8 SCIENTIFIC HIGHLIGHT OF THE MONTH: "Linear-scaling DFT calculations with the CONQUEST code"

Linear-scaling DFT calculations with the CONQUEST code

D. R. Bowler[1], A. S. Torralba[1], T. Miyazaki[2], T. Ohno[2] and M. J. Gillan[1]
[1]London Centre for Nanotechnology, UCL, Gordon St.
London WC1H 0AH, UK
[2]National Institute for Materials Science, 1-2-1 Sengen
Tsukuba, Ibaraki 305-0047, Japan

Abstract

We outline the main ideas underlying the CONQUEST code for first-principles modelling of systems containing many thousands of atoms, focusing on the algorithms used to achieve linear-scaling of the cpu and memory requirements with number of atoms, and the strategies for implementing the algorithms so as to achieve good parallel scaling on parallel computers. We note that the code can be run at different levels of precision, ranging from empirical tight-binding, through *ab initio* tight-binding, to full *ab initio*. Very recent technical developments implemented in the code are outlined. We give illustrations of physical systems currently being studied with the code, ranging from biologically important molecules to Ge hut clusters on Si (001), including structural relaxation on systems of over 20,000 atoms using electronically self-consistent density-functional theory. Arrangements for obtaining and learning to use the code are also noted.

1 Introduction

It is now over 15 years since the first proposals were made for doing DFT calculations so that the amount of memory and number of cpu cycles needed are proportional to the number of atoms, rather than scaling as N^2 or worse [1–7]. These ideas stimulated a flurry of activity, and in the middle 1990's it was more or less obligatory for every condensed-matter electronic-structure conference to include a section on 'linear-scaling' or '$O(N)$' methods. This activity rather quickly led to efficient practical codes for linear-scaling tight-binding calculations, but it gradually became clear that there were many practical difficulties in achieving the same thing for density functional theory. Not the least of these difficulties was that of making the calculations run efficiently on large parallel computers, so that they scaled linearly not only with the number of atoms but (inversely) with the number of processors. The consequence was that the effort to develop linear-scaling DFT codes died away to rather a low level, and the subject started to disappear from the conference programmes. Nevertheless, the persistent efforts of a few research groups have recently started to bear fruit, so that practical codes for performing DFT calculations on very large complex systems are now becoming available [8–12]. CONQUEST is one of these codes.

We have published many papers over the past 12 years about the principles underlying CONQUEST [8, 13–18], so the main purpose of this article is to give an update about recent progress, and particularly about the large-scale practical calculations that are now becoming possible. However, to make the article self-contained, we start by recalling the main ideas. We will then give a summary of how the computational effort is distributed across processors on parallel machines. Then we give some recent practical examples from unpublished or only partially published work, including exploratory calculations on the important enzyme dihydrofolate reductase, and large-scale structural relaxation calculations on Ge/Si hut clusters performed on the Earth Simulator on systems of over 20,000 atoms.

2 Principles of the CONQUEST code

2.1 Theory

The reasons why traditional DFT calculations scale poorly with N are well known. The number of occupied Kohn-Sham orbitals must clearly be proportional to N. But each orbital $\psi_n(\mathbf{r})$ extends over the entire volume of the system, which is also proportional to N. This means that the amount of stored information and the number of operations needed to manipulate it are proportional to N^2. However, all the usual implementations require an operation equivalent to calculating the scalar product $\langle \psi_m | \psi_n \rangle$ of all pairs of occupied orbitals, and the cpu time for this is proportional to N^3. The prefactor is small, but for very large systems this N^3 scaling will dominate. However, Kohn's 'near-sightedness' principle [19] tells us that it should be possible to do much better than this, and that $O(N)$ performance should be achievable. The amount of information stored in an N-atom system is not really proportional to N^2; it is just that the usual manner of doing things incurs an enormous degree of redundancy in the way the information is represented.

With DFT, the near-sightedness principle is expressed by the locality of the Kohn-Sham density matrix $\rho(\mathbf{r}, \mathbf{r}')$. Recall that if the Kohn-Sham occupied orbitals (eigenfunctions of the Kohn-Sham equation) are already known, then $\rho(\mathbf{r}, \mathbf{r}')$ is defined as:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_n \psi_n(\mathbf{r}) \psi_n(\mathbf{r}')^\star \ . \tag{1}$$

The nearsightedness principle says that there is quantum coherence only between nearby positions, or, more exactly: $\rho(\mathbf{r}, \mathbf{r}') \to 0$ as $|\mathbf{r} - \mathbf{r}'| \to \infty$. But the variational principle of DFT can be formulated in terms of the density matrix [7]: the DFT ground state is obtained by minimising the total energy E_{tot} with respect to $\rho(\mathbf{r}, \mathbf{r}')$, subject to the 'weak idempotency' condition that the eigenvalues of ρ should all lie between 0 and 1. Linear scaling is then obtained by minimising E_{tot} with respect to ρ, subject to the constraint that $\rho(\mathbf{r}, \mathbf{r}') = 0$ for $|\mathbf{r} - \mathbf{r}'| > r_c$, where r_c is a chosen cut-off distance. The amount of information stored in $\rho(\mathbf{r}, \mathbf{r}')$ is then manifestly $O(N)$. These ideas are implemented in CONQUEST, with the additional constraint that $\rho(\mathbf{r}, \mathbf{r}')$ be 'separable' (the number of its non-zero eigenvalues is finite), so that:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i\alpha, j\beta} \phi_{i\alpha}(\mathbf{r}) K_{i\alpha, j\beta} \phi_{j\beta}(\mathbf{r}') \ . \tag{2}$$

The functions $\phi_{i\alpha}(\mathbf{r})$, which we refer to as 'support functions', are chosen to be non-zero only within spherical regions of radius R_{reg} centred on the atoms ($\phi_{i\alpha}(\mathbf{r})$ is the αth support function on atom i). Effectively, the matrix $K_{i\alpha, j\beta}$ is the density matrix in the (non-orthogonal) basis of support functions.

In practice, then, the idea is to express the total energy in terms of the density matrix given by eqn (2), and to minimise it with respect to the K-matrix and the $\phi_{i\alpha}(\mathbf{r})$ support functions, subject to the conditions of (i) weak idempotency, and (ii) constant electron number. In doing this, the support functions should in principle be freely varied within their spherical regions, and for this purpose they need to be represented in terms of some chosen basis functions. Before discussing that, though, we address the more difficult question of how to ensure weak idempotency. It goes without saying that we are not allowed to diagonalise the density matrix, because that would be an $O(N^3)$ operation, and we would have achieved nothing. There are several ways of enforcing weak idempotency, but the present implementation in CONQUEST is a combination of the technique of Li, Nunes and Vanderbilt (LNV) [33] and Palser and Manolopoulos [34], both of which are related to McWeeny's 'purification' scheme [35]. In the LNV technique, the matrix K is represented in terms of an 'auxiliary' density matrix L as:

$$K = 3LSL - 2LSLSL \, , \tag{3}$$

where S is the overlap matrix of support functions: $S_{i\alpha,j\beta} = \langle \phi_{i\alpha} | \phi_{j\beta} \rangle$. In order to ensure $O(N)$ scaling, a spatial cut-off is imposed on the L-matrix, so that $L_{i\alpha,j\beta} = 0$ when the distance between atoms and i and j exceeds a chosen cut-off R_L. Alternative methods for enforcing weak idempotency could, of course, also be used.

In order to obtain a scheme that is equivalent to standard DFT, we must allow the support functions $\phi_{i\alpha}(\mathbf{r})$ to be freely varied within their spherical regions. This means that they must be represented in terms of some basis set. We have two completely different ways of doing this in CONQUEST, and which basis set one chooses depends on what one is trying to achieve. If plane-wave precision is desired, then it is essential to use a basis set that is in some sense equivalent to plane waves. The obvious difficulty is that the support functions are localised, so if one literally uses plane waves, then it is undesirable that they should extend over the entire system. The solution we have adopted is to use a finite-element basis that is quite closely equivalent to plane waves. (It is interesting here to compare with the plane-wave methods that have been used to solve this same problem in the ONETEP CODE [10].) On the other hand, if plane-wave precision is not needed, the other option on CONQUEST is to use numerical pseudo-atomic orbitals, as is done in some other codes, notably SIESTA [9, 29] and PLATO [39, 40].

The finite-element scheme we use to obtain plane-wave precision represents the $\phi_{i\alpha}(\mathbf{r})$ in terms of piecewise continuous polynomials, using a technique sometimes referred to as B-splines. Full details of the scheme, with demonstrations of its effectiveness, are presented in a published report [20], so here we give only a brief summary. Suppose first that we have a continuous function $f(x)$ in one dimension, which we wish to represent. The B-spline basis consists of localised functions $\theta_s(x)$, centred on the points of a grid, whose nodes are at positions $X_s = sa$, where a is the grid spacing. The basis functions are all images of each other, displaced by an integer number of grid spacings, so that $\theta_s(x) = \theta_0(x - X_s)$. The basis function $\theta_0(x)$ vanishes identically outside the range $-2a < x < 2a$. Inside this range, it is assembled from cubic polynomials:

$$\theta_0(x) = \begin{cases} 1 - \frac{3}{2}(x/a)^2 + \frac{3}{4}|x/a|^3 & \text{if} \quad 0 < |x| < a \\ \frac{1}{4}(2 - |x|/a)^3 & \text{if} \quad a < |x| < 2a \\ 0 & \text{if} \quad 2a < |x| \end{cases} \tag{4}$$

and has the property that it and its first two derivatives are continuous everywhere. In fact, the only discontinuities are in the third derivative at the points $|x| = 0$, a and $2a$. The representation of a continuous function

$$f(x) \simeq \sum_s b_s \theta_s(x) \tag{5}$$

can be made arbitrarily precise by systematically reducing the grid spacing a. This is exactly analogous to increasing the plane-wave cut-off $G_{\max}$ when taking a plane-wave calculations to convergence.

In fact, there is a close relationship between B-spline and plane-wave basis sets. The $\theta_s(x)$ basis functions can be used to form Bloch-like functions $\chi_k(x)$ by the unitary transformation:

$$\chi_k(x) = \sum_s e^{ikX_s} \theta_s(x) \; . \tag{6}$$

To obtain the full set of distinct χ_k functions, k should be restricted to the range $-\pi/a < k < \pi/a$. As $|k| \to 0$, the functions become identical to plane waves, and in fact they rather precisely reproduce plane waves except near the ends of the interval $(-\pi/a, \pi/a)$. This means that B-splines with grid spacing a are nearly equivalent to plane waves with cut-off $G_{\max} = \pi/a$.

In practice, of course, we work in three dimensions, and the three-dimensional B-splines $\Theta_s(\mathbf{r})$ are defined as Cartesian products:

$$\Theta(\mathbf{r} - \mathbf{R}_s) = \theta(x - X_s)\theta(y - Y_s)\theta(z - Z_s) \; , \tag{7}$$

where (X_s, Y_s, Z_s) are the Cartesian components of $\mathbf{R}_s$, and the support functions are represented as:

$$\phi_{i\alpha}(\mathbf{r}) = \sum_s b_{i\alpha s} \Theta_s(\mathbf{r} - \mathbf{R}_{is}) \; . \tag{8}$$

In the current scheme, the blip-grid on which the $\Theta_s(\mathbf{r})$ are sited is defined separately for each atom, and moves with that atom. To enforce the vanishing of $\phi_{i\alpha}(\mathbf{r})$ outside the support region, we include in eqn (8) only those $\Theta_s(\mathbf{r})$ that are non-zero only for points within the region. The reason for making the blip-grid move with the atom is that this ensures that each $\phi_{i\alpha}(\mathbf{r})$ is represented always in terms of the same set of basis functions.

Blip functions therefore give us a scheme that is closely related to plane waves, but at the same time respects the strict localisation of the support functions. It also shares another feature with plane waves, and that is that as the blip spacing is decreased, the computational effort grows linearly only with the number of blip functions. This is because the number of blip functions that are non-zero at each point in space does not increase as a decreases.

The alternative basis set of numerical pseudo-atomic orbitals provided in CONQUEST is similar in spirit to the ones used in the SIESTA [9, 29] and PLATO [39, 40] codes.

2.2 Implementation

In CONQUEST, the search for the ground-state is organised into three loops. In the innermost loop, the support functions and electron density are fixed and the ground-state density matrix is found, either by varying L or by diagonalisation. In the middle loop, self-consistency is achieved by systematically reducing the electron-density residual, i.e. the difference between the input and output density in a given self-consistency cycle. In the outer loop, the energy is minimised with respect to the support functions, $\phi_{i\alpha}$. This organisation corresponds to a hierarchy of approximations: when the inner loop alone is used, we get the scheme known as non-self-consistent *ab initio* tight binding (NSC-AITB), which is a form of the Harris-Foulkes approximation [22–25]; when the inner two loops are used, we get self-consistent *ab initio* tight binding (SC-AITB); finally, if all loops are used, we have full *ab initio*. In this last case, we recover the exact DFT ground state as the region radius R_{reg} and the L-matrix cut-off R_L are increased. For non-metallic systems, the evidence so far is that accurate approximations to the ground state are obtained with quite modest values of the cut-offs [9, 13]. For the non-self-consistent ground-state search of the inner loop, as well as operating in $O(N)$ mode, CONQUEST can find the ground state directly by diagonalisation, using the SCALAPACK package, which allows efficient parallelisation of the diagonalisation. Since this scales as $O(N^3)$, this will only be appropriate for relatively small systems, but it provides an important tool for testing the outer parts of the ground-state search, and for exploring the convergence of the $O(N)$ algorithm with the cut-off on the L-matrix.

For calculations at the level of full *ab initio* accuracy, the convergence of the outer loop (optimising the support functions with respect to their basis functions) is well-conditioned provided appropriate pre-conditioning measures are taken; these have been discussed both for blips in the context of CONQUEST [26, 27] and for psinc functions in the context of ONETEP [28]. We note that CONQUEST can be run in a mode analagous to SIESTA, where pseudo-atomic orbitals are used and no optimisation is performed; in this case, the outer loop is not performed.

We have recently found that the self-consistency search (the middle loop described above) can be accelerated by use of the Kerker preconditioning. This idea, which is well-known in the plane-wave community, removes long wavelength changes in the charge density during mixing. It is applied in reciprocal space, as a prefactor:

$$f(q) = \frac{q^2}{q^2 + q_0^2} \tag{9}$$

Then the charge is mixed using a Pulay or Broyden (or related) scheme [21] with the prefactor applied to the residual or output charge after transformation to reciprocal space. The mixing includes a parameter, A, which determines how aggressive the mixing is (with the input charge density for iteration $n+1$ given by $\rho_{in}^{n+1} = \rho_{in}^{n} + Af(q)R_n$, with R_n the residual from iteration n).

While performing the search for self-consistency, we must monitor the residual. We define the following dimensionless parameter which is used to monitor the search:

$$d = \frac{\langle |R(\mathbf{r})|^2 \rangle^{1/2}}{\bar{\rho}}, \tag{10}$$

$$\langle |R(\mathbf{r})|^2 \rangle = \frac{1}{V} \int \mathrm{d}\mathbf{r}\, |R(\mathbf{r})|^2 \,, \tag{11}$$

where V is the simulation cell volume and we use the usual definition of residual, $R(\mathbf{r}) = \rho_{\mathrm{out}}(\mathbf{r}) - \rho_{\mathrm{in}}(\mathbf{r})$, the difference between the output and input charge densities. The quantity d is then the RMS value of $R(\mathbf{r})$ normalised by dividing by the *average* charge density in the system, $\bar{\rho}$. Note that, for systems containing large amounts of vacuum, the criterion for convergence will need to be altered when compared to bulk-like environments. This criterion may be coupled with a monitor on the largest value of residual on an individual grid point $\mathbf{r}_l$, $R_{\mathrm{max}} = \max_l |R(\mathbf{r}_l)|$

The scheme we have outlined is closely related to the methods used in SIESTA [9, 29], OpenMX [11] and ONETEP [10]. The main differences are: (i) the basis sets chosen (SIESTA uses fixed PAOs, while OpenMX uses optimized orbitals and ONETEP psinc functions); (ii) the method of finding the ground state density matrix (Siesta uses the constrained search technique [3–5], OpenMX the divide-and-conquer [30] or BOP [31] and ONETEP either penalty functional [19, 32] or LNV [33]); (iii) the technique of 'neutral-atom potentials' [9, 29], used by SIESTA and OpenMX, which allows calculation of matrix elements to be performed very efficiently for localised, atomic-like basis sets.

2.3 Forces on the ions

In order to perform structural relaxation or molecular dynamics of materials with an electronic structure technique, the algorithms for calculating the forces $\mathbf{F}_i$ on the ions must be the exact derivatives of the total ground state energy, E_{GS}, with respect to the positions, $\mathbf{r}_i$, such that $\mathbf{F}_i = -\nabla_i E_{\mathrm{GS}}$. One of the advantages of DFT, within the pseudopotential approximation, is that it is easy, in principle, to achieve this relationship between the forces and the energy. Since the CONQUEST formalism allows the calculation of the total energy at different levels of accuracy, some care is needed in the formulation of the forces to develop a scheme that works at all levels of this hierarchy. It is also important to ensure that it works equally well (and accurately) for both the diagonalisation and $\mathcal{O}(N)$ modes of operation implemented in CONQUEST.

We recall the Harris-Foulkes expression [22, 23] for the total energy, which is often applied when self-consistency is not sought, but which at self-consistency is identical to the standard Kohn-Sham expression for total energy. The expression is:

$$E_{\mathrm{GS}} = E_{\mathrm{BS}} + \Delta E_{\mathrm{Har}} + \Delta E_{\mathrm{xc}} + E_{\mathrm{C}}, \tag{12}$$

with E_{C} the Coulomb energy between the ionic cores, and the band-structure energy, the double-counting Hartree and exchange-correlation energies defined as:

$$\begin{aligned}
E_{\mathrm{BS}} &= 2\sum_n f_n \epsilon_n && (13)\\
&= 2\mathrm{Tr}[KH] && (14)\\
\Delta E_{\mathrm{Har}} &= -\frac{1}{2}\int d\mathbf{r}\, n^{\mathrm{in}}(\mathbf{r}) V^{\mathrm{in}}_{\mathrm{Har}}(\mathbf{r}) &&\\
\Delta E_{\mathrm{xc}} &= \int d\mathbf{r}\, n^{\mathrm{in}}(\mathbf{r}) \left(\epsilon_{\mathrm{xc}}(n^{\mathrm{in}}(\mathbf{r})) - \mu_{\mathrm{xc}}(n^{\mathrm{in}}(\mathbf{r}))\right) . && (15)
\end{aligned}$$

Here, $n^{\mathrm{in}}(\mathbf{r})$ is the *input* charge density used (normally a superposition of atomic charge densities if a non-self-consistent scheme is used, or the self-consistent charge density if self-consistency is used). This expression is very useful when comparing forces at different levels of approximation.

At the empirical TB level, the ionic force is a sum of the band-structure part $\mathbf{F}_i^{\mathrm{BS}}$ and the pair-potential part $\mathbf{F}_i^{\mathrm{pair}}$, the former being given by [24]:

$$\mathbf{F}_i^{\mathrm{BS}} = -2\mathrm{Tr}\left[K\nabla_i H - J\nabla_i S\right] , \tag{16}$$

where K and J are the density matrix and energy matrix respectively [24]. It is readily shown that in the $\mathcal{O}(N)$ scheme of LNV, and in some other $\mathcal{O}(N)$ schemes, the same formula for $\mathbf{F}_i^{\mathrm{BS}}$ is the exact derivative of the $\mathcal{O}(N)$ total energy.

In NSC-AITB (Harris-Foulkes), the forces can be written in two equivalent ways. The way that corresponds most closely to empirical TB is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{BS}} + \mathbf{F}_i^{\Delta\mathrm{Har}} + \mathbf{F}_i^{\Delta\mathrm{xc}} + \mathbf{F}_i^{\mathrm{ion}} , \tag{17}$$

where $\mathbf{F}_i^{\mathrm{BS}}$ is given by exactly the same formula as in empirical TB. The contributions $\mathbf{F}_i^{\Delta\mathrm{Har}}$ and $\mathbf{F}_i^{\Delta\mathrm{xc}}$, which arise from the double-counting Hartree and exchange-correlation parts of the NSC-AITB total energy, have been discussed elsewhere [24]. The final term $\mathbf{F}_i^{\mathrm{ion}}$ come from the ion-ion Coulomb energy. This way of writing $\mathbf{F}_i$ expresses the well-known relationship between NSC-AITB and empirical TB that in the latter the pair term represents the sum of the three contributions $\Delta\mathrm{Har} + \Delta\mathrm{xc} + \mathrm{ion} - \mathrm{ion}$. The alternative, and exactly equivalent, way of writing $\mathbf{F}_i$ in NSC-AITB is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{ps}} + \mathbf{F}_i^{\mathrm{Pulay}} + \mathbf{F}_i^{\mathrm{NSC}} + \mathbf{F}_i^{\mathrm{ion}} . \tag{18}$$

Here, $\mathbf{F}_i^{\mathrm{ps}}$ is the "Hellmann-Feynman" force exerted by the valence electrons on the ion cores; $\mathbf{F}_i^{\mathrm{Pulay}}$ is the Pulay force that arises in any method where the basis set depends on ionic positions; $\mathbf{F}_i^{\mathrm{NSC}}$ is a force contribution associated with non-self-consistency, and is expressed in terms of the difference between output and input electron densities; $\mathbf{F}_i^{\mathrm{ion}}$, as before, is the ion-ion Coulomb force. Exactly the same formulas represent the exact derivative of E_{tot} in both diagonalisation and $\mathcal{O}(N)$ modes.

In both SC-AITB and full AI, the force formula is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{ps}} + \mathbf{F}_i^{\mathrm{Pulay}} + \mathbf{F}_i^{\mathrm{ion}} , \tag{19}$$

which differs from the second version of the NSC-AITB formula eqn (18) only by the absence of the non-self- consistent contribution $\mathbf{F}_i^{\mathrm{NSC}}$, as expected.

The above hierarchy of force formulas has been implemented in CONQUEST, and extensive tests have ensured that the total energy and the forces are exactly consistent within rounding-error precision [17].

2.4 Parallel operation

The principle of near-sightedness and the idea of parallel computation fit each other as a glove fits a hand. Since different regions of space are independent of each other as far as quantum coherence is concerned, there is a natural mapping of $O(N)$ calculations onto an array of processors. CONQUEST was written from the outset as parallel code, and a large part of the development effort has been concerned with techniques for achieving good parallel scaling. The parallelisation techniques have been described in detail elsewhere [8, 14, 36], so we give only a brief summary. There are three main types of operation that must be distributed across processors:

- the storage and manipulation of support functions, e.g. the calculation of $\phi_{i\alpha}(\mathbf{r})$ on the integration grid starting from blip- or PAO-coefficients, and the calculation of the derivatives of E_{tot} with respect to these coefficients, which are needed for the ground-state search;
- the storage and manipulation of elements of the various matrices (H, S, K, L, etc...);
- the calculation of matrix elements by summation over domains of points on the integration grid, or by analytic operations (for certain integrals involving PAOs and blips).

Efficient parallelisation of these operations, and the elimination of unnecessary communication between processors, depend heavily on the organisation of both atoms and grid points into small compact sets, which are assigned to processors [36]. When the code runs in $\mathcal{O}(N)$ mode, matrix multiplication takes a large part of the computer effort, and we have developed parallel multiplication techniques [36] that exploit the specific patterns of sparsity on which $\mathcal{O}(N)$ operation depends.

2.5 Recent technical progress

The implementation outlined above was already in place, and the practical $O(N)$ performance of the code was demonstrated several years ago. However, until fairly recently the range of systems to which the code could easily be applied was rather limited. However, in the past two years we have greatly enhanced its functionality and ease of use, in preparation for public release later this year. We have now standardized the pseudopotentials used in the code on the Troullier-Martins form [37]. The reason for this choice is that these pseudpotentials are used in a number of plane-wave/pseudopotential codes, such as ABINIT [38]. This makes it rather convenient to cross-check CONQUEST results against standard plane-wave calculations. If one chooses to use PAO basis sets in CONQUEST, it is necessary to generate the PAO's using the standard pseudopotentials. The code for doing this has been adapted from the PLATO code [39, 40]

Any practical DFT code needs to be able to use a range of available exchange-correlation functionals. To make this possible, we have recently implemented the PBE form of generalized-gradient approximation (GGA) [41], with PW92 parameterization [42] for the local part. The gradient calculations are done following the scheme of White and Bird [43], which is formally exact on a grid, and involves the computational of only four Fast Fourier Transforms (FFT's). The linearity of the scheme preserves $O(N)$ operation. In order to keep the ability of the code to perform structural relaxation with non-self-consistent Harris-Foulkes calculations, the original computation of forces had to be adapted to the newly implemented GGA functional. As we will report elsewhere [44], we are able to maintain the condition that the forces are exact derivatives of the total energy, and the number of FFT's remains equal to four.

We mentioned above that the division of atoms and grid points into compact groups is important in achieving good parallel efficiency. The way this was done in early versions of the code is summarised above. However, those methods turned out to be inefficient for problems in which a significant part of the system consists of empty space – a common situation when dealing with surface problems. We have now been able to develop more sophisticated procedures, which significantly improve the efficiency.

In applications of nano-devices, a crucial physical effect is often the transport of electrons, and the exchange of energy between the ionic and electronic sub-systems. These are effects that are not included in conventional first-principles molecular dynamics (m.d.) techniques, which explicitly or implicitly enforce the Born-Oppenheimer approximation, that the electronic sub-system adiabatically follows the motion of the ions. Recently, an important extension of m.d. has been developed, known as “correlated electron-ion dynamics” (CEID) [45], in which the quantum spread of the ions is included *via* a small-amplitude moment expansion. With CEID, it is possible to make direct numerical simulations of, for example, inelastic current-voltage spectroscopy in atomic wires. An ambition for the future is to implement CEID within the CONQUEST code, and we are currently formulating the strategies needed to do this.

3 Technical tests

Much of the hard work involved in developing any large code goes into demonstrating that it really achieves what it is intended to achieve, that it is reasonably robust, and that it runs efficiently on appropriate platforms. In the case of a linear-scaling DFT code like CONQUEST, the issues that must be addressed include the following: (i) does the code actually achieve parallel scaling with respect to the number of atoms? (ii) does it achieve good parallel scaling on parallel computers, and at what typical numbers of processors does the quality of the scaling start to deteriorate? (iii) if we go to basis-set convergence, and if we go to the limit of large support-region radius R_{reg} and large L-matrix cut-off, does it recover standard plane-wave results (using the same pseudopotentials, obviously)? (iv) how rapid is the convergence with respect to R_{reg} and R_L? (v) how rapid is the ground-state search for large systems? (vi) how rapid is the search for self-consistency for large systems? (vii) how rapid is structural relaxation for large systems?

The issue of scaling with respect to number of atoms and number of processors on large parallel computers was studied already 10 years ago, when we demonstrated excellent scaling of both kinds on systems of up to $\sim 15,000$ atoms using computers having up to ~ 512 processors [14]. More recently, we have done extensive tests on the Earth simulator, the results of which will be published soon. We have also presented the results of test on convergence with respect to R_{reg} and R_L. As an illustration of the search for self-consistency, we show in Fig. 1 the decrease of the self-consistency residual as a function of iteration number for an amorphous Si cluster of 343 atoms, which was specifically designed as a challenge to self consistency, because it is close to being metallic. The results show rather rapid monotonic convergence to self consistency, and generally we find similar behaviour also for much larger systems. As an illustration of the ground-state search for very large systems, we show in Fig. 2 the deviation of the energy from the exact ground-state value as a function of iteration number for a 23,000-atom Ge hut cluster on Si (001) (see below for more details of this system). For fuller discussion of the many other technical issues referred to above, our published papers should be consulted.

4 Scientific applications

In the immediate future, we expect the most important applications of CONQUEST to be in the area of biomolecular systems and nano-systems (there are, of course, close links between the two types of systems). In all cases, it is clearly essential to build up experience with $O(N)$ methods, starting with relatively small systems, where we can cross-check against the results of more standard codes. For nano-systems, we started this learning process with simple tests on semiconductor surfaces [17], and we are now making exploratory calculations on much larger and more complex systems. For biomolecules, we are still at the stage of tests on systems of a few hundred atoms.

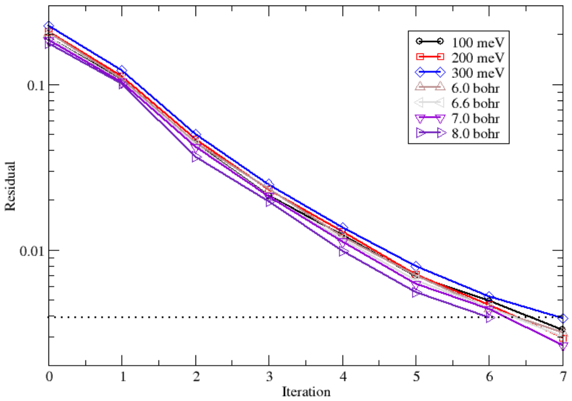

Figure 1: Residual during self-consistency search for different $R_{\rm reg}$. System treated is an amorphous Si cluster (see text).

In preparation for large-scale calculations on DNA systems, we are currently making extensive tests with CONQUEST on single DNA bases and on DNA base pairs, and comparing with results obtained with other codes, including SIESTA, VASP and GAUSSIAN. As expected, we find excellent agreement for the equilibrium bond lengths of covalently bonded atoms. Results of these tests will be published in the near future [46].

We are also performing tests on the important enzyme dihydrofolate reductase (DHFR), whose function in living organisms is to catalyze the reduction of dihydrofolate to produce tetrahydrofolate. The latter is an important molecule in metabolism. In particular, it is an essential cofactor in one-carbon transfer reactions. As a consequence, DHFR, which is the only enzyme that synthesizes it, has receive much attention, for example as a target for anti-malarial drugs. Although the specific substrate for DHFR is dihydrofolate (DHF), in some species the enzyme also catalyzes, very inefficiently and less specifically, the reduction of folate, a precursor of DHF.

The reasons why DHFR is specific for DHF remain unclear. LDA DFT calculations of the active site suggested that enzyme-induced polarization of the substrates may be a cause for the preference, at least in the *Escherichia coli* enzyme. Indeed, one study [47] found large electron density differences (EDD) between the density of DHF when bound to the enzyme with respect to that in vacuum. However, results from MP2 calculations, although qualitatively supportive for a role of polarization, are less conclusive [48, 49].

All existing studies used a point-charge model for the bulk protein, restricting the quantum mechanical (QM) calculations to a few atoms at the active site. Hence, the quantitative discrepances between different studies may be due to that limitation of the models, rather than to the different QM methods employed. Since DHFR is a relatively small protein (159 amino acids in *Escherichia coli*, or about 3000 atoms), we decided to assess such posibility by using Conquest to perform LDA DFT calculations in extended models of the active site, with the ultimate goal of including the whole of the protein. Thus, we did not model bulk protein in any way, since its effect was expected to become obvious as the size of the model increased.

Our preliminary results on portions of the protein of up to 300 atoms show that indeed larger models are quantitatively closer to MP2 results than to the original LDA calculations. We found larger polarization on DHF than on folate, and only DHF displayed polarization on the bond susceptible of hydrogenation, consistent with the observed specificity (see Fig. 3). Furthermore, calculations on different conformations of the protein agree with experimental evidence regarding the mechanism.

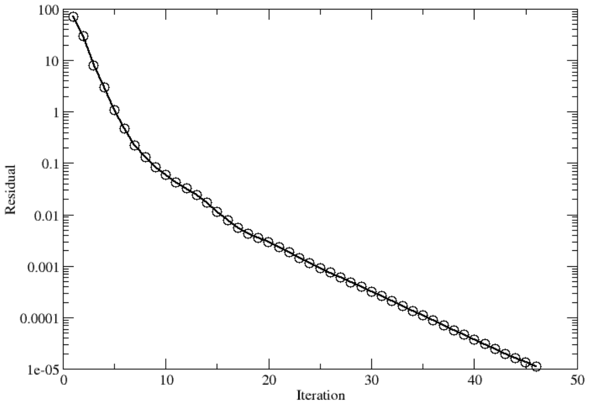

Figure 2: Convergence during energy minimisation with respect to density matrix elements (inner loop of ground-state search) for 23,000-atom Ge hut cluster on Si (001).

In particular, the presence of some amino acids of the so-called Met20 loop seems to be essential for catalysis, as represented by polarization on the hydrogenable bond of the substrate. Moreover, those amino acids must be occluding the active site for polarization to be observed, as expected in the proposed mechanism [50].

Turning now to the application of CONQUEST to nano-systems, we summarise our recent progress in investigating the three-dimensional (3D) structures formed when Ge is deposited on the Si (001) surface. The Ge/Si (001) has been extensively studied, because it is a prototypical example of hetero-epitaxial Stranski-Krastanov growth. When Ge atoms are deposited on Si (001), growth initially occurs layer by layer, up to a critical thickness of about three monolayers (ML). Strain due to the lattice mismatch is relieved by the formation of regularly spaced rows of dimer vacancies in the two-dimensional (2D) structure, resulting in the $2 \times N$ structure. Deposition of further Ge leads to another strain-relief structure, 3D pyramid-like structures known as "hut clusters" [51]. Recently, we have studied this transition from 2D to 3D structures, using CONQUEST.

Usually, the stability of 3D structures is governed by (i) the lowering of strain energy in the clusters and the underlying substrate, and (ii) the energy increase arising from the formation of facets. Theoretical approaches used so far have used continuum elasticity theory to describe the strain energy, with DFT being used only for the surface energies [52, 53]. For the Ge/Si system, the four facets of the hut cluster are well established to be $\{105\}$ surfaces, and the structure of these surfaces has recently been clarified by DFT calculations [54, 55]. Note that the typical side-length of hut clusters is about 150 Å, and deposition of additional Ge leads to the formation of other 3D structures called "domes", having steeper facets. Interestingly and importantly, the DFT calculations show that the strained Ge (105) surface is more stable than strained Ge (001). This means that the surface energy may actually *stabilise* the structure. If the surface contribution to the overall energy is small or favours the 3D structure, contributions from the edges where the facets meet each other and the wetting layer may also affect the stability of the 3D structure. In addition, as the area of the facets of the experimentally observed Ge hut cluster is not large, the evaluation of the surface part itself is doubtful. For these reasons, the validity of previous theoretical approaches is uncertain, especially for small hut clusters. To overcome these problems, we are using CONQUEST to model the entire hut cluster, together with the wetting layer and the Si substrate.

In preparation for CONQUEST calculations on the full system, we first performed DFT calculations on the Ge (105) surface, including test calculations also on the unstrained and strained Ge

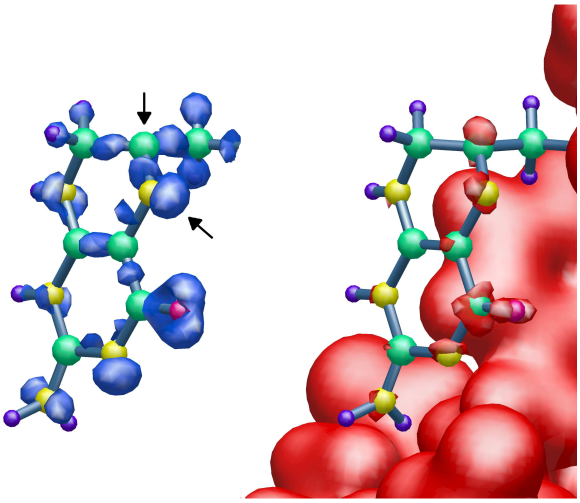

Figure 3: Electronic density difference plot for dihydrofolate (DHF) upon binding to the enzyme dihydrofolate reductase (± 0.001 electron/bohr). Charge deficiency (with respect to the density of DHF in vacuum) is shown in blue; charge excess is red. The enzyme induces polarization on N5 and C6 atoms (marked with arrows), and electronic density withdrawal from the bond linking them. These effects are consistent with the catalyzed reaction, namely, protonation of N and hydride transfer to the bond, and are much weaker for the very inefficient reduction of folate, a secondary substrate (not shown).

bulk [56]. Since the size of this system is relatively small, we can employ diagonalisation in this case. We have clarified the accuracy of the various DFT methods explained above for the unstrained and strained Ge systems. We have also confirmed that full DFT calculations performed with CONQUEST using cubic-spline basis sets are accurate enough for the study of the strained Ge (105) surface. The conditions need for $O(N)$ calculations to achieve good accuracy for this system have also been established.

Using these results, we have performed $O(N)$ DFT calculations on the entire Ge/Si (001) hut clusters. At the non-self-consistent level, we have performed structual optimisation on systems of different sizes. The largest system treated so far, shown in Fig. 4, contains ~ 23000 atoms, and we found that structure optimisation is robust even for such large systems. We have examined three structural models of the Ge hut cluster having different facet or edge structures, and we have compared their energies with those of the $2 \times N$ reconstructions with $N = 4$, 6 and 8. The results, to be reported in detail elsewhere [56,57], show that the 2D structure is more stable for small coverages of Ge atoms, but the 3D hut structure becomes more stable when the coverage exceeds 2.6 monolayers, in agreement with experimental observation.

5 Distribution of the CONQUEST code

We plan that the CONQUEST code will be released under a GNU General Public License by the end of 2007. At the time of writing (end of May 2007), we are in the beta-testing phase, and the code has been released to a small set of carefully chosen users, who will work with us to apply the code to their own scientific problems. A short tutorial course on the practical use of CONQUEST will be held at CECAM $7-8$ September 2007, and there is funding to support the attendance of participants. For more details, please go to `www.cecam.fr` and click on 'tutorials'.

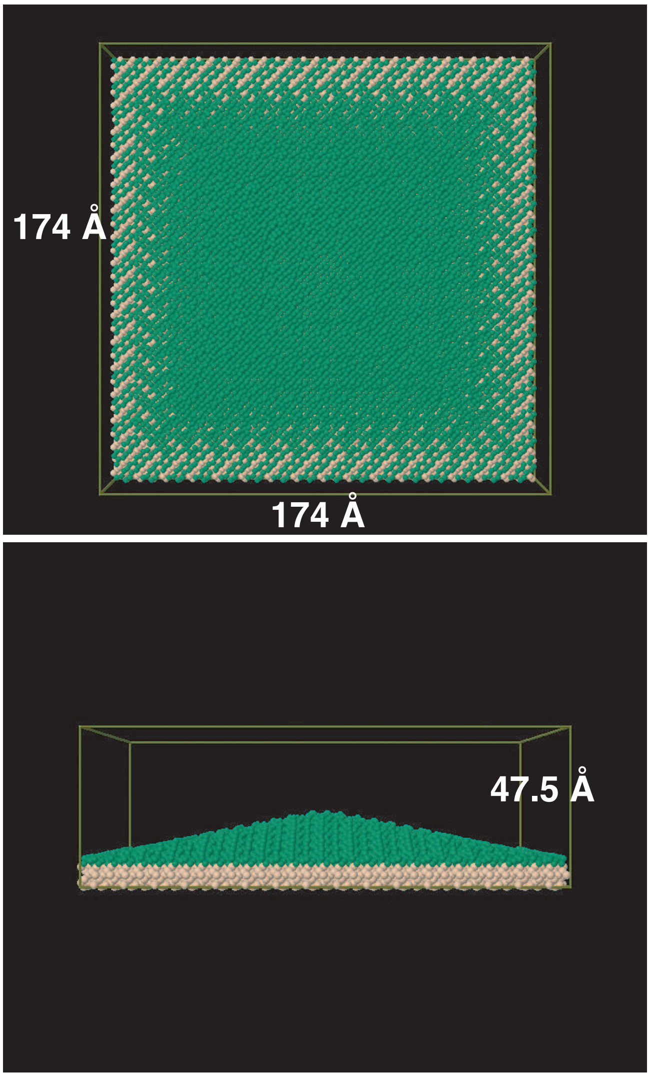

Figure 4: Atomic geometry of largest Ge/Si (001) hut cluster used for structural relaxation with CONQUEST DFT calculations. Upper and lower panels shows plan and side views, respectively. Pink and green spheres represent Si and Ge atoms. Dimensions of periodically repeated cell in surface plane and normal to surface are marked.

6 Summary

The main ideas underlying linear-scaling DFT were established in the early 1990's. However, the realisation of these ideas in practical codes has required the solution of a large number of technical problems concerning basis sets, the enforcement of linear scaling in the calculation of the ground-state density matrix, efficient manipulation of sparse matrices having the patterns of sparsity associated with spatial locality in three dimensions, and implementation of the algorithms on large parallel computers. Some of these problems admit of more than one solution, and the codes that have appeared so far, including CONQUEST, SIESTA, ONETEP and OPEN-MX, differ in important ways. We have tried to show here how the CONQUEST code has now passed beyond the stage of feasibility studies, and can now be applied to real scientific problems concerning biomolecular and nanoscale systems. Comparisons with the results of standard codes for relatively small systems of a few hundred atoms are demonstrating the realibility of the methods. At the same time, it is clear that structural relaxation at different levels of precision, using both self-consistent and non-self-consistent calculations, is becoming a practical proposition for systems containing more than 20,000 atoms.

Acknowledgments

DRB is supported by a Royal Society University Research Fellowship, The CONQUEST project is partially support by Special Coordination Funds for Promoting Science and Technology from the MEXT, Japan. Some of the calculations were carried out on the Earth Simulator supercomputer.

References

[1] S. Baroni and P. Giannozzi, *Europhys. Lett.*, **17**, 547 (1991).

[2] G. Galli and M. Parrinello, *Phys. Rev. Lett.*, **69**, 3547 (1992).

[3] F. Mauri, G. Galli and R. Car, *Phys. Rev. B*, **47**, 9973 (1993).

[4] J. Kim, F. Mauri and G. Galli, *Phys. Rev. B*, **52**, 1640 (1995).

[5] P. Ordejón, D. Drabold, M. Grumbach and R. Martin, *Phys. Rev. B*, **48**, 14646 (1993).

[6] P. Ordejón, D. Drabold, R. Martin and M. Grumbach, *Phys. Rev. B*, **51**, 1456 (1995).

[7] E. Hernández and M. J. Gillan, *Phys. Rev. B*, **51**, 10157 (1995).

[8] D. R. Bowler, T. Miyazaki and M. J. Gillan, *J. Phys: Condens. Matter*, **14**, 2781 (2002)

[9] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón and D. Sánchez-Portal, *J. Phys: Condens. Matter*, **14**, 2745 (2002).

[10] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, *J. Chem. Phys.*, **122**, 084119 (2005).

[11] T. Ozaki, *Phys. Rev. B*, **74**, 245101 (2006).

[12] The BigDFT project: `http://www-drfmc.cea.fr/sp2m/L_Sim/BigDFT/index.en.html`

[13] E. Hernández, M. J. Gillan and C. M. Goringe, *Phys. Rev. B*, **53**, 7147 (1996).

[14] C. M. Goringe, E. Hernández, M. J. Gillan and I. J. Bush, *Comput. Phys. Commun.*, **102**, (1997).

[15] D. R. Bowler and M. J. Gillan, *Comput. Phys. Commun.*, **120**, 95 (1999).

[16] D. R. Bowler, I. J. Bush and M. J. Gillan, *Int. J. Quantum Chem.*, **77**, 831 (2000).

[17] T. Miyazaki, D. R. Bowler, R. Choudhury and M. J. Gillan, *J. Chem. Phys.*, **121**, 6186 (2004).

[18] D. R. Bowler, R. Choudhury, M. J. Gillan and T. Miyazaki, *phys. stat. sol.*, **243**, 989 (2006).

[19] W. Kohn, *Phys. Rev. Lett.*, **76**, 3168 (1996).

[20] E. Hernández and M. J. Gillan, *Phys. Rev. B*, **55**, 13485 (1997).

[21] D. D. Johnson, *Phys. Rev. B*, **38**, 12807 (1988).

[22] J. Harris, *Phys. Rev. B*, **31**, 1770 (1985).

[23] W. M. C. Foulkes and R. Haydock, *Phys. Rev. B*, **39**, 12520 (1989).

[24] O. F. Sankey and D. J. Niklewski, *Phys. Rev. B*, **40**, 3979 (1989).

[25] A. P. Horsfield and A. M. Bratkowsky, *J. Phys. Condens. Matter*, **12**, R1 (2000).

[26] D. R. Bowler and M. J. Gillan, *Comput. Phys. Commun.*, **112**, 103 (1998).

[27] M. J. Gillan, D. R. Bowler, C. M. Goringe and E. Hernández, in *The Physics of Complex Liquids*, ed. F. Yonezawa, K. Tsuji, K. Kaji, M. Doi and T. Fujiwara, (World Scientific, 1998).

[28] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris and M. C. Payne, *J. Chem. Phys.*, **119** (2003).

[29] P. Ordejón, E. Artacho and J. M. Soler, *Phys. Rev. B*, **53**, R10441 (1996).

[30] W. Yang, *Phys. Rev. Lett.*, **66**, 1438 (1991).

[31] T. Ozaki and K. Terakura, *Phys. Rev. B*, **64**, 195126 (2001).

[32] P. D. Haynes and M. C. Payne, *Comput. Phys. Commun.*, **102**, 17 (1999).

[33] X.-P. Li, R. W. Nunes and D. Vanderbilt, *Phys. Rev. B*, **47**, 10891 (1993).

[34] A. H. R. Palser and D. Manolopoulos, *Phys. Rev. B*, **58**, 12704 (1998).

[35] R. McWeeny, *Rev. Mod. Phys.*, **32**, 335 (1960).

[36] D. R. Bowler, T. Miyazaki and M. J. Gillan, *Comput. Phys. Commun.*, **137**, 255 (2001).

[37] N. Troullier and J. L. Martins, *Phys. Rev. B*, **43**, 1993 (1991).

[38] X. Gonze, J.-M. Beuken, R. Caracas, F. Detraux, M. Fuchs, G.-M. Rignanese, L. Sindic, M. Verstraete, G. Zérah, F. Jollet, M. Torrent, A. Roy, M. Mimaki, P. Ghosez, J.-Y. Raty, D. C. Allan, *Comput. Mater. Sci.*, **25**, 478 (2002).

[39] A. P. Horsfield, *Phys. Rev. B*, **56**, 6594 (1997).

[40] S. D. Keny, A. P. Horsfield, H. Fujitani, *Phys. Rev. B*, **62**, 4899 (2000).

[41] J. P. Perdew, K. Burke and M. Enzerhof, *Phys. Rev. Lett.*, **77**, 3865 (1996).

[42] J. P. Perdew and Y. Wang, *Phys. Rev. B*, **45**, 13244 (1992).

[43] J. A. White and D. M. Bird, *Phys. Rev. B*, **50**, 4954 (1994).

[44] A. S. Torralba, D. R. Bowler and M. J. Gillan, in preparation.

[45] A. P. Horsfield, D. R. Bower, A. J. Fisher, T. N. Todorov, and C. G. Sanchez, *J. Phys. Condens. Matter*, **17**, 4793 (2005).

[46] T. Otsuka, T. Miyazaki, T. Ohno, D. R. Bowler and M. J. Gillan, in preparation.

[47] J. Bajorath, J. Kraut, Z. Li, D. H. Kitson, A. T. Hagler, *Proc. Natl. Acad. Sci. USA*, **88**, 6423 (1991).

[48] S. P. Greatbanks, J. E. Gready, A. C. Limaye and A. P. Rendell, *Proteins*, **37**, 157 (1999).

[49] M. García-Vicola, D. G. Truhlar and J. Gao, *J. Mol. Biol.*, **327**, 549 (2003).

[50] D. H. J. Schnell, R. Jason and P. E. Wright, *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 119 (2004).

[51] Y. W. Mo, D. E. Savage, B. G. Schwartzentruber and M. G. Lagally, *Phys. Rev. Lett.*, **65**, 1020 (1990).

[52] O. E. Shklyaev, M. J. Beck, M. Asta, M. J. Miksis and P. W. Vorhees, *Phys. Rev. Lett.*, **94**, 176102 (2005).

[53] G.-H. Lu and F. Liu, *Phys. Rev. Lett.*, **94**, 176103 (2005).

[54] Y. Fujikawa, K. Akiyama, T. Nagao, T. Sakurai, M. G. Lagally, T. Hashimoto, Y. Morikawa and K. Terakura, *Phys. Rev. Lett.*, **88**, 176101 (2002).

[55] P. Raiteri, D. B. Migas, L. Kiglio, A. Rastelli and H. von Känel, *Phys. Rev. Lett.* **88**, 256103 (2002).

[56] T. Miyazaki, D. R. Bowler, R. Choudhury and M. J. Gillan, *Phys. Rev. B*, submitted.

[57] T. Miyazaki, D. R. Bowler, M. J. Gillan and T. Ohno, in preparation.