into analysis. [g] Geometry gradient differences; percent to absolute value. [h] Average calculation time.

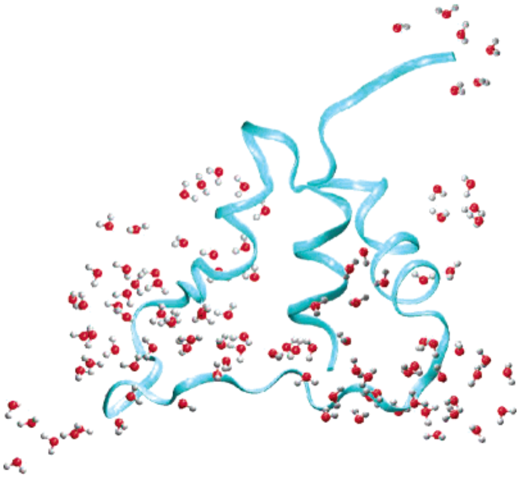

***Figure 1.*** Insulin conformation #1 solvated by 100 molecules of water (1247 atoms total), which are located in the vicinity of the ionized amino acids, creating the total charge of +1.

$$\mathrm{RMS} = \sqrt{\frac{\sum_{c}^{M}\sum_{i}^{N}|\bar{V}_{ic}^{\mathrm{MOPAC}} - \bar{V}_{ic}^{\mathrm{test}}|^{2}}{MN}} \tag{2}$$

LocalSCF systematically underestimates the total energy, whereas MOZYME shows a good agreement with the matrix diagonalization. The smallest LocalSCF RMS difference of 3.50 kcal/mol is observed for the MNDO Hamiltonian. The largest difference of 10.50 kcal/mol is observed for the PM5 Hamiltonian, with AM1 and PM3 results being in the middle. This difference in total energy between LocalSCF and MOZYME is attributed to the difference in size of LocalSCF and MOZYME LMOs, which are 30 and 150, respectively. Supposedly, obtaining the energy limit requires adopting larger LMOs by LocalSCF; however, shorter LMOs are beneficial from a performance perspective. To find out whether LocalSCF is able to produce meaningful computational results with short LMOs, other molecular properties besides total energy have to be studied as well. The accuracy in the prediction of conformational energy is one such example. Here, the situation is quite different, with LocalSCF energies being systematically more accurate than the MOZYME ones. This might be because LocalSCF LMOs are better converged. Indeed, because LocalSCF solves the task of linear scalability variationally, the produced wave function is the closest possible approximation to the target wave function under the given constraint of LMO size, whereas MOZYME LMO tail oscillations are harder to control. The ability to accurately reproduce conformational energies is extremely important when we think about the prospective performance of the molecular dynamics simulation of proteins at the semiempirical level. Overall, both methods show good agreement with the matrix diagonalization on conformational energies.

Another important aspect of comparison is to study how well linear scaling methods reproduce electrostatic properties and, namely, the electric dipole moment and partial atomic charges, in comparison to the matrix diagonalization. Because of the long-range nature of electrostatic forces, their accurate description is especially critical for large biological systems. In our analysis, the dipole moment differences are calculated in the vector form using eq 2. This approach is shown to be more sensitive to the errors in calculations of the dipole moment.[22] Because proteins have large dipole moments, even a 1% error in orientation of the dipole vector may have serious implications. Because insulin has a charge of +1, the center of mass was taken as the origin for the dipole moment calculation. In this study, both LocalSCF and MOZYME show very good agreement with the matrix diagonalization. LocalSCF is slightly better for the MNDO, AM1, and PM3 Hamiltonians, whereas MOZYME shows better results for the PM5 Hamiltonian. The situation with partial atomic charges is also good, although LocalSCF is in a little better agreement with the matrix diagonalization. In turn, MOZYME is slightly better on the prediction of bond orders. In all of these cases, the errors are well below the accuracy of semiempirical methods.

The next property presented in Table 2 is geometry gradients. Their accurate representation by linear scaling methods is especially important, keeping in mind forthcoming semiempirical MD simulations of proteins. Because the absolute value of the gradient strongly depends on the geometry of studied molecules, we perform an error analysis in relative units of percent value. The differences are calculated according to eq 3 in the vector form

$$\mathrm{RMS} = \sqrt{\frac{\sum_{c}^{M}\sum_{i}^{N}|\bar{V}_{ic}^{\mathrm{MOPAC}} - \bar{V}_{ic}^{\mathrm{test}}|^2/|\bar{V}_{ic}^{\mathrm{MOPAC}}|^2}{MN}} \qquad (3)$$

Here, LocalSCF shows more accurate geometry gradients than MOZYME for heavy atoms for each Hamiltonian, except PM3. For hydrogen atoms, MOZYME gradients are systematically more accurate than LocalSCF for all Hamiltonians. In both of the cases, the LocalSCF and MOZYME gradient errors are below 2% of the corresponding target values, which is a good indication of reliability of both linear scaling methods.

The last factor considered in this comparison is CPU time. According to Table 2, the LocalSCF program using its default settings is slower than MOZYME by a factor of 1.2−2.0. This is because LocalSCF is optimized for large systems and is, in part, due to difficulties of making two computer programs run at the same level of accuracy, with the LocalSCF program settings being tuned up for a better agreement with matrix diagonalization and for a maximal savings of computer memory. Moreover, the performance comparison conducted on a small system may not be representative of the real performance when applied to a larger system. Because calculation time is critical for the calculation of large molecules, the faster program modes and larger systems will be considered further in this study.

To complete our analysis of the data presented in Table 2, we consider how well LocalSCF and MOZYME work with particular Hamiltonians. In MOZYME calculations, the MNDO, AM1, and PM3 Hamiltonians score equally with slightly higher errors obtained for PM5. The difference between the Hamiltonians is sharper in the case of LocalSCF. Here, the least errors are observed for the MNDO Hamiltonian, followed by AM1, then PM3, and at last, PM5. Because the AM1 Hamiltonian is considered to be of primary interest among other semiempirical Hamiltonians for biological applications,[22−26] it is especially important that our results confirm that AM1 does not introduce any particular complications for the linear scaling regimen. Regarding the PM5 Hamiltonian, both LocalSCF and MOZYME results indicate that PM5 makes more difficult in comparison to other Hamiltonians the task of achieving the linear scaling regimen.

**3.2. Fast Mode Calculation of Small Proteins.** In the above tests, LocalSCF calculations were performed with Coulomb integrals treated explicitly. However, the straightforward calculation time of Coulomb integrals scales quadratically with the number of atoms. This becomes unacceptable for larger systems. This problem is addressed in the LocalSCF program with the help of the fast multipole method. The next series of LocalSCF and MOZYME calculations on the insulin snapshots consisting of 1247 atoms are performed in the fast program settings. LocalSCF calculations were performed with the FMM option turned on. The comparable faster mode of MOZYME is obtained by setting the NDDO cutoff to 6 Å. FMM is not available in MOPAC2002 program. Because the studied system is still the same as the one used to collect data for Tables 1 and 2, the existing matrix diagonalization data were used as reference data here as well. The obtained results are summarized in Table 3.

As in the previous tests, the total energy is underestimated by LocalSCF; however, all of the relative properties are well-reproduced. MOZYME energy differences in the fast mode also grew larger. Again, both linear scaling methods show the largest differences with the matrix diagonalization when the PM5 Hamiltonian is employed. Here, the difference is larger for MOZYME, which is unusual. Similar to the previous test, the LocalSCF conformational energies are in better agreement with matrix diagonalization than the ones calculated by MOZYME for all of the Hamiltonians, including PM5. LocalSCF dipole moments are more accurate for all Hamiltonians except PM5, where MOZYME is a little more accurate. Partial atomic charge differences are smaller for LocalSCF. Bond orders are equally good for both methods. A comparison of the data shown in Tables 2 and 3 indicates that, with the exception of the PM5 Hamiltonian, LocalSCF in its fast mode (Table 3) still provides more accurate conformational energies, dipole moments, and partial charges than MOZYME in its regular mode (Table 2). For the programs in their fast mode, the gradients are systematically more accurate in MOZYME with LocalSCF errors twice as large on average. The largest RMS differences in LocalSCF gradients are observed for the PM5 Hamiltonian.

In contrast to LocalSCF, where regular and fast modes are different by turning off and on the FMM mode, the MOZYME program fast mode employs a reduced criterion for NDDO approximation (6 vs 12 Å), and therefore, the increase of MOZYME gradient errors comes entirely from the SCF part of the calculation. This allows us to compare LocalSCF gradient errors from Table 2 to MOZYME gradient errors from Table 3, because both are coming from the corresponding linear scaling replacement of the matrix diagonalization. As the data show, the gradient errors due to the MOZYME alternative to diagonalization in the fast mode (Table 3) are larger than the LocalSCF errors (Table 2). The errors are also subject to computer implementation of the algorithms, the factor which complicates the analysis of the differences. Overall, the errors due to LocalSCF and MOZYME linear scaling solutions are relatively small. Our work to reduce gradient errors is in progress and will be reported in a subsequent publication. As in the case of the default program settings (Table 2), LocalSCF in fast program settings is slower than the corresponding fast mode of MOZYME (Table 3); however, the LocalSCF program in the fast mode is in better agreement with matrix diagonalization for conformational energies, dipole moments, and partial atomic charges than MOZYME in default (accurate) settings for all Hamiltonians except PM5.

**3.3. Fast Mode Calculation of Protein in a Water Box.** The hydrated insulin samples calculated above consisted of 1247 atoms each. These are quite small systems in comparison to the number of atoms normally treated in biomolecular simulations by modern macromolecular force fields. It might be expected that, for such a relatively small system of 1000 atoms, many linear scaling methods may perform equally well. Indeed, LocalSCF and MOZYME in the above

***Table 3.*** LocalSCF (LSCFF)[a] and MOZYME (MOZ6)[b] RMS Differences for 20 Insulin Conformations Containing 1247 Atoms in Comparison with Matrix Diagonalization in Their Fast Mode Settings

| | MNDO | | AM1 | | PM3 | | PM5 | |
|---|---|---|---|---|---|---|---|---|
| | LSCFF | MOZ6 | LSCFF | MOZ6 | LSCFF | MOZ6 | LSCFF | MOZ6 |
| $E_{tot}$ | 4.39 | 4.15 | 7.73 | 2.37 | 8.25 | 1.14 | 18.35 | 23.71 |
| $E_{conf}$ | 0.37 | 1.39 | 0.55 | 1.38 | 0.37 | 1.23 | 1.40 | 1.74 |
| $m$ | 0.03 | 0.17 | 0.06 | 0.16 | 0.08 | 0.15 | 0.12 | 0.11 |
| $Q_{\text{non-hydrogen}}$ | 0.0005 | 0.0014 | 0.0005 | 0.0016 | 0.0006 | 0.0020 | 0.0019 | 0.0021 |
| $Q_{\text{hydrogen}}$ | 0.0002 | 0.0008 | 0.0002 | 0.0009 | 0.0002 | 0.0009 | 0.0007 | 0.0009 |
| $BO_{AB}$ | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| $\nabla_{\text{non-hydrogen},\%}$ | 3.49 | 1.56 | 3.73 | 1.64 | 4.03 | 1.59 | 7.73 | 3.09 |
| $\nabla_{\text{hydrogen},\%}$ | 4.81 | 2.51 | 6.87 | 2.59 | 4.44 | 2.64 | 9.03 | 3.36 |
| CPU time, sec | 102 | 49 | 104 | 60 | 111 | 75 | 111 | 68 |

[a] FMM turned on. [b] Cutoff for NDDO approximation is 6 Å.

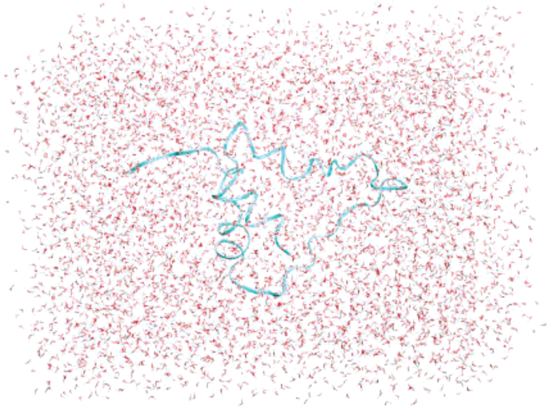

***Figure 2.*** Snapshot #1 of insulin in a water box (20 058 atoms) with total charge 0 neutralized by chloride counterions.

tests showed similar performances and comparable levels of agreement with matrix diagonalization. To see a difference between these two methods by CPU time and memory consumption, one should consider a bigger system for the tests. One of the practical challenges driving the development of linear scaling QM methods is to enable realistic calculations of large biological systems in their physiological environment, for example, quantum-mechanical simulations of biological macromolecules in the condensed phase. Although this goal is not yet quite achieved, it would be interesting to compare the performance of the LocalSCF and MOZYME methods on the insulin placed in a large water box, which we used in a classical MD simulation for the purpose of the generation of insulin conformations. A total of 20 snapshots containing 20 058 atoms (Figure 2) were considered in the test. The LocalSCF and MOZYME programs were used in their fast modes as described above. These results are compared with conformational energies, dipole moments, and partial atomic charges calculated by LocalSCF in default settings (LSCFR) because the latter shows the best agreement with matrix diagonalization according to Tables 2 and 3. The reference LSCFR calculations required 198 MB of memory and 34 826 s of CPU time per box for the AM1 Hamiltonian, whereas MOZ12 calculation required 1500 MB of memory, which did not fit into 1 GB of the available memory.

The analysis of the performed calculations is presented in Table 4. Use of the FMM in a LocalSCF calculation resulted in significant deviations of LSCFF results from the reference data for the PM5 Hamiltonian, whereas existing FMM implementation worked well for all other Hamiltonians. LSCFF conformational energies are in much better agreement with reference data than the MOZ6 results. Similarly, LSCFF dipole moments are closer to the reference data than the MOZ6 ones. Deviations of partial atomic charges are all below 0.002 electron units for both methods, with MNDO and AM1 Hamiltonians favoring the MOZYME method and PM3 favoring the LocalSCF one. LSCFF calculation is about 3-fold faster than MOZ6 and 3-fold more economical in memory requirement.

The computational advantage of the VFL approximation, LocalSCF linear scaling solution, and short LMOs extends beyond the test cases discussed in this work. For example, performing LocalSCF calculations with disabled LMO expansion on the insulin in a water box provides an additional 3-fold speedup for the AM1 Hamiltonian with an average LMO size of 23 atomic centers. However, it is not yet clear whether the existing semiempirical Hamiltonians will remain

***Table 4.*** LocalSCF (LSCFF) and MOZYME (MOZ6) in Their Fast Mode Settings; RMS Differences for 20 Insulin−Water Boxes Containing 20 058 Atoms in Comparison with LocalSCF in Accurate Settings (LSCFR)

| | MNDO | | AM1 | | PM3 | | PM5 | |
|---|---|---|---|---|---|---|---|---|
| | LSCFF | MOZ6 | LSCFF | MOZ6 | LSCFF | MOZ6 | LSCFF | MOZ6 |
| $E_{conf}$, kcal/mol | 4.07 | 32.75 | 5.42 | 25.56 | 1.94 | 23.82 | 54.06 | 16.33 |
| $\mu$, Debye | 0.86 | 3.71 | 1.13 | 3.50 | 0.22 | 3.40 | 10.15 | 2.00 |
| $Q_{\text{non−hydrogen}}$, $e$ | 0.0013 | 0.0010 | 0.0016 | 0.0010 | 0.0003 | 0.0012 | 0.0157 | 0.0013 |
| $Q_{\text{hydrogen}}$, $e$ | 0.0009 | 0.0008 | 0.0011 | 0.0008 | 0.0002 | 0.0008 | 0.0109 | 0.0009 |
| CPU time, sec | 3707 | 10039 | 4018 | 11682 | 5824 | 19035 | 5560 | 14388 |
| memory, MB | 241 | 816 | 256 | 816 | 283 | 816 | 270 | 816 |

valid for such extremely short LMOs. New semiempirical parameters are necessary in order to utilize in full the performance advantages of the extremely short LMOs.

## Conclusions

Linear scaling LocalSCF and MOZYME methods were studied in this work with respect to computational performance and the ability to reproduce matrix diagonalization results. Both diagonalization alternatives were found to reliably reproduce the target data for the MNDO, AM1, and PM3 Hamiltonians with the differences staying below a typical accuracy of semiempirical methods. Somewhat, larger differences between linear scaling and diagonalization results were observed for the PM5 Hamiltonian, pointing to the higher level of difficulty introduced by the PM5 Hamiltonian for the linear scaling LocalSCF and MOZYME methods. In the advent of modern linear scaling methods, future efforts on the development of new semiempirical Hamiltonians for biological applications should ideally target the task of facilitating the linear scaling. The performed tests indicate that MOZYME is more accurate in the prediction of total energy, which is systematically underestimated by LocalSCF due to the employment of relatively short LMOs. However, the LocalSCF method is shown to be more reliable in the prediction of conformational energies, dipole moments, and atomic charges, thereby supporting the conclusion about the predictive abilities of the LocalSCF approximation for the linear scaling problem. The increase of the LocalSCF geometry gradient errors in the FMM mode points to the need for the better optimization of the LocalSCF−FMM interface. For the large box of insulin in water, containing 20 058 atoms, the LocalSCF computer program is about 3-fold faster and 3-fold more economical by memory in comparison to the MOZYME program. However, the relative performances obtained are only rough estimations of the true capabilities of the LocalSCF and MOZYME theoretical methods because the details of their computer implementations cannot be entirely eliminated from influencing the test results. Also, the performed tests are limited to the studied systems, and the proposed methodology of the comparison should be applied to other molecular systems before the final conclusions can be drawn for a general case.

**Acknowledgment.** The authors thank Fujitsu Limited for granting the access to the LocalSCF and MOPAC2002 computer programs.

**Supporting Information Available:** Tables listing computational details for each individual snapshot and particular Hamiltonian. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Anikin, N. A.; Anisimov, V. M.; Bugaenko, V. L.; Bobrikov, V. V.; Andreyev, A. M. LocalSCF Method for Semiempirical Quantum-Chemical Calculation of Ultralarge Biomolecules. *J. Chem. Phys.* **2004**, *121* (3), 1266−1270.

(2) Bugaenko, V. L.; Bobrikov, V. V.; Andreyev, A. M.; Anikin, N. A.; Anisimov, V. M. *LocalSCF2 User Manual*; Fujitsu Ltd.: Tokyo, Japan, 2005.

(3) Stewart, J. J. P. Application of Localized Molecular Orbitals to the Solution of Semiempirical Self-Consistent Field Equations. *Int. J. Quantum Chem.* **1996**, *58* (2), 133−146.

(4) Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23* (2), 69−89.

(5) Hall, G. G. The Molecular Orbital Theory of Chemical Valency VIII: A Method for Calculating Ionization Potentials. *Proc. R. Soc. London* **1951**, *A205*, 541−5552.

(6) Lee, T.-S.; York, D. M.; Yang, W. Linear-Scaling Semiempirical Quantum Calculations for Macromolecules. *J. Chem. Phys.* **1996**, *105* (7), 2744−2750.

(7) Dixon, S. L.; Merz, K. M., Jr. Semiempirical Molecular Orbital Calculations with Linear System Size Scaling. *J. Chem. Phys.* **1996**, *104* (17), 6643−6649.

(8) Millam, J. M.; Scuseria, G. E. Linear Scaling Conjugate Gradient Density Matrix Search as an Alternative to Diagonalization for First Principles Electronic Structure Calculations. *J. Chem. Phys.* **1997**, *106* (13), 5569−5577.

(9) Seijo, L.; Barandiaran, Z. Parallel, Linear-Scaling Building-Block and Embedding Method Based on Localized Orbitals and Orbital-Specific Basis Sets. *J. Chem. Phys.* **2004**, *121* (14), 6698−6709.

(10) Goedecker, S.; Scuseria, G. E. Linear Scaling Electronic Structure Methods in Chemistry and Physics. *Comput. Sci. Eng.* **2003**, *5* (4), 14−21.

(11) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(12) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586−3616.

(13) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **1995**, *103* (11), 4613−4621.

(14) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327−341.

(15) York, D. M.; Darden, T. A.; Pedersen, L. G. The Effect of Long-Range Electrostatic Interactions in Simulations of Macromolecular Crystals: A Comparison of the Ewald and Truncated List Methods. *J. Chem. Phys.* **1993**, *99* (10), 8345−8388.

(16) Greengard, L. *The Rapid Evaluation of Potential Fields in Particle Systems*; MIT Press: Cambridge, MA, 1988.

(17) White, C. A.; Head-Gordon, M. Derivation and Efficient Implementation of the Fast Multipole Method. *J. Chem. Phys.* **1994**, *101* (8), 6593−6605.

(18) Dewar, M. J. S.; Thiel, W. Ground States of Molecules. 38. The MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99* (15), 4899−4907.

(19) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902−3909.

(20) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.* **1989**, *10* (2), 209−220.

(21) Stewart, J. J. P. *MOPAC2002*, release 2.1; Fujitsu Limited: Tokyo, Japan, 2002.

(22) Anisimov, V. M.; Anikin, N.; Bugaenko, V.; Bobrikov, V.; Andreyev, A. Accuracy Assessment of Semiempirical Molecular Electrostatic Potential of Proteins. *Theor. Chem. Acc.* **2003**, *109* (4), 213-219.

(23) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. A Semiempirical Quantum Model for Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Chem. Theory Comput.* **2005**, *1* (6), 1275−1285.

(24) Villar, R.; Gil, M. J.; García, J. I.; Martínez-Merino, V. Are AM1 Ligand−Protein Binding Enthalpies Good Enough for Use in the Rational Design of New Drugs? *J. Comput. Chem.* **2005**, *26* (13), 1347−1358.

(25) Metzger, T. G.; Ferguson, D. M.; Glauser, W. A. A Computational Analysis of Interaction Energies in Methane and Neopentane Dimer Systems. *J. Comput. Chem.* **1998**, *18* (1), 70−79.

(26) Vasilyev, V.; Bliznyuk, A. Application of Semiempirical Quantum Chemical Methods as a Scoring Function in Docking. *Theor. Chem. Acc.* **2004**, *112* (4), 313−317.

CT600222T