Self-Consistent Field Convergence for Proteins: A Comparison of Full and Localized-Molecular-Orbital Schemes

Christian R. Wick,[a] Matthias Hennemann,[a] James J. P. Stewart[b] and Timothy Clark[a,c,*]

[a] Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Department Chemie und Pharmazie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, 91052 Erlangen, Germany.

[b] *Stewart Computational Chemistry, 15210 Paddington Circle, Colorado Springs, CO 80921-2512, USA.*

[c] Centre for Molecular Design, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth PO1 2DY, United Kingdom.

Abstract

Proteins in the gas phase present an extreme (and unrealistic) challenge for self-consistent-field iteration schemes because their ionized groups are very strong electron donors or acceptors, depending on their formal charge. This means that gas-phase proteins have a very small band gap but that their frontier orbitals are localized compared to “normal” conjugated semiconductors. The frontier orbitals are thus likely to be separated in space so that they are close to, but not quite, orthogonal during the SCF iterations. We report full SCF calculations using the massively parallel EMPIRE code and linear scaling localized-molecular-orbital (LMO) calculations using Mopac2009. The LMO procedure can lead to artificially over-polarized wavefunctions in gas-phase proteins. The full SCF iteration procedure can be very slow to converge because many cycles are needed to overcome the over-polarization by inductive charge shifts. Example molecules have been constructed to demonstrate this behavior. The two approaches give identical results if solvent effects are included.

Keywords:

Self-consistent field; linear scaling; NDDO; proteins; LMO-SCF

Introduction

MNDO-like [1] NDDO (Neglect of Diatomic Differential Overlap) self-consistent-field (SCF) calculations are now being applied routinely to systems of thousands of atoms. Both linear-scaling techniques such as divide and conquer (D&C) [2,3,4] and the localized-molecular-orbital (LMO) technique [5,6] or conventional but highly parallel calculations [7] are now available that can handle tens of thousands of atoms easily, so that we are now able for the first time to compare the wavefunctions obtained for very large systems with linear-scaling and conventional algorithms. During the development of the EMPIRE code, [7] it became evident that SCF convergence is very slow for physically unrealistic but testing gas-phase calculations on zwitterionic (i.e. almost all) proteins, whereas such calculations converge very effectively using the LMO-SCF technique. [5,6] Closer investigation of this phenomenon suggested that the very slow inductive charge-transfer process that made the conventional SCF calculations so slow to converge is prevented in the LMO-SCF scheme, so that we might expect the two procedures to converge to different wavefunctions. We now report a detailed study of this phenomenon and specify the types of system for which the results of LMO- and conventional SCF schemes may be expected to give different results.

Theoretical Background and Computational Details

The EMPIRE [7] and Mopac2009 [8] programs were used for all calculations, which used the AM1 Hamiltonian. [9] EMPIRE uses a conventional SCF scheme in which the initial guess is obtained by diagonalization of an extended-Hückel-like matrix [7] and the SCF-iterations simply involve a parallel pseudodiagonalization [10] step, possibly combined with a separate calculation of the Eigenvalues of the Fock matrix. This procedure has been demonstrated to converge to the same wavefunction as conventional SCF iterations using full diagonalizations and is terminated by a full diagonalization of the Fock matrix to obtain canonical molecular orbitals (MOs).

LMO-SCF calculations using Mopac2009 start with an LMO initial guess and achieve linear scaling by ensuring that the MOs remain local during the SCF iterations. [5] This enforced locality reduces both the numbers of virtual/occupied pairs of MOs to be rotated during the SCF and the number of atomic orbitals involved in each virtual/occupied rotation, thus ensuring linear scaling. Mopac2009 also allows the use of cutoffs for, for instance the two-electron integrals, in order to speed up calculations. In order to retain compatibility with

EMPIRE, cutoffs were not used in LMO-SCF calculations with Mopac2009 where possible (standard cutoffs were applied in calculations with explicit solvent).

Even without cutoffs, Mopac2009 and EMPIRE are not exactly comparable because they use different convergence criteria and some physical constants are marginally different between the two programs. This would not normally be important but for very large systems (with heats of formation of several thousand kcal mol^{-1}) these small differences can lead to noticeably different calculated energies. We therefore used EMPIRE to compare the energies given by different wavefunctions by reading the converged MOs from Mopac2009 calculations into EMPIRE as the initial guess and either calculating the electronic energy non-iteratively (i.e. simply calculating the energy given by the initial guess) or allowing the calculation to converge using the conventional EMPIRE SCF-procedure. All energies reported below are calculated in this way unless otherwise noted.

The protein structure was based on the X-ray structure of hNur77 [11] and was taken from another study in which the protein was subjected to molecular-dynamics simulations with AMBER. [12] The protein was placed in an octahedral water box with TIP4PEw water. [13] The AMBER 1999 force field (ff99SB, [14]) was used for the simulations. The Particle Mesh Ewald (PME, [15]) technique was used to treat long-range electrostatics and constant-pressure periodic boundary conditions were applied. SHAKE [16] was used to constrain bonds to hydrogen atoms, allowing an integration step size of 2 fs. After initial unconstrained minimization, the system was equilibrated for 200 ps at 300 K after slowly heating over a period of 100 ps by coupling to a heat bath with Cartesian restraints on backbone atoms.

A force-field optimized structure starting from a snapshot from the equilibrated simulation was used for the AM1 calculations. In one case, the "gas-phase" protein was constructed by removing all the solvent molecules, and in the other, the periodic water box was truncated to a non-periodic water shell surrounding the protein for the AM1 calculation. The former comprised 3,707 atoms and the latter 9,929. We emphasize that protein calculations without solvent are artificial; our purpose here is to investigate and define the behavior of the alternative SCF-procedures.

hNur77: A Test Protein

AM1 single-point calculations on the snapshot geometry for the gas-phase protein given in the Supporting Material gave wavefunctions that correspond to an EMPIRE Heat of Formation of

–9,845.39 kcal mol^{-1} for the conventional SCF calculation and –9,803.28 kcal mol^{-1} (42.11 kcal mol^{-1} less stable) for the Heat of Formation for the LMO-SCF wavefunction. Because of the program differences outlined above, the original Mopac2009 Heat of Formation was –9,833.70 kcal mol^{-1}. A full SCF calculation with EMPIRE using the LMO-SCF wavefunction as initial guess converged to the same energy as the pure EMPIRE calculation using the standard initial guess.

An analysis of the Coulson net atomic charges for the two calculations is shown in Figure 1.

(Figure 1 here)

Figure 1 shows that charged groups have higher numerical charges (i.e. are more highly charged) in the LMO-SCF calculation than in the full SCF. The nature of the discrepancy is clearer in the histogram of the differences in Coulson atomic charges for the two calculations shown in Figure 2.

(Figure 2 here)

This phenomenon explains the slow convergence of gas-phase proteins in EMPIRE compared to the very fast convergence in Mopac2009. The lowest unoccupied MOs are localized on formally positively charged groups and the highest occupied ones on formally negative groups. When these centers are far from each other, there is no direct overlap but shifting electrons from the negatively charged group to the positive one will nonetheless result in a reduction of the charge separation and of the total energy by reducing the charge separation. Within the SCF iteration scheme, this charge transfer can only occur inductively in a stepwise fashion through the intervening atoms. This process is slow and results in the large number of iterations needed for the full SCF calculation. Figure 1 not only shows the highly charged groups that differ strongly between the two procedures, but also indicates the paths by which the charges wander through the protein during the SCF-iterations as small residual charge differences. Exactly this result would be expected from the interpretation given above.

In the LMO-SCF scheme, the initial charges on the charged groups are high (as is also the case for the extended Hückel-like initial guess used in EMPIRE [5]). However, when the slow charge-transfer by induction begins to occur in the SCF calculation, it is negated by the re-localization procedure used in Mopac2009. [4] Thus, long-range inductive charge transfer cannot occur within the LMO-SCF scheme, which results in the differences observed for

hNur77 above. The progress of the charge transfer is illustrated for an EMPIRE SCF calculation on hNur77 in Figure 3 for selected charged residues.

(Figure 3 here)

The slow migration of charge to decrease the charge separation in the final converged solution can be seen clearly. This charge migration is prevented by the combination of the LMO initial guess and the re-localization step in the LMO-SCF procedure.

Test Molecules

In order to test exactly when the full and the LMO-SCF procedures deviate from one another, we constructed the zwitterionic test molecules **1** and **2** (see Scheme 1).

(Scheme 1 here)

Figure 4 shows the observed differences in the heats of formation (calculated as outlined above) between the LMO-SCF and full schemes.

(Figure 4 here)

Model compound **1** gives identical results for the two programs, whereas compound **2** with more highly charged separated groups exhibits the same behavior as found for the protein. The results agree for short alkane chains and then deviate to give a constant deviation of approximately 80 kcal mol^{-1} for n=35 and larger. We thus conclude that pairs of singly charged zwitterionic centers do not lead to differences between the LMO-SCF procedure and the full SCF, but that more highly charged residues separated by large distances do. The energetic effect reaches a plateau value at a distance between the charged centers of approximately 50 Å. The smallest molecule for which a significant difference (3 kcal mol^{-1}) is found between the two procedures is **3**, in which the charge centers are eight bonds, or approximately 10 Å, apart (see Scheme 2). The geometry used for the calculations on **3** is included in the Supporting Information.

(Scheme 2 here)

The number of bonds separating the highly charged groups is the important factor because it determines whether the MOs are relocalized during the LMO-SCF procedure.

The Effect of Solvent

The above results apply to the gas phase and are thus not relevant for real-world protein calculations. As no implicit solvent model is yet implemented in EMPIRE, we chose to compare the two SCF-formalisms by calculating hNur77 in an explicit water box taken from a classical molecular-dynamics simulation. The system consisted of the protein and 2,074 water molecules to give 9,929 atoms. In this case (when the localized charges are stabilized by the solvent environment), both programs converge to the same wavefunction, as shown in Figure 5, a plot of the differences in Coulson charges for the protein atoms given by the two programs.

(Figure 5 here)

The largest deviations in Coulson charges found are well below 0.01 electrons, indicating that the two different SCF schemes have converged to the same wavefunction.

Conclusions

It is possible to construct molecules with highly charged groups that do not converge to the variational wavefunction using the LMO-SCF procedure. The LMO-SCF wavefunction is "more polar" than the variational one because remote charged groups cannot transfer charge from one to the other if the molecular orbitals are re-localized during the SCF calculation. This effect does not arise for simple zwitterions with one positive and one negative center, but is likely for most proteins, which have several charged groups of each polarity. The result is that, for instance, local properties [17] calculated from the LMO-SCF wavefunction will not be correct for problem molecules, even after a full diagonalization of the Fock matrix to calculate the canonical molecular orbitals.

However, no difference between LMO-SCF and the full (pseudodiagonalization-based) procedure is found for solvated proteins, so that, protein calculations that use an implicit solvent model will converge to the variational wavefunction.

The above discussion is restricted to restricted Hartree-Fock (RHF) SCF calculations. In many of the examples discussed, the global minimum wavefunction has significant open-shell character (i.e. the RHF calculations exhibit UHF instability). In such cases, the variational RHF wavefunction represents an electronic stationary point but not the global minimum. We have not investigated the effect described for unrestricted calculations because such large molecules are likely to exhibit many almost degenerate UHF wavefunctions, so that comparisons between different iteration schemes become very difficult.

Supporting Information

The molecular structures used for hNur77 and 3 (.xyz) files are available as electronic Supporting Information.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft as part of the Excellence Cluster *Engineering of Advanced Materials* and by the *Bundesministerium für Bildung und Forschung* as part of the high-performance Computer-Aided Drug Design (hpCADD) project.

References

[1] Clark T, Stewart JJP (2011) MNDO-like Semiempirical Molecular Orbital Theory and its Application to Large Systems. In: Reimers JJ (ed) Computational Methods for Large Systems. Wiley, Chichester, Chapter 8 (ISBN: 978-0-470-48788-4)

[2] Yang W (1991) Direct calculation of electron density in density-functional theory. Phys Rev Lett 66:1438–1441

[3] Dixon SL, Merz Jr KM (1997) Fast, accurate semiempirical molecular orbital calculations for macromolecules. J Chem Phys 107:879–893

[4] Ababoua A, van der Vaart A, Gogonea V, Merz Jr KM (2007) Interaction energy decomposition in protein-protein association: a quantum mechanical study of barnase-barstar complex. Biophys Chem 125:221–236

[5] Stewart JJP (1996) Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. Int J Quant Chem 58:133–146

[6] Stewart JJP (2009) Application of the PM6 method to modeling proteins. J Mol Model 15:765–805

[7] Hennemann M, Clark T (2013) EMPIRE. Universität Erlangen-Nürnberg and Cepos InSilico Ltd (http://www.ceposinsilico.de/products/empire.htm), accessed January 19th 2014

[8] Stewart JJP (2008) MOPAC2009. Stewart Computational Chemistry, Colorado Springs, CO, USA. http://OpenMOPAC.net, accessed January 19th 2014

[9] Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc 107:3902–3909

[10] Stewart JJP, Császár P, Pulay P (1982) Fast semiempirical calculations. J Comput Chem 3:227–228

[11] Min JR, Schuetz A, Loppnau P, Weigelt J, Sundstrom M, Arrowsmith CH, Edwards AM, Bochkarev A, Plotnikov AN (2014) Human NR4A1 ligand-binding domain. http://www.pdb.org/pdb/explore/explore.do?structureId=2QW4, accessed January 19th 2014

[12] Case DA, Darden TA, Cheatham III TE, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Hayik S, Roitberg A, Seabra G, Swails J, Goetz AW, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wolf RM, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh M-J, Cui G, Roe DR, Mathews DH, Seetin MG, Salomon-Ferrer R, Sagui C, Babin V, Luchko T,

Gusarov S, Kovalenko A, Kollman PA (2012) AMBER 12. University of California, San Francisco. http://ambermd.org/, accessed January 19th 2014

[13] Horn HW, Swope WC, Pitera JD, Madura JD, Dick TJ, Hura GL, Head-Gordon T (2004) Development of an improved four-site water model for biomolecular simulations: TIP4P–Ew. J Chem Phys 120:9665–9678

[14] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65:712–725

[15] Darden T, York D, Pederson L (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092

[16] Yoneya M, Berendsen HJC, Hirasawa K (1994) A Noniterative Matrix Method for Constraint Molecular-Dynamics Simulations. Molecular Simulations 13:395–405

[17] Ehresmann B, Martin B, Horn AHC, Clark T (2003) Local molecular properties and their use in predicting reactivity. J Mol Model 9:342–347

Figure 1. *Space-filling model:* Differences in Coulson net atomic charges between the LMO-SCF and conventional wavefunctions for hNUR77. Atoms with charge differences larger than ± 0.01 are shown as spheres. Red space-filling atoms and positive numbers indicate less negative charges for the full SCF calculation and blue atoms and negative numbers less positive.

Figure 2. Histogram of the charge differences between the full SCF calculation and LMO-SCF for the hNur77 structure shown in Figure 1. Charges were summed over all atoms for each amino-acid residue.

Figure 3. Coulson charges for each cycle during a full SCF calculation for hNur77. The sum of the residues His10, Lys94 and 132 is represented by the blue line, and that of residues Asp218, Asp232 and Thr233 is shown in red. Dashed lines represent the final Coulson charges obtained with the LMO-SCF method for corresponding residues.

Figure 4. Difference between heats of formation calculated with EMPIRE and with the LMO-SCF formalism for compounds **1** and **2** with increasing numbers of CH_2 groups (from n=0 to 58). The LMO-SCF heat of formation was subtracted from the full SCF heat of formation, thus negative numbers indicate a higher (less stable) energy for LMO-SCF.

Figure 5. Histogram of the charge differences between the full SCF calculation and LMO-SCF for the hNur77 structure shown in Figure 1 in an explicit water box summed over all atoms of each residue.

Scheme 1. General structure of the model compounds

Scheme 2. Smallest model compound that shows a significant difference between LMO-SCF and Full-SCF.