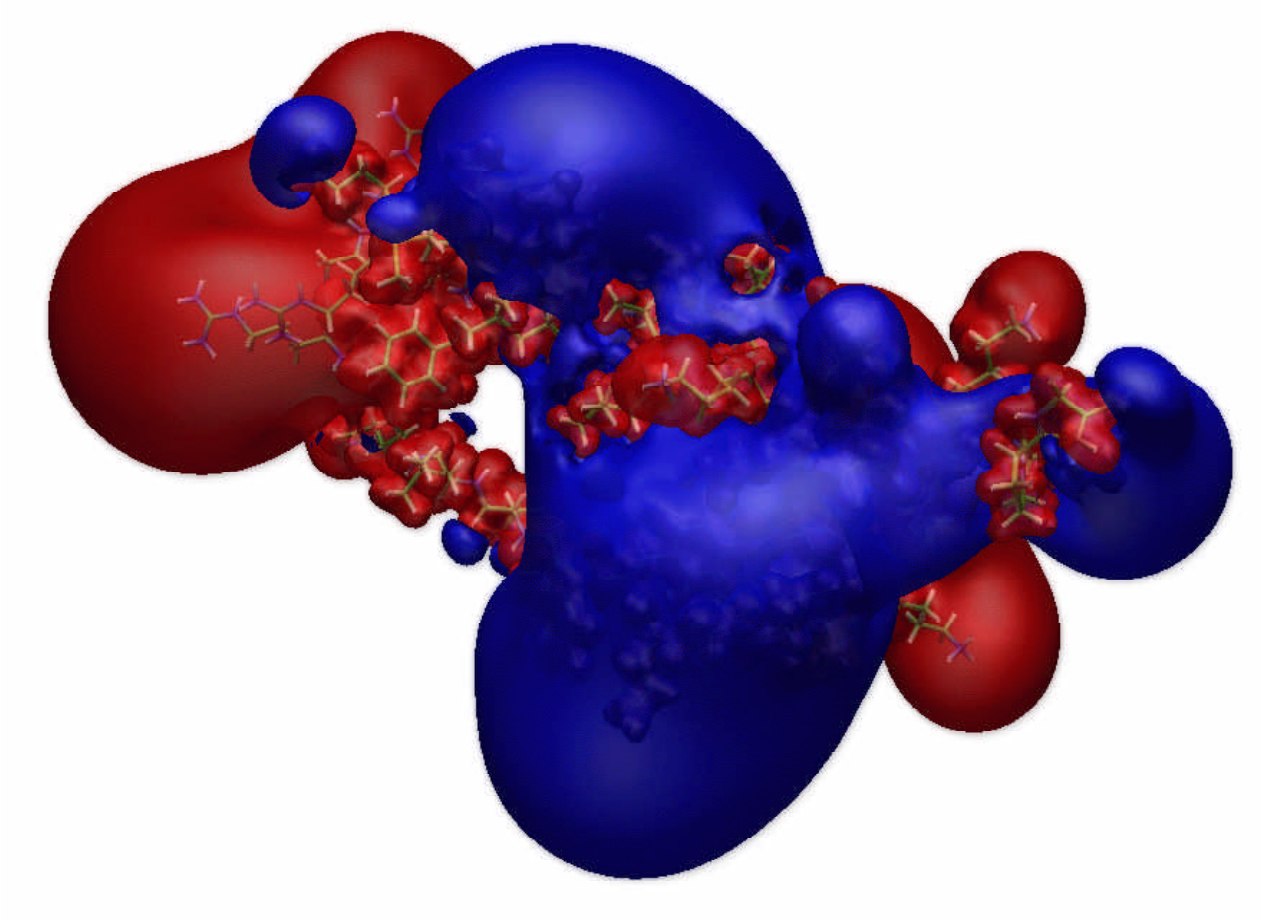
Multipole Methods for Solving Electronic Wavefunctions

A Method for Achieving Linear Scaling in Electronic Structure Calculations



A Third Year Chemistry Technical Report
June 1999

Supervisor: Dr. I. Gould

Ross Walker
(96Chem518)
M.Sci. Chemistry(F103)

Imperial College of Science, Technology & Medicine

Table of Contents

Cover Illustration[1]*:* Image showing iso-surfaces of the electrostatic potential of the p53 tumour suppresser tetramerization monomer. The result was obtained by M.Challacombe *et al.*[2] using the RHF/3-21G level of theory with newly developed linear scaling Hartree-Fock methods. The calculation involved a total of 698 atoms and employed a total of 3836 Gaussian basis functions

1. Introduction

> *"The underlying laws necessary for the mathematical theory of large parts of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."*
>
> **(Paul Dirac - 1929)**[3]

Since the advent of computers in the second half of this century mankind has become more or less dependent on them for numerous tasks. Since the 1940's computers have vastly increased in power, in recent times effectively doubling in computational speed every 18 months,[4] to the point where todays fastest computers can execute more than a trillion (1×10^{12}) floating point operations per second![a] At the same time as computational speed has increased so has the storage capacity of modern computers such that it is now common for modern supercomputers to have gigabytes of primary (RAM) storage and terabytes of secondary (disc) storage.

These rapid advances in computing power have brought with them a whole new field of theoretical science termed computational science. This field employs the use of computers for modelling and simulating a range of different properties from the structural forces in a jet aircraft or the gravitational interactions between two galaxies to the electron distribution within single molecules.

One such field brought about by these advances in computational power is that of computational quantum chemistry whereby molecules and reactions are modelled by solving approximations of the Schrödinger equation using computers. Such procedures can yield a large amount of valuable data about the way in which molecules behave and interact so that computer based calculations are now routinely used to supplement experimental techniques.

[a] *Intel ASCI Red*. Sandia National Laboratories, Albuquerque (Nov 1998). *Source* - http://www.top500.org.

However, the big problem facing researchers in this field is the complexity of the calculations involved. Quantum mechanics states that the energy, and related properties, of a given molecule can be obtained by solving the time-independent Schrödinger equation (*eq. 1.1*).[5]

$$H\Psi = E\Psi \qquad 1.1$$

Where: H is the Hamiltonian operator.
Ψ is the wavefunction describing the system.
E is the systems energy.

However, exact solutions to equation 1.1 are not possible for systems larger than H_2^+ so in order to find solutions for larger chemical systems theoretical chemists have been forced to devise a range of different approximations to this equation that are theoretically solvable. The development of these various approximations have had such an impact on the field of computational chemistry that the 1998 Noble prize for chemistry was awarded, jointly, to Walter Kohn (*fig. 1.1*) for his development of density functional theory and John Pople (*fig. 1.2*) for his development of computational methods in quantum chemistry, in particular his work on the program Gaussian which is now the de facto standard for *ab initio* quantum chemistry calculations.[6]

Figure 1.1 Walter Kohn joint winner of 1998 Nobel prize for chemistry.

Figure 1.2 John Pople joint winner of 1998 Nobel prize for chemistry.

As discussed in section 5 the main problem with quantum chemistry calculations using approximations to the Schrödinger equation is trying to achieve an approximation which gives acceptably accurate results but which scales linearly with problem size. This is the so called "*Holy Grail*" of quantum chemistry[7] and it is one such promising linear scaling approximation method, the use of fast multipole methods,[8] which will be discussed in this report.

2. Quantum Chemistry - Historical Background[9]

When quantum mechanics was formulated some 70 years ago it laid the theoretical foundation for modern physics and chemistry.[10] This made possible, in principle, the understanding of how electrons and nuclei interact to form chemical bonds. However, as stated by Paul Dirac in 1929[11]:

> *"The underlying laws necessary for the mathematical theory of large parts of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."*

Although very true in 1929 with the advances in computers made in the early 60-ies scientists began to challenge this pessimistic view employing computers to look upon the complex equations from new angles.

Initial attempts were made using the so called independent particle model or Hartree-Fock (HF) method that was originally developed in the 1930's by Hartree, Fock, Slater and others and successfully applied to atoms. In this model the famous many-body problem is overcome and reduced to a set of single particle (orbital) problems by making the assumption that each electron moves independently of the others present and experiences a mean field from the fixed nuclei and other electrons.

An important contribution to this theory was made by C. C. J. Roothaan who published a paper in 1951[12] where it was suggested that the orbitals could be expanded into a set of basis functions, also termed atomic orbitals, which could be expressed as *Slater Type Orbitals*[13] of the functional form, in spherical harmonics, given in equation 2.1.

$$\chi_{\varsigma,n,l,m}(r,\theta,\vartheta) = NY_{l,m}(\theta,\vartheta)r^{n-1}e^{-\varsigma r} \qquad 2.1$$

Where: N is a normalisation constant.
$Y_{l,m}$ is the spherical harmonic function.

This procedure allowed the differential equations to be expressed as a matrix problem which is well suited to solving via computer. A further important contribution to this theory was made by S. F. Boys who suggested that the basis functions, instead of being described by an e^r

function as in *Slater Type Orbitals* could instead be in the form of *Gaussian type*[14] functions with an e^{r^2} dependence as in equation 2.2.

$$\chi_{\varsigma,n,l,m}(r,\theta,\vartheta) = NY_{l,m}(\theta,\vartheta)r^{(2n-2-l)}e^{-\varsigma r^2}$$
$$\chi_{\varsigma,l_x,l_y,l_z}(x,y,z) = Nx^{l_x}y^{l_y}z^{l_z}e^{-\varsigma r^2} \qquad 2.2$$

Expressing the orbitals in this form leads to very large simplifications in calculating the necessary integrals.

The HF method for modelling of molecules was then developed in the 60-ies and works as an approximation to the Schrödinger equation in which the wave function Ψ is written as the product of one-electron functions (orbitals). HF theory acts as the starting point for a hierarchy of methods that aim to obtain as accurate solutions as possible to the Schrödinger equation.

Originally many believed that non-empirical methods, such as HF theory, would never compete with the semi-empirical methods since the computer resources required would be far too large. John Pople was one such person who held this view but it was he who, over the next decade, changed this situation.

The Hartree-Fock method effectively consists of two major computational steps. In the first step the molecular orbitals are expanded, as suggested by Boys[15], into a basis set composed of Gaussian functions. This allows the one-electron Schrödinger equation to be replaced by a matrix eigen-value problem which yields the orbital energies and expansion coefficients. The elements of this matrix consist of integrals over the basis functions that describe the various energy components, such as the nuclear attraction energy, the kinetic energy and the electron-electron repulsion energy. The second step consists of diagonalising the HF matrix which is an iterative procedure that has to be repeated until self consistency is achieved.

The most demanding in terms of computer resources is the calculation of the integrals, especially the electron repulsion integrals. The problem with these integrals is that there are a very large number of them (10^6 - 10^9) and they are six dimensional. Pople devised an approach which reduced the computational costs by up to two orders of magnitude. This was a decisive achievement and made possible the use of HF calculations for real chemical

applications. Following this came a number of improvements and differing theories, one being the so called Density Functional Theory (DFT) devised by Walter Kohn where it was proven that the exact ground state electron density uniquely specifies the acting one-electron potential $v(r)$. Since the kinetic energy of the electrons and their coulomb interaction cannot be modified, it was concluded that the ground state density specifies the Hamiltonian of the system and hence the properties of the ground state.[16,17] Further work expanded on this theory arriving, in 1992, with a computational implementation that could treat systems with hundreds of atoms.

From this point forward computational chemistry developed considerably from a stage 30 years ago where many ridiculed the subject as a futile undertaking with little effect on chemistry to a point today where the general consensus of opinion is that computational chemistry is one of the most important developments in chemistry over the last two decades. However, one problem still remains. Although the methods briefly discussed above greatly simplify the computational complexity of the problem they still scale as the cube, or greater, of system size. Since computers only double in power typically every 18 months, as discussed in section 5, there is theoretically a maximum system size that can be solved. Hence in order to progress further and make computational chemistry a truly versatile tool that can accurately model anything from simple systems to proteins or DNA it is necessary to develop methods that scale linearly with system size. Many people are currently working in this field developing several methods with results that, to date, are encouraging. It is these methods that will be discussed in detail later in this report.

3. Background Theory[b]

In the previous two sections a brief overview of the developments that have taken place in the field of quantum chemistry has been given. Before going on to discuss the fast multipole method for achieving linear scaling in electronic structure calculations it is necessary to describe some of the background theory behind quantum chemistry.

3.1 Wavefunctions

The Schrödinger equation is a well known entity in modern science. In its barest, time independent, form it states.

$$H\Psi = E\Psi \qquad 3.1$$

In this equation H is the short hand notation for the Hamiltonian operator which operates on the mathematical function Ψ, which represents the wave function of the system, to yield the energy E. Writing the Schrödinger equation in this way disguises the fact that this equation is in fact a set of differential equations each with a function Ψ_n corresponding to each allowed energy E_n. Thus only for the case of a hydrogen atom, with a single electron outside a single positively charged nucleus, is it possible to solve the equation exactly.

Once the wave function is known for a particular state of a system it is then possible, in theory, to determine any physical observable using equation 3.2.

$$\text{Observable} = \frac{\int \Psi^* \langle \text{operator} \rangle \Psi \, d\tau}{\int \Psi^* \Psi \, d\tau} \qquad 3.2$$

The operator used is that which is appropriate to the observable required. i.e The Hamiltonian H for energy, another for dipole moment, charge density etc.

[b] Presented here is a very brief overview of the theory involved in calculating electronic structure. For a more in depth discussion of quantum mechanics the reader is referred to P. W. Atkins & R. S. Friedman, "Molecular Quantum Mechanics". Chapter 9 of this book also gives a well written discussion of the methods behind calculating electronic structure.

3.2 Atomic Orbitals and Spin

Wavefunctions which satisfy the Schrödinger equation are often called *orbitals*. An atomic orbital is therefore just a mathematical function which is a solution to the time independent Schrödinger equation.

When dealing with polyelectronic atoms the Schrödinger equation cannot be solved analytically hence the so called *orbital approximation* is used. This approximation works by treating each electron independently, each having its own one-electron wave function or orbital. This is more commonly encountered in the standard procedure for describing orbital configurations of elements i.e.

$$\mathrm{O} : 1s^2 2s^2 2p^4$$

Using this approximation the total wave function of an atom, Ψ, is merely the product of the one-electron wavefunctions (χ_n) for each electron, i.e. (*eq. 3.3*).

$$\Psi = \chi_1 \chi_2 \chi_3 \chi_4 \ldots\ldots \chi_n \qquad 3.3$$

When writing the orbital configuration of an atom as in the example of oxygen above this is understood to be shorthand for the true representation.

$$1s^{\alpha} 1s^{\beta} 2s^{\alpha} 2s^{\beta} 2p_x{}^{\alpha} 2\mathrm{p_x}^{\beta} 2p_y{}^{\alpha} 2p_z{}^{\alpha}$$

where α and β represent the two possible orientations of electron spin (up & down).

Including the electron spin in equation 3.3, denoting spin α as χ and spin β as $\overline{\chi}$, yields, for the simpler case of helium, the wave function for the whole atom as a product of *spin-orbitals* (*eq. 3.4*).

$$\Psi_{He} = \chi_{1s}(1)\overline{\chi}_{1s}(2) \qquad 3.4$$

where the numbers in parentheses refer to the electron associated with each spin-orbital.

It can be seen, however, from equation 3.4 that this does not satisfy the Pauli exclusion principle. This states that[19]:

"the wave function for the system (Ψ) must change sign if any pair of electrons are interchanged, since electrons are identical fermion particles."

Thus writing the total wave function as a simple product of one electron orbitals is not sufficient, as in equation 3.4, since:

$$\Psi_{He} = \chi_{1s}(1)\overline{\chi}_{1s}(2) \quad \& \quad \Psi'_{He} = \chi_{1s}(2)\overline{\chi}_{1s}(1) \qquad 3.5$$

Ψ' is therefore not the negative of Ψ.

In order to satisfy the Pauli exclusion principle it is necessary to express the wavefunction for the helium atom, using the orbital approximation, as in equation 3.6.

$$\Psi_{He} = \frac{1}{\sqrt{2}}\left[\chi_{1s}(1)\overline{\chi}_{1s}(2) - \chi_{1s}(2)\overline{\chi}_{1s}(1)\right] \qquad 3.6$$

(The $1/\sqrt{2}$ acts as a normalisation constant such that $\Psi^*\Psi d\tau = 1$. Integration is carried out over the element $d\tau$ since there are effectively 4 dimensions, dx, dy, dz and spin.)

However, on going to a more complex system such as beryllium it is necessary to allow for all possible interchanges between the four electrons present giving equation 3.7.

$$\begin{aligned}\Psi_{Be} = {} & \chi_{1s}(1)\overline{\chi}_{1s}(2)\chi_{2s}(3)\overline{\chi}_{2s}(4) - \overline{\chi}_{1s}(1)\chi_{1s}(2)\chi_{2s}(3)\overline{\chi}_{2s}(4) + \\ & \chi_{2s}(1)\overline{\chi}_{1s}(2)\overline{\chi}_{1s}(3)\chi_{2s}(4) - \overline{\chi}_{2s}(1)\chi_{1s}(2)\overline{\chi}_{1s}(3)\chi_{2s}(4) + \text{.....etc.} \\ & \text{in total 24 products.}\end{aligned} \qquad 3.7$$

It can quickly be seen that the number of products required will rise rapidly as the atomic number increases. Fortunately all these products are simply the expanded form of determinants. Thus equation 3.6 is simply the expanded form of the determinant given in equation 3.8 below.

$$\Psi_{He} = \frac{1}{\sqrt{2!}}\begin{vmatrix}\chi_{1s}(1) & \chi_{1s}(2) \\ \overline{\chi}_{1s}(1) & \overline{\chi}_{1s}(2)\end{vmatrix} \qquad 3.8$$

and for beryllium the long equation given in 3.7 can simply be expressed as:

$$\Psi_{Be} = \frac{1}{\sqrt{4!}} \begin{vmatrix} \chi_{1s}(1) & \chi_{1s}(2) & \chi_{1s}(3) & \chi_{1s}(4) \\ \overline{\chi}_{1s}(1) & \overline{\chi}_{1s}(2) & \overline{\chi}_{1s}(3) & \overline{\chi}_{1s}(4) \\ \chi_{2s}(1) & \chi_{2s}(2) & \chi_{2s}(3) & \chi_{2s}(4) \\ \overline{\chi}_{2s}(1) & \overline{\chi}_{2s}(2) & \overline{\chi}_{2s}(3) & \overline{\chi}_{2s}(4) \end{vmatrix} \qquad 3.9$$

Since only the diagonals of the determinants are required to define them the expanded spin-orbital wavefunctions are usually written just as the simple products in equation 3.10.

$$\begin{aligned} \Psi_{He} &= \chi_{1s}\overline{\chi}_{1s} \text{ or } {\chi_{1s}}^2 \\ \Psi_{Be} &= {\chi_{1s}}^2{\chi_{2s}}^2 \end{aligned} \qquad 3.10$$

Thus the wave function of an atom (Ψ) can be written generally as in equation 3.11 remembering that this is not just a simple product.

$$\Psi = \chi_1\chi_2\chi_3\chi_4\cdots\cdots\chi_n \qquad 3.11$$

The problem, therefore, is reduced to finding χ_1, χ_2 etc. These are each one-electron functions in polar co-ordinates multiplied by the electron spin (α *or* β). Each individual orbital (χ_n) can be expressed as a numerical function that assigns a numerical value to each point defined by the 3 co-ordinates.

Hartree produced some very accurate atomic functions of this type that were later analytically expressed by Slater,[20] of the form given in equation 2.1, and were termed Slater atomic wavefunctions. The variation of the spherical harmonic part is identical to the variation for the hydrogen atom wavefunctions so the differences from atom to atom are found only in the r-dependent part of the orbital.

3.3 Molecular Orbitals

The previous two sections have concentrated on the wavefunctions for atoms rather than molecules. In principle the wave function for a molecule is no different than that for an atom hence the orbital approximation given in equation 3.12 can be made where Ψ represents the molecular wave function and each function ϕ_n represents a three

dimensional function determining the properties of each individual electron in the molecule.

$$\Psi = \phi_1\phi_2\phi_3\phi_4....\phi_n \qquad 3.12$$

As discussed previously it is necessary to remember that this is not a simple product but rather the diagonal of a determinant since the wave function has to be antisymmetric with respect to electron interchange.

In quantum chemical calculations the goal is generally to derive the wave function, Ψ, for a given molecule. This can be done if the constituent molecular orbitals, ϕ_n, are known. The method normally employed for finding each ϕ_n is to expand each of the unknown molecular orbitals, ϕ_n, as a linear combination of known atomic orbitals as shown in equation 3.13.

$$\phi_n = \sum_k c_{nk}\chi_k \qquad 3.13$$

where each χ_k will be a function of the form

$$\chi_k = \text{constant} \times (\text{function of r}) \times (\text{spherical harmonic finction in } \theta \text{ and } \phi)$$

Thus the problem of calculating the wave function for a molecule reduces to finding the expansion coefficients c_{nk}. A simple example is that of the symmetric molecule Li_2 which has the molecular orbitals as shown in figure 3.1.

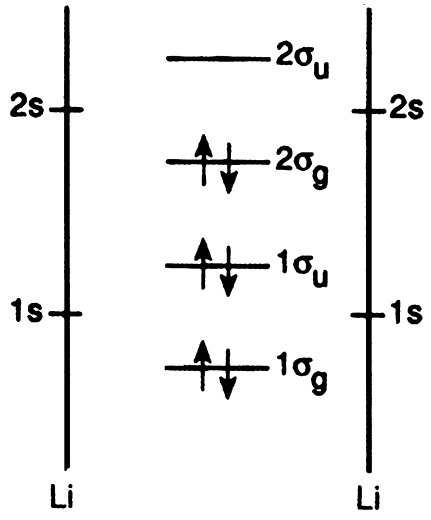


Figure 3.1: The molecular orbitals for diatomic Li_2. Adapted from G.H. Grant & W. G. Richards "Computational Chemistry" 1995, p.10.

The ground state molecular wave function for Li_2 will therefore be:

$$\Psi_{Li_2} = \phi_{1\sigma_g}\phi_{1\sigma_g}\phi_{1\sigma_u}\phi_{1\sigma_u}\phi_{2\sigma_g}\phi_{2\sigma_g} \qquad 3.14$$

with $\phi_{1\sigma_g}$ being defined by equation 3.15.

$$\begin{aligned}\phi_{1\sigma_g} &= c_{11}\chi_{1sLi} + c_{12}\chi_{1sLi} \\ \phi_{1\sigma_u} &= c_{21}\chi_{1sLi} + c_{22}\chi_{1sLi}\end{aligned} \qquad 3.15$$

Since Li_2 is symmetric to inversion for $1\sigma_g$ and antisymmetric for $1\sigma_u$ then

$$\begin{aligned}c_{11} &= c_{12} \\ c_{21} &= -c_{22}\end{aligned} \qquad 3.16$$

The variation principle states that the more flexible the wave function describing a system the lower the energy of that system will be. When all particles are an infinite distance apart the energy of the system is defined as zero hence all calculated energies are negative numbers with the lower the energy the larger the number. For Li_2 a lower energy can be obtained by using a more flexible wave function by extending the molecular orbital expansion.

$$\phi_n = c_{n1}\chi_{1sLi} + c_{n2}\chi_{2sLi} + c_{n3}\chi_{2pLi} \qquad 3.17$$

The problem, however, remains the same. In order to find the molecular wave function, Ψ, it is necessary to find each molecular orbital ϕ_n in terms of known functions (atomic orbitals) multiplied by coefficients that have to be determined. This is effectively what *ab initio*[c] computational methods attempt to find by varying the coefficients until an energy minimum is found.

[c] The term *ab initio* comes from Latin meaning 'from the beginning.'

4. The Hartree-Fock Self-Consistent Field Method

The HF-SCF method is a way of finding the approximate wave function of an atom or molecule as described in section 3 above. The HF method makes use of the Born-Oppenheimer approximation, where electron and nucleon movement are independent, and attempts to solve the Schrödinger equation (*eq. 3.1*) using, for the Hamiltonian, equation 4.1.[21]

$$H = -\frac{\hbar^2}{2m_e}\sum_i^n \nabla_i^2 - \sum_i^n \sum_I^N \frac{Z_I e^2}{4\pi\varepsilon_0 r_{Ii}} + \frac{1}{2}\sum_{i,j}^n \frac{e^2}{4\pi\varepsilon_0 r_{ij}} \qquad 4.1$$

4.1 The General HF Approach

The repulsion felt between two electrons is significant and must be included in any accurate electronic structure method. However, evaluating the two electron integrals is inherently very difficult so various approximations are used. With one such approximation, the Hartree-Fock method (HF), electron-electron repulsions are treated in an average way with each electron considered to be moving in the field generated by the nuclei and an average field from the other *n-1* electrons. The spin orbitals that give rise to the lowest energy wavefunction are then found using variation theory.[d]

This procedure leads to the Hartree-Fock equations[22] for the individual spin orbitals. For spin orbital $\phi_a(1)$ where electron 1 has been assigned to spin orbital ϕ_a the equation is:

$$F_i\phi_a(1) = \varepsilon_a\phi_a(1) \qquad 4.2$$

where ε_a represents the spin orbital energy and F_i, if each spin orbital is expanded in the form of a set of basis functions as discussed in section 2, is the Fock matrix (*eq. 4.3*).

$$F = h + J - \frac{1}{2}K \qquad 4.3$$

[d] The precise nature of variation theory will not be covered here. For a good description the reader is referred to section 6.9 of "Molecular Quantum Mechanics" Edn. 3 by P.W. Atkins and R.S. Friedman.

where h is the core Hamiltonian, J_{ab} the coulomb matrix (*eq. 4.4*) and K_{ab} the exchange matrix (*eq. 4.5*).

$$J_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_b \middle| \phi_c \phi_d \right) \qquad 4.4$$

$$K_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_c \middle| \phi_b \phi_d \right) \qquad 4.5$$

Where D_{cd} is the density matrix (*eq. 4.6*).

$$D_{cd} = 2 \sum_{\mu} C_{d\mu} C_{c\mu}^{*} \qquad 4.6$$

Each spinorbital for the molecule under study can be found by solving an equation of the form given in *eq. 4.2* above using the corresponding Fock matrix F_i. However, F_i depends on the spinorbitals of all the other electrons present so the solution can only be found using an iterative procedure that stops when the solutions are self consistent, hence the name self-consistent field (SCF). This procedure, illustrated graphically in figure 4.1, involves taking a trial set of spinorbitals which are used to construct the Fock matrix. The HF equations are then solved to obtain a new set of spinorbitals which are then fed back into the Fock matrix and so on. The cycle is repeated until the pre-defined convergence criteria are fulfilled.

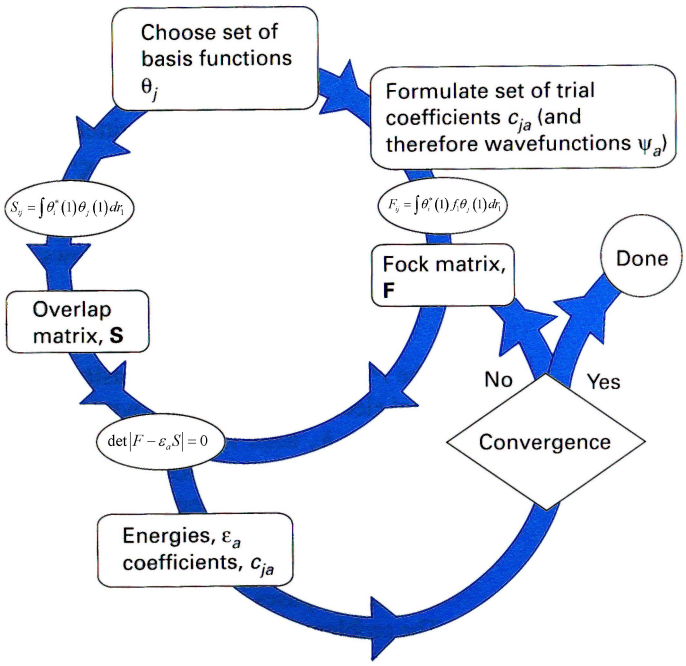


Figure 4.1: Schematic of the iterative procedure in the Hartree-Fock self-consistent field method. Adapted from Atkins, "Molecular Quantum Mechanics" Edn. 3, p. 283.

The number of spinorbitals that can exist are technically infinite but in practice it is necessary to be content with solving for a finite number *m*, where *m* must be greater than the total number of electrons present. The *m* optimised spinorbitals, obtained when the HF-SCF procedure is complete, are generally arranged in order of increasing energy with the *n* lowest energy orbitals being termed *occupied orbitals* and the remaining spinorbitals being termed *virtual orbitals*. From these spinorbitals can be found the ground state wavefunction for the molecule and the molecular orbitals. An example of the molecular orbitals obtained from the HF procedure are given in figure 4.2.

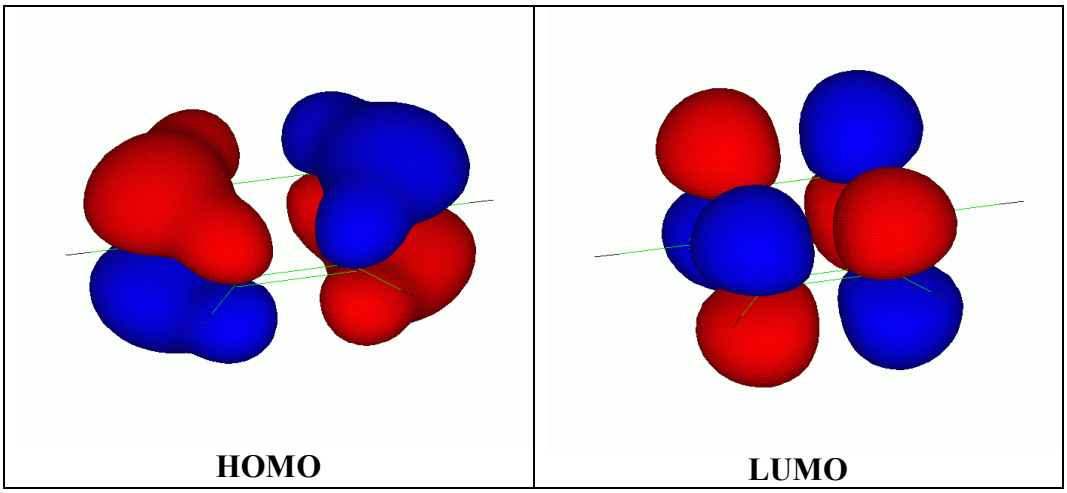


Figure 4.2: Example of highest occupied and lowest unoccupied molecular orbitals of benzene generated using the Hartree-Fock SCF procedure with the STO-3G basis set. Source:-[rd]

5. Scaling in HF-SCF Calculations

Solution of the HF equations by the linear combination of atomic orbitals (LCAO) method,[23,24] as discussed in section 4 above, leads to a set of equations, the complexity of which scales formally as N^4. Where N is the number of basis functions describing the molecule. The origin of this scaling is due to the computation of the two electron repulsion integrals (ERI's),[25] (*eq. 5.1*).

$$(\phi_a\phi_b|\phi_c\phi_d) = \iint dr dr' \phi_a^*(r)\phi_b(r)\frac{1}{|r-r'|}\phi_c^*(r')\phi_d(r') \qquad 5.1$$

which are required to construct the Fock matrix (*eq. 4.3*).

Hence the current major bottleneck in Hartree-Fock Self Consistent Field electronic structure calculations is the iterative construction of the Fock matrix. In the 1970's it was found that, while the number of ERI's grows formerly as N^4, in large systems a sizeable number are small so avoiding their calculation leads to algorithms that scale as $N^2 \text{Log } N$ or even N^2.[26,27](*fig. 5.1*).

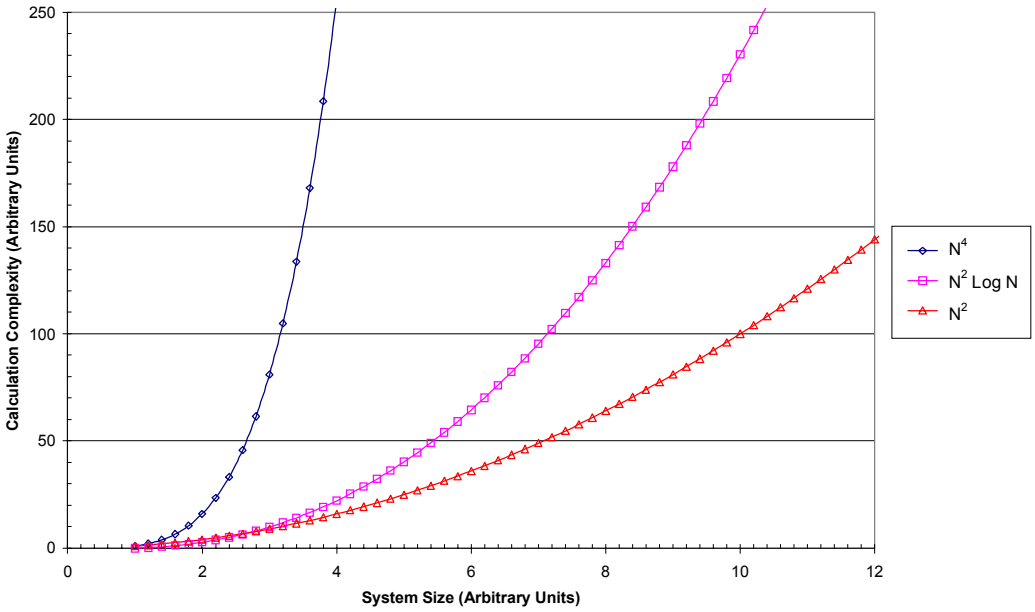


Figure 5.1: Arbitrary plot showing the scaling relationship between various power curves.

However, the size of an ERI is not always a good measure of its importance to the Fock matrix. Hence a method was developed that uses density weighted thresholding of the ERI's to determine which are relevant to the construction of the Fock matrix.[28] The use of these methods and other evaluation schemes has allowed direct SCF (a method whereby the pre-factors are calculated on the fly and then discarded rather than swapping to disk and then

retrieving when next needed) calculations on large molecular systems to obtain an N^2 dependency in construction of the Fock matrix.

Although other steps in the calculation procedure, scale greater than N^2, such as diagonalisation of the relevant matrices which scales as N^3, it is the construction of the Fock matrix which is the most computationally intense. Recent work has produced a direct method that gives a linear (N) routine for the minimisation of the density matrix D (*eq. 4.6*).[29] Hence large scale HF-SCF calculations are therefore limited by the calculation of the ERI's which at best can be made to scale as N^2.

5.1 The Problems of Non Linear Scaling

The problem with performing calculations that scale as a power dependence is that, in theory, there is a maximum system size that you can solve regardless of computer power. With N^2 scaling doubling the power of the computer used will only increase the complexity of problem that can be solved in a given time by a factor of $\sqrt{2}$. Thus by arbitrarily plotting the complexity of problem that can be solved in a given time against computer power (or time if computers are indeed doubling in power every 18 months) the problem becomes obvious (*fig. 5.2*).

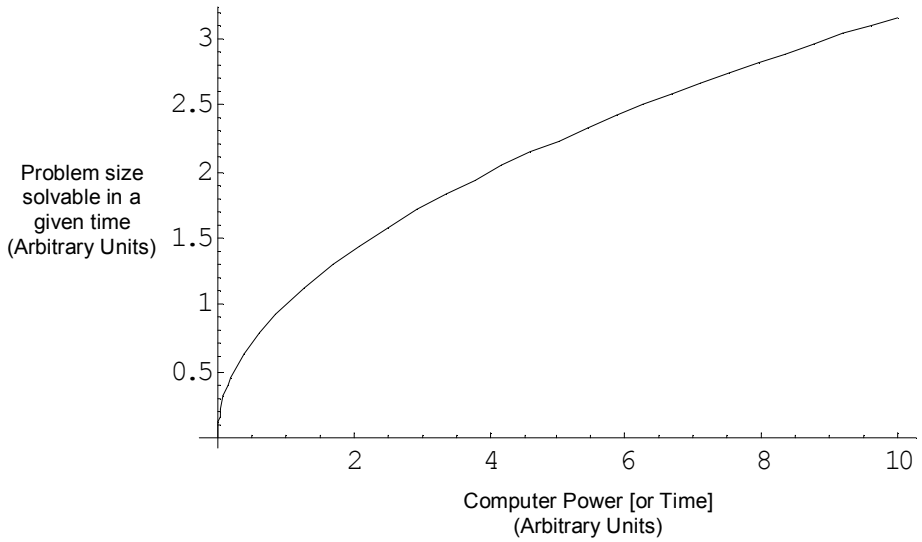


Figure 5.2: Arbitrary plot showing the problem of non linear scaling.

Hence the law of diminishing returns means that the curve becomes asymptotic at a given complexity thus unless calculation procedures, that give accurate results, can be devised that scale linearly with problem size there will come a point where, regardless of computer

power, it will not be possible to find the electronic structure, in a given amount of time, for systems greater than a given size.

There is therefore a large amount of research being directed at methods for computing the coulomb and exchange matrices that scale linearly. One promising area of research is in the use of multipole expansions which are discussed in section 6.

6. The Fast Multipole Method

Over the past 15 years there has been widespread activity on a new class of fast numerical methods for applying linear operators. Collectively these have come to be known as "*Fast Multipole Methods (FMM's)*".[30] The name *fast* originates from the idea of Fast Fourier Transform.

The essential property of the fast multipole method is that the sparse decompositions used to render applications of the operators are fundamentally approximate. Another way to state this is that the method effectively makes use of analytic, rather than algebraic, properties of the operators. The fact that the fast multipole method is fundamentally approximate is not, however, a limitation because, if correctly implemented, the errors can be made as small as necessary at a reasonable cost.[31]

6.1 History of FMM Methods

The mathematics underlying the new fast multipole methods is of 19th century origin hence one is bound to ask why the availability of such methods for sparse decompositions of linear operators has not been recognised until the close of the 20th century. The answer is due to the nature of the scaling of these methods. A method which grows slowly with problem size also, by virtue of the way it works, diminishes less rapidly as the problem size is decreased. Thus there is a minimum problem size, or break even point, at which using a linear scaling (fast) method becomes more economical than traditional slower methods. This is illustrated graphically in figure 6.1.

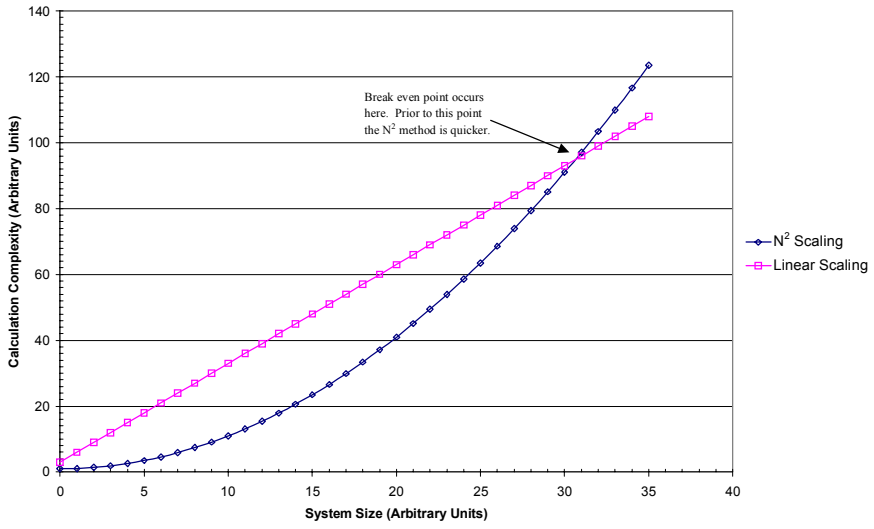


Figure 6.1: Arbitrary plot showing the concept of a break even system size.

In general it is only recently that computers have become powerful enough to study problems of break even size.[32] Thus prior to the 1990's fast multipole methods were purely academic curiosities that were uneconomical for solving problems of the time. It is only recently with the rise in computational power that such methods have become economically viable.

The first fast multipole method was originally devised to solve N-body problems using particle in cell methods.[33,34] Since then a number of variants on this approach have been made. One of the most useful in terms of solving electron repulsion integrals was published in 1990 in a paper entitled "Multilevel matrix multiplication and fast solution of integral equations" by A.Brandt & A.A.Lubrecht.[35] This and other work in the field encouraged computational chemists to begin to look at the ways in which FMM could be adapted to achieve linear scaling HF-SCF electronic structure calculations for large systems.

6.2 General FMM Theory[e]

The fast multipole method was originally devised for calculating interactions between a system of classical particles interacting via two-body forces.[36] This is an *N*-body problem which requires, for a direct summation over all pairs, N^2 work.[37] The fast multipole method works by splitting the problem into near and far fields. The near field is calculated exactly (by direct SCF methods) while the far field is divided into a number of boxes where the interaction between all the charges in one box and all the charges in another is represented by the interaction between two multipoles centred in each box.

As the distance between boxes increases it is possible, for a given level of accuracy, to use larger and larger boxes (*fig. 6.2*). Thus for larger systems where the far field is substantial in comparison to the near field the work required is reduced from N^2 scaling to something which approaches linear scaling with respect to problem size.

[e] The mathematics underlying the fast multipole method will not be discussed in detail in this report. For information on the precise details of the mathematics involved the reader is referred to references 36 and 39.

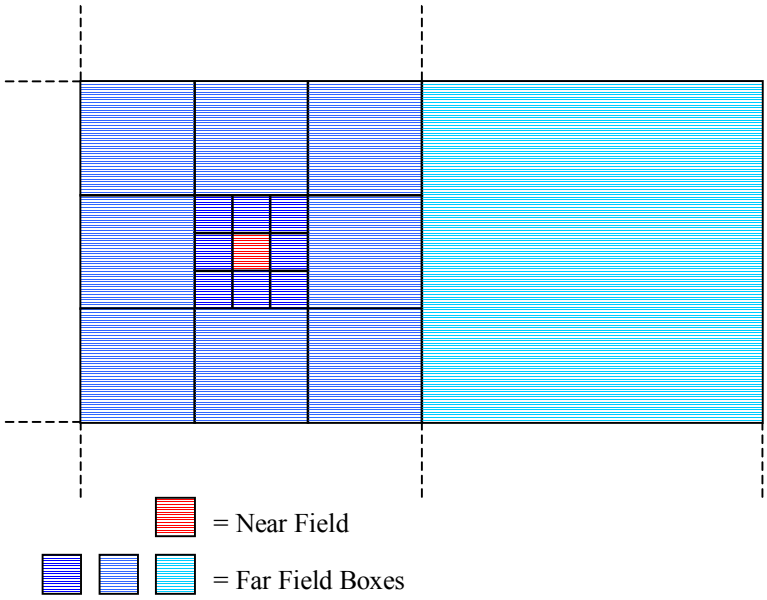


Figure 6.2: Illustration of the hierarchical box structure of the fast multipole method. Adapted from F. Jensen, "Introduction to Computational Chemistry", 1999 p. 387.

The size of the errors introduced into the calculations can be adjusted by varying the size of the boxes and the length at which the multipole expansion representing the interaction between multipoles is truncated.

This method, while easily applied to systems of point charges, is not immediately applicable to quantum chemical calculations because electrons are represented by probability distributions in quantum mechanics rather than discrete particles as in classical mechanics. Hence a large amount of work was required to adapt the original fast multipole method of Greengard and Rokhlin[38] for use in HF-SCF calculations.

6.3 The Coulomb and Exchange Problems

Research in this field has effectively followed two paths. Early work addressed the so called "electronic quantum coulomb problem" which is related to the formation of the coulomb matrix (*eq. 6.1)*.

$$J_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_b \middle| \phi_c \phi_d \right) \qquad 6.1$$

This component of the Fock matrix deals with the electrostatic repulsions between electrons and, due to its $\frac{1}{r}$ dependence on the distance between two electrons, has an effect that extends a substantial distance before becoming negligible (*fig. 6.3*).

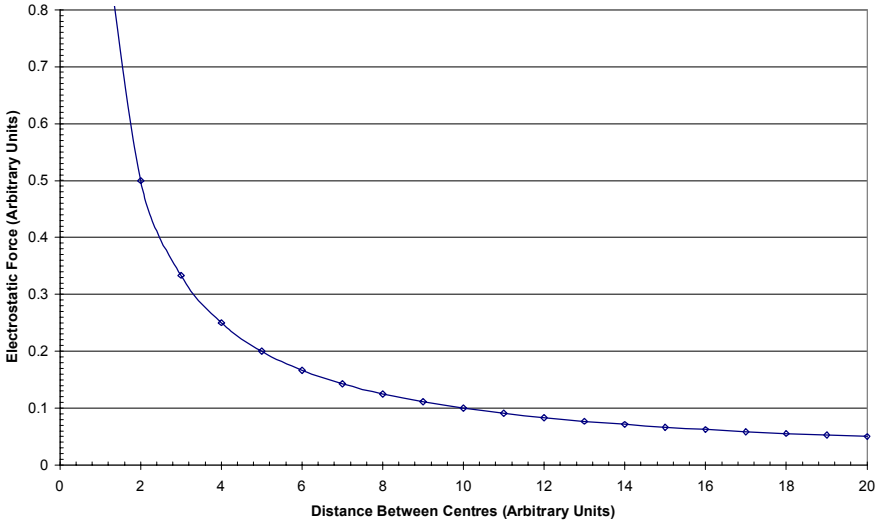


Figure 6.3: Arbitrary plot showing the 1/r scaling relationship of the coulomb interaction.

This problem therefore has to be addressed in a different way to the second component of the Fock matrix, the exchange matrix (*eq. 6.2*).

$$K_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_c \middle| \phi_b \phi_d \right) \qquad 6.2$$

The exchange matrix is essentially responsible for the formation of chemical bonds in molecules as it represents the exchange of electrons between bonded atoms. The nature of this matrix is very different to the coulomb matrix because the effect, in non metallic systems, is very localised. Hence this has to be treated in a different way to the coulomb problem. More recent work has therefore been directed towards developing methods for fast assembly of the exchange matrix which, when coupled with fast builds of the coulomb matrix, will yield a linear scaling solution to the construction of the Fock matrix.

Since the two problems are very different and hence have been treated as essentially separate entities they will be covered separately in the following two sections.

7. The Electronic Quantum Coulomb Problem

7.1 The N^2 Limit

The N^2 limit for direct SCF calculations can be explained in terms of the Gaussian product theorem which states that an uncontracted distribution $\rho_{ab} = \phi_a \phi_b$ can be expressed as a finite sum of cartesian Gaussian type functions with exponent $\zeta_p = \zeta_a + \zeta_a$.[40] E.g. the product of two s-type Gaussians is:-

$$\exp\left[-\zeta_a (r-A)^2\right] \times \exp\left[-\zeta_b (r-B)^2\right] = \exp\left[-\xi (A-B)^2\right] \times \exp\left[-\zeta_p (r-P)^2\right] \qquad 7.1$$

As the size of the system increases the radial overlap ($\exp\left[-\xi (A-B)^2\right]$) falls off exponentially with the distance between A and B and the number of significant distributions to the density approaches N from above. Similarly the number of coulomb matrix elements larger than a certain threshold also increases as N. Hence computation of the coulomb matrix by conventional methods is at best N^2 and typically around $N^{2.7-3}$.

7.2 The Multipole Expansion in Cartesian Co-ordinates

As system size increases the dominant component in the two electron energy term becomes due to simple electrostatics. This electrostatic energy is given, when the overlap of the two densities is minimal, by equation 7.2

$$E_{el}^{AB} = \int dr \int dr' \, \rho_A (r) |r - r'|^{-1} \rho_B (r') \qquad 7.2$$

This can be approximated accurately by expressing $|r-r'|^{-1}$ in the form of a multipole expansion. In cartesian co-ordinates this is in the form of a Taylor series expansion giving equation 7.3.

$$E_{el}^{AB} = \sum_{LMN} \sum_{L'M'N'} (-1)^{L'+M'+N'} Q_{LMN}^{A} T_{L+L',M+M',N+N'}^{AB} Q_{L'M'N'}^{B}$$

where *7.3*

$$Q_{LMN}^{A} = \frac{1}{L!M!N!} \int dr\, x^{L} y^{M} z^{N} \rho_{A}(r)$$

Q represents an unmodified cartesian multipole moment. Truncation of this expansion allows the electrostatic energy to be evaluated to a desired degree of accuracy determined by the length of the expansion used.

7.3 A Practical Implementation of Multipole Methods for Solving the Electronic Quantum Coulomb Problem

An example of early work using this approach is discussed in a paper entitled "Achieving linear scaling for the electronic quantum coulomb problem" by M. Strain *et al.*[41] In this paper they report that:

> "a generalisation of the fast multipole method to Gaussian charge distributions achieves near linear scaling for the quantum coulomb problem."

They also report that:

> "the method becomes faster that standard analytic evaluation of Gaussian two electron integrals for systems containing as few as 300 basis functions."

The method employed by Strain *et al.* works by embedding the system under consideration in a hierarchy of 8^n cubic boxes, as discussed in section 6.2, where n represents the number of tiers used (in this case $N < 7$). All charge distributions in a given box were then represented by multipole expansions about the centre of the box. The near-field section, defined as interactions within a given box and between neighbouring boxes, was treated exactly while the far-field was treated through multipole expansions.

Since the charge distribution in quantum chemical problems is a continuum it is necessary to define the range of a charge distribution. Truncating the distribution too early leads to

large errors in the results while truncating it at a much longer distance leads to the need for a large amount of extra computation for negligible improvement in accuracy.

This problem was addressed by Strain *et al.* using equation 7.4 which represents the range definition derived from the basic coulomb integral between two s-type functions.[42]

$$r = (2s)^{-\frac{1}{2}} Erfc^{-1}(\varepsilon) \qquad 7.4$$

Here *s* represents the exponent of the product Gaussian distribution and *Erfc* represents the error function ε which is the desired level of accuracy. A given interaction is then only included in the far field if the number of boxes between interactions is greater than the sum of the ranges of the distributions.

The second problem is deciding at which point to truncate the multipole expansion since, being derived from a Taylor series, it effectively has infinite terms. As with the charge distribution above, truncating too early gives large errors while truncating at a higher number of terms necessitates unnecessary work. This problem was addressed by Strain *et al.* using equation 7.5.

$$\varepsilon = k\left(\frac{a}{R}\right)^{l_{eff}} \qquad 7.5$$

where l_{eff} is the length at which the expansion is truncated, R is the number of boxes, a is a constant, k is an adjustment factor and ε, as above, is the desired accuracy.

It was found that using values of $l_{eff} = 12$ and $\varepsilon = 10^{-6}$ gave errors of $\approx 10^{-6}$ Hartrees for the coulomb energy which is small enough to have negligible effect on the chemistry under study.

The computational times required for calculating the entire (NF+FF) coulomb matrix with standard analytic two electron integral evaluation between contracted Gaussian functions and calculation using fast multipole methods (the specific method developed was termed GvFMM in the paper) were compared for a range of graphitic sheets of the form $C_{6m^2}H_{6m}$

for m = 1 to 8. These sheets were modelled using the 3-21G basis set with $\varepsilon = 10^{-6}$, $l_{eff} = 12$ and the number of boxes set at 8^5 for the FMM calculations.

It was found that the fast multipole calculations became competitive with analytic integration when 80 % of the interactions were treated in the far field, in this case for as few as 300 basis functions ($C_{24}H_{12}$). The results obtained are shown in figure 7.1 where each of the curves have been fitted with an equation of the form N^{Ω}.

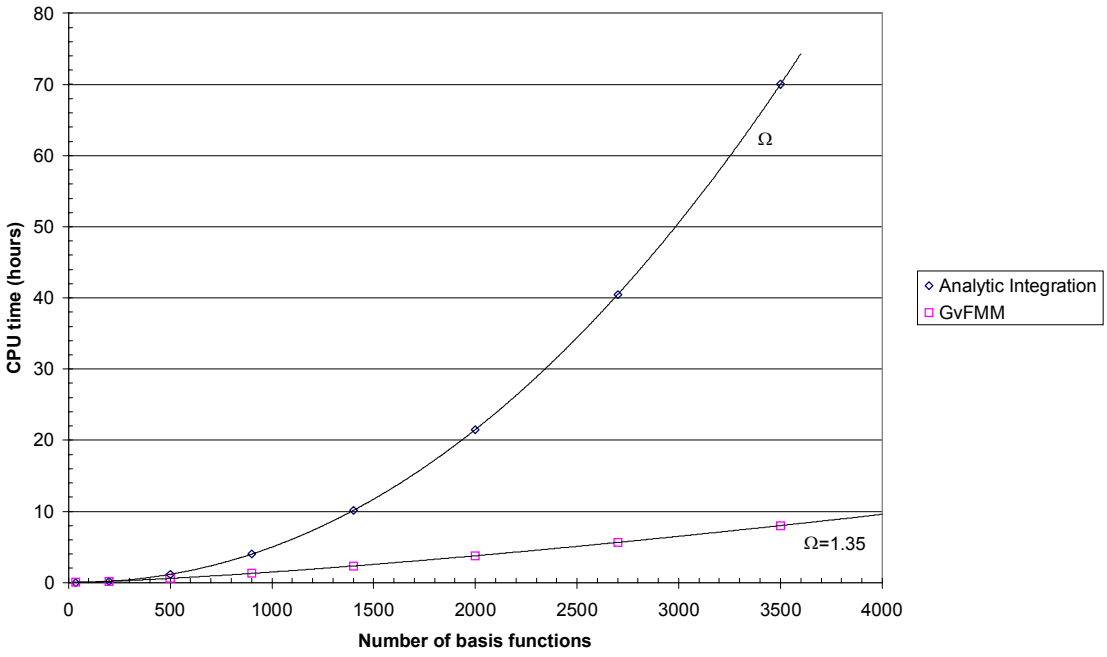


Figure 7.1: CPU Times (IBM/RS6000-370) for the formation of the coulomb matrix (first iteration of the SCF procedure) for a series of graphitic sheets using 3-21G basis set with analytic and GvFMM calculations. Adapted from M.C. Strain *et al. Science*, 271, **1996**, 51.

The effective scaling for analytic integration was found to be ≈ 2.11 while for the GvFMM it was found to be only ≈ 1.35. This is substantially lower than the quadratic behaviour for analytic integration. Thus for the largest calculation carried out ($C_{384}H_{48}$ ≈ 3500 basis functions) the GvFMM was over eight times faster than direct methods and with an absolute error of only ≈ 10^{-6} Hartrees.

A breakdown of the NF and FF components of the coulomb problem calculated with GvFMM for 8^5 boxes showed that the computational time was dominated by the analytic integration of the NF portion (*fig. 7.2*) even though for $C_{384}H_{48}$ it represented only 3 % of the total interactions.

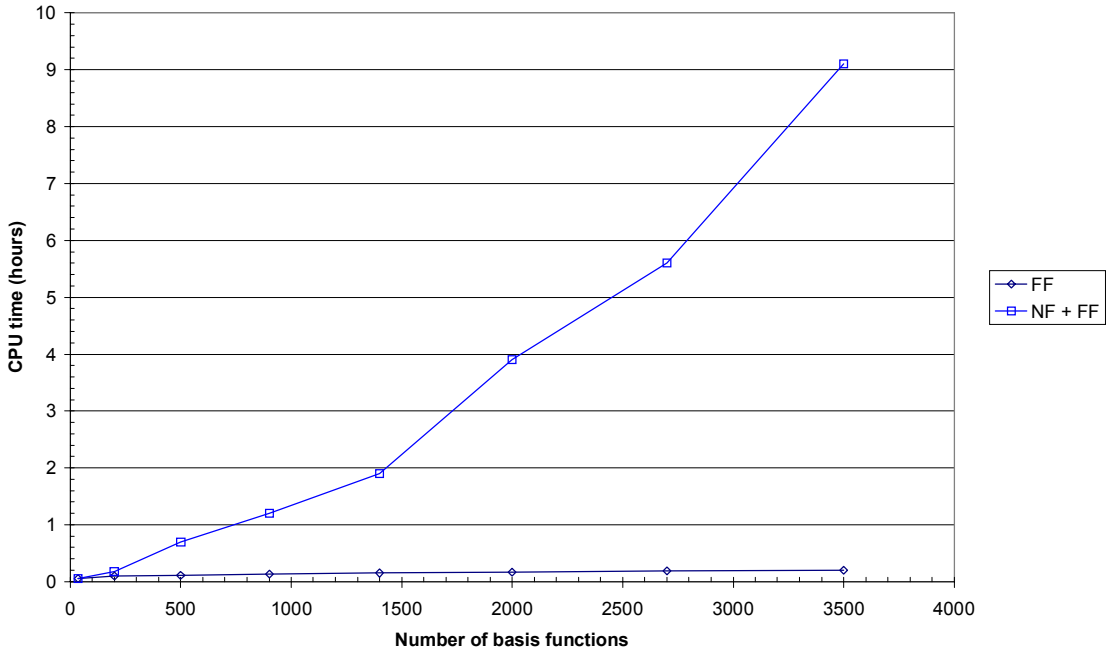


Figure 7.2: CPU Times (IBM/RS6000-370) for computing the NF and FF components of the coulomb matrix for a series of graphitic sheets using 3-21G basis set with analytic and GvFMM calculations. Adapted from M.C. Strain *et al.* *Science*, 271, **1996**, 51.

Hence in the large molecule limit where the ratio of NF interactions to FF interactions becomes ever smaller the computational time required for FMM calculations of the coulomb matrix should approach linear scaling.

The authors conclude that while the results obtained were limited to graphene-sheets the procedures used are equally valid for complex materials and that given the speed, accuracy and scaling properties of the GvFMM in practical, high accuracy, calculations the method is very promising for future electronic structure calculations on large molecular systems.

Further work on the problem has refined the technique by adjusting the size of the hierarchy of boxes used and the point at which the expansion is truncated to obtain more accurate results. Strain *et al.* have also gone on to expand the GvFMM method for the calculation of Kohn-Sham analytic energy second derivatives in density functional theory.[f]

The major problem, however, with the GvFMM method is that it is unlikely to scale as favourably with 3 dimensional systems since the separations between atoms grow much more slowly in 3-d.[43] Hence while a fast method for 2-d structures, such as graphitic sheets, the work of Strain *et al* will not necessarily give linear scaling when applied to complex molecules such as proteins.

[f] This topic will not be covered in this report. For further details the reader is referred to references 44 & 45.

Work by M. Challacombe and E. Schwegler, who were the first to demonstrate true linear scaling for 3 dimensional systems, found that if Hermite Gaussian type functions (*eq. 7.6*) are used instead of cartesian Gaussian type functions the complexity of the FMM calculations can be greatly reduced.[46]

$$\Lambda^{p}_{LMN}(r) = \frac{\partial^{L}}{\partial P_x^{L}} \frac{\partial^{M}}{\partial P_y^{M}} \frac{\partial^{N}}{\partial P_z^{N}} \exp\left[-\varsigma_p (r-P)^2\right] \qquad 7.6$$

Use of this idea and appropriate thresholding with a quantum chemical tree code that uses a multipole and penetration acceptability criterion to determine near and far field on the fly gives linear scaling for construction of the coulomb matrix. The tree code (*fig 7.3*) works by starting with the root (parent) cell which contains all of the particles. Space is then sub divided in a recursive manner into smaller cells (children). Computation is then simplified by representing the particle distributions within each cell as a multipole expansion.

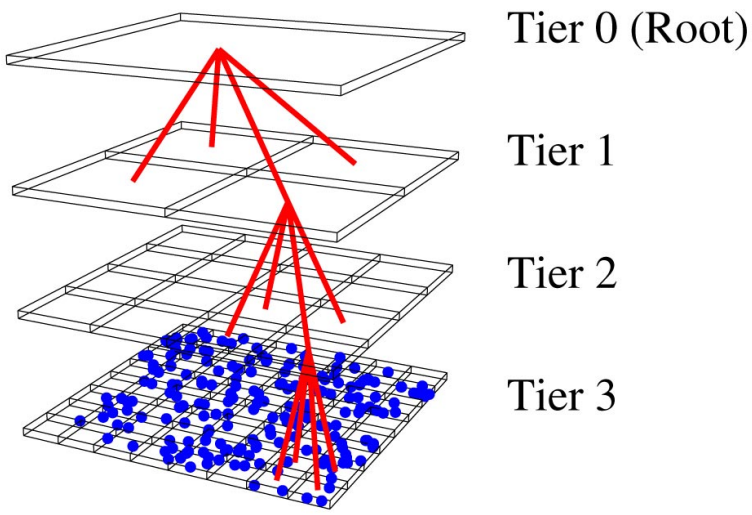


Figure 7.3: Illustrative example of the tree-code data structure showing how space is sub divided into a hierarchy of cells. Kindly provided by M.Challacombe, *personal correspondence.*

In Challacombe *et al.*'s implementation of this method[47] the tree code represents the electron density using a hierarchical multipole representation. The tree is then traversed for each of the matrix elements and a multipole acceptability criterion and penetration acceptability criterion to ascertain if the error that arises from the multipole approximation is acceptable. If the error falls below a certain threshold and is thus acceptable the interaction of the "bra" charge distribution is calculated using the multipole representation at that level of resolution. If the error is greater than a certain threshold it is considered unacceptable and sub-branches of the tree continue to be traversed, going to greater and

greater resolution until the errors are acceptable. Any near field interactions are then finally evaluated accurately using a modified McMurchie-Davidson integral code.[48]

Challacombe and Schwegler tested the scaling properties of their method by performing SCF calculations at the RHF/3-21G level of theory on a range of water clusters (*fig. 7.4*).

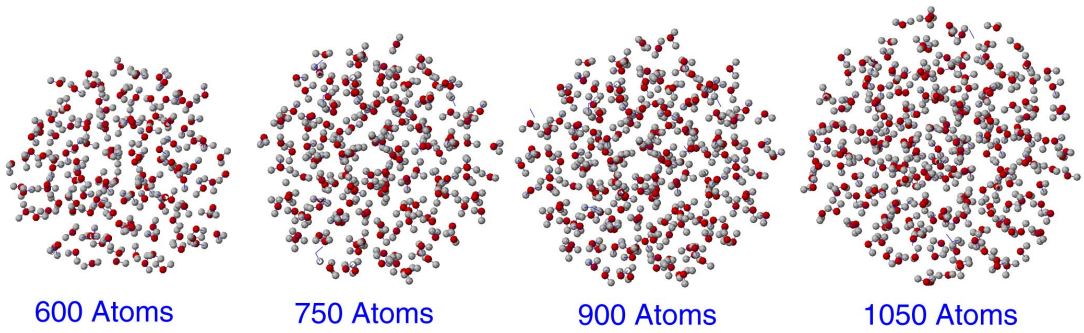


Figure 7.4: Example configuration of water clusters used for testing FMM calculations. Adapted from M. Challacombe *et al. J. Chem. Phys.* 110, **1999**, 2332.

The results obtained for construction of the coulomb matrix (***J*** matrix) are given in figure 7.5 showing the comparison between direct ***J*** builds and *fast* ***J*** builds.

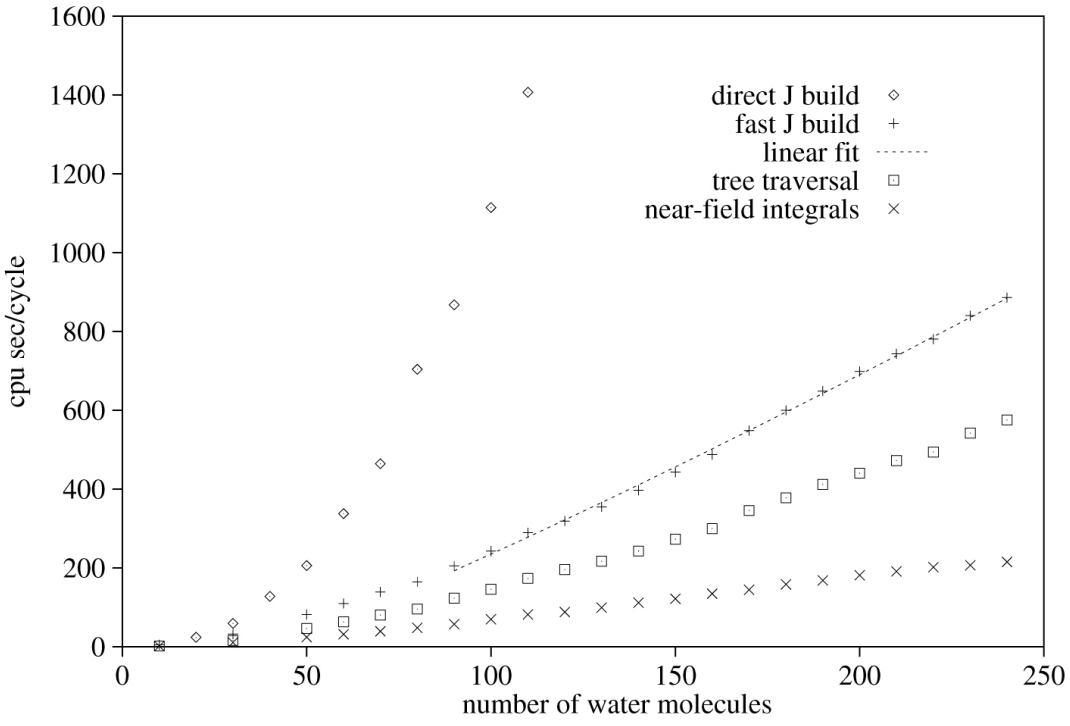


Figure 7.5: CPU Times (IBM SP2 66 MHz, 256 Mb RAM) for the formation of the coulomb matrix by direct and fast methods on a sequence of water clusters. Adapted from M. Challacombe *et al. J. Chem. Phys.*, 106, **1997**, 5526.

Figure 7.5 shows that for more than ≈ 80 water molecules the quantum chemical tree code-FMM method of building the ***J*** matrix scales linearly with system size.

The errors in the converged total energies of the clusters were kept as small as those for direct methods indicating that this method for building the coulomb matrix, while giving linear scaling, is suitably accurate for practical computational modelling. Thus linear scaling of the electronic quantum coulomb problem has been achieved and the next step in achieving linear scaling for building the Fock matrix is the construction of the exchange matrix (***K*** matrix) covered in section 8.

8. Linear Scaling Exchange Matrix Builds

Since exchange interactions are, in non metallic systems, very short range in comparison to coulomb interactions the methods used for fast evaluation of the coulomb matrix are not directly applicable to building the exchange matrix with linear scaling. Hence recent work by E. Schwegler, M. Challacombe and others has centred on methods for achieving linear scaling for the exchange problem which when coupled with fast methods for J matrix builds and a simplified density matrix minimisation[49] will allow building of the Fock matrix in a way that scales linearly with problem size.

The first specialised methods to be developed for building the exchange matrix worked by truncating the density and exchange matrix with distance dependence cut-offs.[50] Unfortunately these methods at best scale as N^2.

The first linear scaling method, termed ONX (order N exchange), was developed by Schwegler *et al*.[51] and is based on the assumed exponential decay of the density matrix contributing to the exchange matrix (*fig 8.1*).

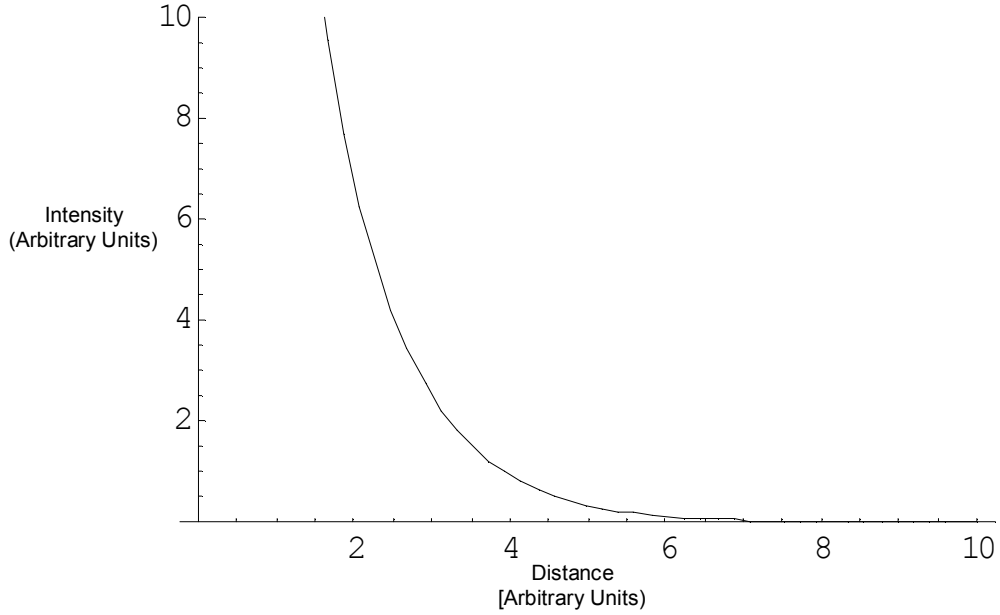


Figure 8.1: Arbitrary plot showing the concept of exponential decay with distance.

This method calculates the insignificant elements of $\boldsymbol{K}$, and so simplifies the calculations, by thresholding using equation 8.1.

$$\tilde{K}_{ab} = \int dr \int dr' \phi_a(r) \phi_b(r') |r - r'|^{-1} e^{\frac{-|r-r'|}{l}} \qquad 8.1$$

The success of these methods, however, is dependent on a knowledge of the density matrix decay to achieve a balance between accuracy and speed. Central is the assumption that the exchange interactions become negligible when the separation of basis functions exceeds a certain thresholding parameter.

An improved version of this method has since been developed[52] that uses a rigorous method for linear scaling computation of the exchange matrix and does not assume anything about the decay of the density matrix. Thus if given an insulating system, such as a protein, which has approximate exponential decay the computation time scales linearly, as in figure 8.2 below, while if given a metallic density matrix, as in graphite, the calculation naturally reverts to quadratic scaling (*fig 8.3*).

The main breakthrough in this later version of ONX was the realisation that the integral screening scales as N^2, while the evaluation of the integrals scales only as N. Thus methods were developed that sort the integral estimates in such a way that the screening overhead is also reduced to linear scaling.

8.1 Multipole Accelerated Exchange

Work by E. Schwegler and M. Challacombe has found that an adaptation of their ONX method that uses multipole approximations can give an extremely competitive approach for computation of the HF exchange matrix of large systems.

When computing the exchange matrix contracted Gaussian type functions of the form given in equation 8.2 are used.

$$\phi_a(r) = \sum_i^{k_a} d_{ai} \Psi_{ai}(r) \qquad 8.2$$

However, this leads to four fold contraction loops (*eq. 8.3*) when evaluating the electron repulsion integrals which are computationally very expensive.

$$(\phi_a \phi_b \,|\, \phi_a \phi_b) = \sum_i^{K_a} \sum_j^{K_b} \sum_k^{K_c} \sum_l^{K_l} d_{ai} d_{bj} d_{ck} d_{dl} (\Psi_{ai} \Psi_{bj} \,|\, \Psi_{ck} \Psi_{dl}) \qquad 8.3$$

It has been known for some time that it is possible to move a fraction of the electron repulsion integral evaluation outside of the contraction summations. It is also possible to reduce the average contraction length ($\tilde{K}$) of the basis function products, by pre-screening, to:

$$\phi_a(r)\phi_b(r) = \sum_i^{K_a}\sum_j^{K_b} \Psi_{ai}(r)\Psi_{bj}(r) \qquad 8.4$$

This, however, still results in an amount of work that scales as $\tilde{K}^4$.

Schwegler *et al.* have found that by decoupling distributions $(\phi_a\phi_b|$ and $|\phi_c\phi_d)$ that are well separated and using a multipole approximation enables the contractions to be performed independently of the ERI. Multipole approximations, used in conjunction with a multipole acceptability criterion[53] which controls errors due to truncation of the expansions, can be used with the ONX method to eliminate much of the work associated with basis set contraction and so offer an extremely competitive approach to computation of the HF exchange matrix for both large systems and highly contracted basis sets.[54]

The extent to which formation of the exchange matrix can be accelerated by using multipole approximations is dependent on two factors. The first is the percentage of all interactions that can be accurately calculated in a multipole representation. The second is the relative speed of calculating and interaction with multipole versus direct ERI evaluation.

8.2 A Practical Example of Multipole Accelerated Exchange

In a paper entitled "Linear scaling computation of the Fock matrix IV. Multipole accelerated formation of the exchange matrix"[55] Schwegler *et al.* report the results of exchange matrix builds, for a series of water clusters and graphitic sheets, using multipole accelerated versions of their order N exchange (ONX) and symmetrised ONX (SONX).[g]

[g] ONX and SONX are linear scaling methods for computing the HF exchange matrix.

The results obtained (*figs. 8.2 & 8.3*) show that the multipole accelerated exchange methods scale linearly with system size and show between a 4 and 5 fold increase in speed over non accelerated methods (ONX and SONX). The scaling of the ONX routine can be seen to be linear for an insulating system (*fig. 8.2*) and quadratic for a conducting system (*fig. 8.3*).

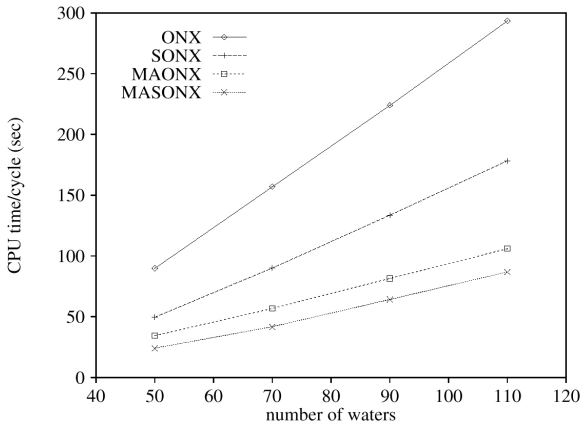


Figure 8.2: CPU Times (PPC 604e 322 MHz) calculation of HF-exchange matrix by direct and fast methods on a sequence of water clusters (RHF/STO-3G). Adapted from M. Challacombe *et al.* Accepted, Pre-print LA-UR-99-578.

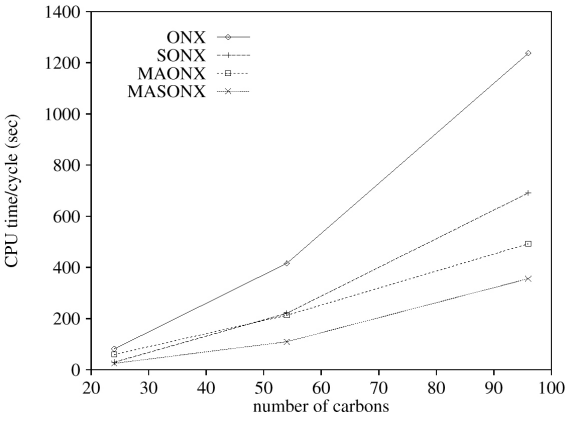


Figure 8.3: CPU Times (PPC 604e 322 MHz) calculation of HF-exchange matrix by direct and fast methods on a sequence of graphitic sheets (RHF/STO-3G). Adapted from M. Challacombe *et al.* Accepted, Pre-print LA-UR-99-578.

The authors also present results of error analysis (*fig. 8.4*) that shows that:

> "the errors incurred by MAONX are indistinguishable from the integral pre-screening errors associated with the Schwartz inequality used in ONX."

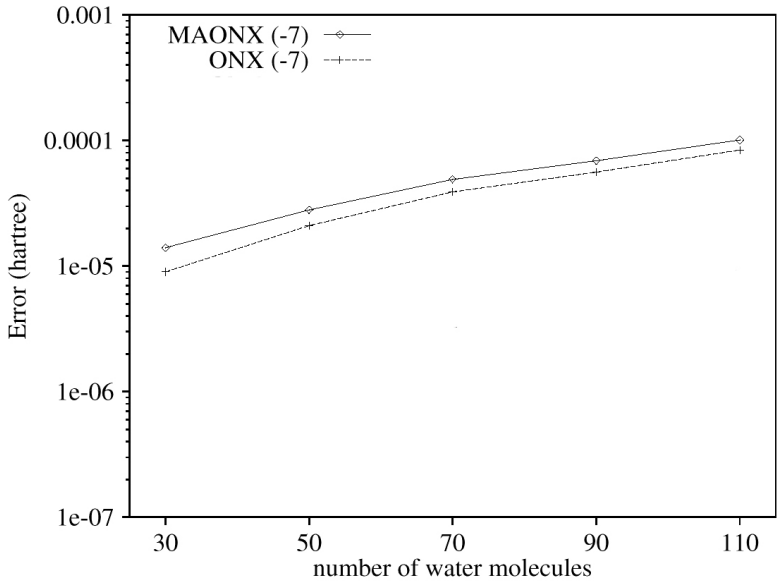


Figure 8.4: Absolute errors in the converged total energy for a series of water clusters. Adapted from M. Challacombe *et al.* Accepted, Pre-print LA-UR-99-578.

The authors conclude that:

> "Implementation [of multipole expansions] in the linear scaling methods MAONX and MASONX for computing the Hartree-Fock exchange matrix indicate that large computational savings are possible when tightly contracted basis sets are used."

MAONX was found to be 4.6 times faster that ONX for exchange matrix builds of water clusters and graphitic sheets with highly contracted basis sets. With less contracted basis sets the speedup is observed to be ≈ 2 times. Thus:

> "…multipole acceleration is expected to greatly increase the efficiency of linear scaling computations of the Fock matrix when highly contracted basis sets are used."

9. Linear Scaling Fock Matrix Builds

As stated in section 4.1 the Fock matrix (*eq. 9.1*) is composed of h, the core Hamiltonian, J the coulomb matrix and K the exchange matrix.

$$F = h + J - \frac{1}{2}K \qquad 9.1$$

By using the principles discussed in sections 7 and 8 it has been possible to formulate a method that gives linear scaling for Fock builds with acceptable levels of accuracy that removes the current major bottleneck in HF-SCF calculations making possible electronic structure calculations on systems of previously unprecedented size.

9.1 Example Calculations using Linear Scaling HF-SCF Theory.

The procedures discussed are very much still in their infancy so most published calculations have only been done on sequences of water clusters, polyglycine chains or graphitic sheets. While useful for calibration and error control these examples are not really of great chemical interest.

The only people to have published calculations involving systems of chemical interest using these methods are M. Challacombe and E. Schwegler who carried out calculations at the RHF/3-21G level of theory on several proteins of interest including endothelin (EDP), charybdotoxin (CRD) and the tetramerization monomer (P53).[56]

The results obtained are given in table 9.1 below.

Table 9.1: Results of single point MONDO RHF/3-21G calculations on selected proteins. Adapted from M. Challacombe *et al. J. Chem. Phys.*, 106, **1997**, 5526.

Protein	Energy (Hartree)	Atoms	Basis Functions	Time to compute K (hours)	Time to compute J (hours)
EDP	-8154.72370 -8154.72323[1]	255	1461	0.85	0.11
CRD	-16737.90643	572	3237	2.70	0.38
P53	-17115.30474	698	3836	2.10	0.38

[1] Energy obtained with GAUSSIAN 94 default settings.

From the two energies given for EDP it can be seen that the linear scaling methods, while allowing fast calculations, give very good accuracy in the results.

Figure 9.1 shows the iso-surfaces of the P53 tetramerisation monomer electrostatic potential, obtained via linear scaling SCF theory.

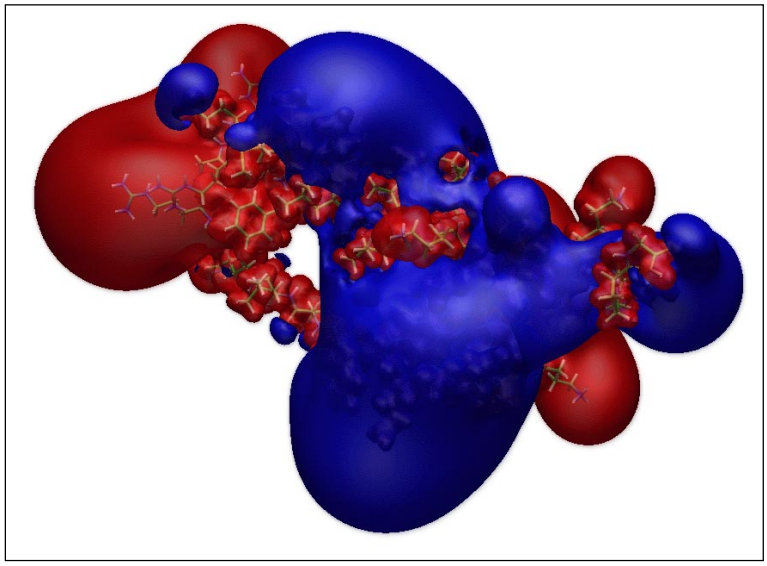


Figure 9.1: Iso-sufaces of the P53 tetramerisation monomer electrostatic potential at the RHF/3-21G level of theory. Obtained using linear scaling SCF theory. Adapted from M. Challacombe *et al. J. Chem. Phys.*, 106, **1997**, 5526.

The P53 calculation, involving a total of 698 atoms and 3836 basis functions, is the largest published Hartree-Fock calculation to date.

Interest in the P53 protein comes from cancer research. P53 is a tumour suppresser, mutations of which are the most frequently observed genetic alterations in human cancer. Thus application of linear scaling HF theory to problems of this type could prove very useful in future medical research.

10.Conclusion

As discussed in section 5.1, current methods for finding the electronic structures of systems, using *ab initio* methods, are hindered by their steep scaling. This places severe limitations on the size of systems that can be modelled. For example a large amount of interest is currently directed towards understanding the methods by which various proteins work in the human body. Modelling of such proteins in the gas phase is currently at the very limit of what is achievable using conventional methods. While the expected increases in computer power will make modelling of reasonable size proteins possible in the near future such calculations will be restricted to gas phase molecules as attempting to incorporate solvent effects will lead to calculations too complex to solve in a reasonable amount of time. Unfortunately the way proteins fold is very important in terms of their reactivity. Hence modelling in the gas phase can yield only limited information.

The examples discussed in this report show that the use of multipole expansions for Hartree-Fock calculations can reduce the scaling for Fock matrix builds to linear, whilst maintaining acceptable accuracy.

Hence, although only in their infancy and yet to be properly implemented in the main commercial quantum chemistry packages, these methods would appear to have a bright future unlocking the way towards modelling of ever more complex systems.

The major bottleneck in HF-SCF calculations would therefore appear to have been removed such that the next limiting factor is diagonalisation of the Fock matrix which scales as N^3. A method which solves the SCF equations in a way that only requires linear scaling CPU time has recently been developed[57] and applied successfully to record breaking RHF/STO-3G calculations with 2000 atoms and 6000 basis functions.[58] This is well beyond the current 5000 by 5000 limit (for serial calculations) with standard methods due to current memory contraints. Thus it would appear that for the HF level of theory quantum chemistry has finally found it's "Holy Grail", linear scaling with system size. The next stage is to go beyond the HF level of theory to develop linear scaling methods for the HF/DFT level which is very good at modelling biological type systems.

10.1 Parallelising of Calculations

The next major hurdle to be overcome in linear scaling SCF theory, before it can realise its full potential, is the conversion to code that can be executed in parallel.

All the examples given in this report were calculated in serial on single processors. Unfortunately the future of fast computing is likely to be in the form of larger and larger distributed arrays of parallel processors using shared memory. Hence all the time the implementations only work on single processors the size of system that can be studied is severely limited.

Greengard discussed the problem of parallel implementation of his original FMM method as follows[59]:-

> "…because all *fast* methods rely on *rearranging* a computation to reduce the operation count, efficient parallel implementation is a daunting problem. That's because transporting the data fast enough that processors are not left idle can be very difficult"

Therefore most research is currently being directed towards efficient parallel implementations of linear scaling SCF theories. Work currently being conducted by Challacombe *et al*. on parallel implementations of the above methods is showing promising results. Speeds of up to 4×10^9 floating point operations per second are being achieved for the minimisation step on 32 nodes of an SGI origin 2000 array. Papers on this field should be forthcoming in the near future.[60]

When this problem is overcome, with the rise in computing power making *fast* methods more and more economical over direct methods, the use of multipole expansions in electronic structure calculations is likely to prove to be a very important development making possible the accurate study of systems of previously unprecedented size.

References

1. M. Challacombe, http://www.t12.lanl.gov/~mchalla/.

2. M. Challacombe & E. Schwegler, "Linear scaling computation of the Fock matrix", *J.Chem.Phys.*, 106, 13, **1997**, pp. 5526-5536

3. Royal Swedish Academy of Sciences - 1998 Nobel Prize Press Release.

4. Top 500 Supercomputer List - http://www.top500.org/

5. Æ. Frisch, Gaussian Inc., "Exploring Chemistry with Electronic Structure Methods", Edn. 2, J.B. Foresman, **1993**.

6. Royal Swedish Academy of Sciences - "Additional Background Material on the Nobel Prize in Chemistry 1998".

7. F. Jensen, "Introduction to Computational Chemistry", Wiley, **1999**.

8. L. Greengard & V. Rokhlin, "A fast algorithm for particle simulations", *J. Comp. Phys.*, 73, **1987**, 325.

9. The historical background described here was adapted in part from
   1) The Royal Academy of Sciences - 1998 Nobel Prize Press Release.
   2) The Royal Academy of Sciences - Additional Background Material on the 1998 Nobel Prize.
   3) F. Jensen - Introduction to Computational Chemistry.
   4) A. Rae - Quantum Physics: Illusion or Reality?
   5) O. Frisch - What Little I Remember.

10. A. Rae., "Quantum Physics: Illusion or Reality?", Canto, **1986**.

11. *Op Cit.*[3]

12. C. C. J. Roothaan, *Rev. Mod. Phys.,* 23, **1951**, 69.

13. J. C. Slater, *Phys. Rev.*, 36, **1930**, 57.

14. S. F. Boys, *Proc. R. Soc. (London)*, A, 200, **1950**, 542.

15. *Ibid.*

16. P. Hohenberg & W. Kohn, *Phys. Rev. B*, 136, **1964**, 864.

17. W. Kohn & L. J. Sham, *Phys. Rev. A*, 140, **1965**, 1133.

18. P. W. Atkins & R. S. Friedman "Molecular Quantum Mechanics", Oxford, Edn. 3, **1997**.

19. G. H. Grant & W. Graham Richards, "Computational Chemistry", OCP, **1996**.

20. *Op Cit.*[13]

21. P. W. Atkins & R. S. Friedman "Molecular Quantum Mechanics", Oxford, Edn. 3, **1997** p. 276.

22. P. W. Atkins & R. S. Friedman "Molecular Quantum Mechanics", Oxford, Edn. 3, **1997** Further Information Section 11, pp. 499 - 502.

23. *Op Cit.*[12]

24. J. A. Pople and R. K. Nesbet, "Self-consistent orbitals for radicals", *J. Chem. Phys.*, 22, **1954**, 571.

25. E. R. Schwegler, Thesis, October 21, **1998**.

26. V. Dyczmons, "No N^4 dependence in the calculation of large molecules", *Theoret. Chim. Acta.*, 28(3), **1973**, pp. 307-310.

27. R. Ahlrichs, "Methods for efficient evaluation of integrals for Gaussian type basis sets", *Theoret. Chim. Acta.*, 33(2), **1974**, pp. 157-167.

28. J. Almlöf, K. Faegri & K. Korsell, "Principles for a direct SCF approach to LCAO-MO *ab initio* calculations", *J. Comp. Chem.* 3(3), **1982**, pp. 385-399.

29. M. Challacombe, "A simplified density matrix minimization for linear scaling SCF theory", Preprint - LA-UR-98-3682.

30. L. Greengard & S. Wandsura, "Fast Multipole Methods - guest editor introduction", *IEEE. Comp. Sci. Eng.*, 5, **1998**, pp. 16-18.

31. *Ibid.*

32. *Op Cit.*[30]

33. R.W. Hackney & J. W. Eastwood, "Computer simulation using particles", McGraw-Hill, New York, **1981**.

34. A.W. Appel, "An efficient program for many body simulation", *SIAM J. Sci. Stat. Comp.*, 6, **1985**, pp. 85-103

35. A. Brandt & A.A. Lubrecht, "Multilevel matrix multiplication and fast solution of integral equations", *J. Comp. Phys.*, 90, **1990**, pp. 348-370.

36. *Op Cit.*[8]

37. *Op Cit.*[7]

38. *Op Cit.*[36]

39. *C.*A. White & M. Head-Gordon, "Derivation and efficient implementation of the fast multipole method", *J. Chem. Phys.*, 101, **1994**, pp. 6593-6606.

40. M. Challacombe, E. Schwegler & J. Almlöf "Modern developments in Hartree-Fock theory:- Fast methods for computing the coulomb matrix", in "Computational Chemistry: Review of current trends", pp. 53-107, Ed. J. Leczynski, World Scientific, **1996**.

41. M.C. Strain, G.E. Scuseria & M.J. Frisch, "Achieving linear scaling for the electronic quantum coulomb problem", *Science*, 271, **1996**, pp. 51-53.

42. *Op Cit.*[14]

43. M. Challacombe, *Personal Correspondence.*

44. J.C. Burant, M.C. Strain, G.E. Scuseria & M.J. Frisch, "Kohn-Sham analytic energy second derivatives with the Gaussian very fast multipole method (GvFMM)", *Chem. Phys. Lett.*, 258, **1996**, pp. 45-52.

45. B.G. Johnson & M.J. Frisch, *J. Chem. Phys*, 100, **1994**, 8448.

46. *Op Cit*[2].

47. M. Challacombe & E. Schwegler, "Linear scaling computation of the Hartree-Fock exchange matrix", *J. Chem. Phys.*, 105, **1996**, 2726.

48. *Op Cit.*[40] P. 8.

49. M. Challacombe, "A simplified density matrix minimization for linear scaling SCF theory", *J. Chem. Phys.*, 110, **1999**, pp. 2332-2342.

50. *Op Cit.*[25]

51. *Op Cit.*[47]

52. E. Schwegler & M. Challacombe, "Linear scaling computation of the Fock matrix. II. Rigorous bounds on exchange integrals and incremental Fock builds", Accepted, Pre-print LA-UR-99-578.

53. E. Schwegler, M. Challacombe & M. Head-Gordon, "A multipole acceptability criterion for electronic structure theory", *J. Chem. Phys.*, 109, **1998**, pp. 8764-8769.

54. *Op Cit.*[25] p. 50

55. E. Schwegler & M. Challacombe, "Linear scaling computation of the Fock matrix. IV. Multipole accelerated formation of the exchange matrix", Accepted, Pre-print LA-UR-99-578.

56. *Op Cit.*[2]

57. *Op Cit.*[49]

58. M. Challacombe, *Personal Correspondence.*

59. *Op Cit.*[30]

60. M. Challacombe, *Personal Correspondence*

Acknowledgements:

Dr. Matt Challacombe (Theoretical Chemistry Division, Los Alamos National Laboratory)
For not hesitating to send me electronic copies of all his research work in the field of linear scaling SCF theory thereby saving me a fortune on photocopying costs. Also for reading draft copies of my report and supplying helpful advice.