

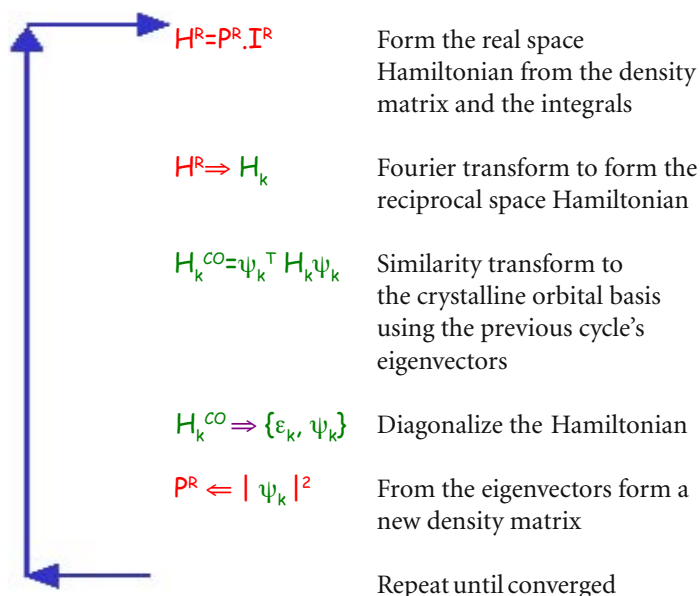
CRYSTAL on HPCx: *ab initio studies of proteins*

Ian J. Bush
HPCx Terascaling Team

CRYSTAL

CRYSTAL performs *ab initio* electronic structure calculations on periodic systems, ie crystals, slabs or polymers. Hartree-Fock, Kohn-Sham or various hybrid Hamiltonians may be used to calculate the wave-functions and properties of such systems, and the wave-functions are expanded in an atom-centred Gaussian-type basis. The code has been developed in a long-standing collaboration between the Theoretical Chemistry group at the University of Torino, Italy, and the Computational Materials Science group at Daresbury Laboratory. More recently, the Advanced Research Computing group, also of Daresbury Laboratory, has also been involved, in particular to enable CRYSTAL to exploit large MPP machines. Further details of the code's applications can be found on the CRYSTAL web pages [1,2].

To solve either the Hartree-Fock or Kohn-Sham equations an iterative procedure is used, which is illustrated below:



The time for large-scale CRYSTAL calculations is totally dominated by two steps, namely the formation of the real space Hamiltonian (H^R) and the diagonalization. As these two steps have different memory requirements and computational characteristics they are parallelized in different ways.

The formation of H^R involves the evaluation of a very large number of integrals. Since each evaluation is independent, task

farming is a natural and effective way to parallelize this portion of the code. This is facilitated by the memory requirements not being very large, since H^R and P^R are sparse, and so it is possible, though not ideal, to replicate all the data required for this stage on all the processors.

On the other hand, the diagonalization step is highly memory intensive and further for large systems there are few independent tasks. It is therefore necessary to use a distributed data algorithm for this stage of the calculation. The BFG package (developed by myself [3]) is used to perform the diagonalization. Here, BFG is preferred over other packages, such as ScaLAPACK [4], for a number of reasons. Firstly, unlike other methods, its workspace requirements are modest, thus allowing it to solve somewhat larger systems in practical applications. Secondly, as it uses Jacobi's algorithm, rather than Householder's, it is very straightforward to use an initial guess at the eigenvectors to speed up the calculation. In an iterative scheme like CRYSTAL's this is very useful as the eigenvectors from the previous cycle can be used as such a guess.



Figure 1: Abyssinian cabbage

Finally, Jacobi's algorithm is more amenable to parallelization than Householder's, and so tends to scale better to large numbers of processors. The net result of these last two points is that, though Jacobi's algorithm is somewhat slower than Householder's in serial, on a large number of processors, with a reasonable guess at the eigenvectors, it becomes competitive.

Figure 2:
The Crambin
unit cell.

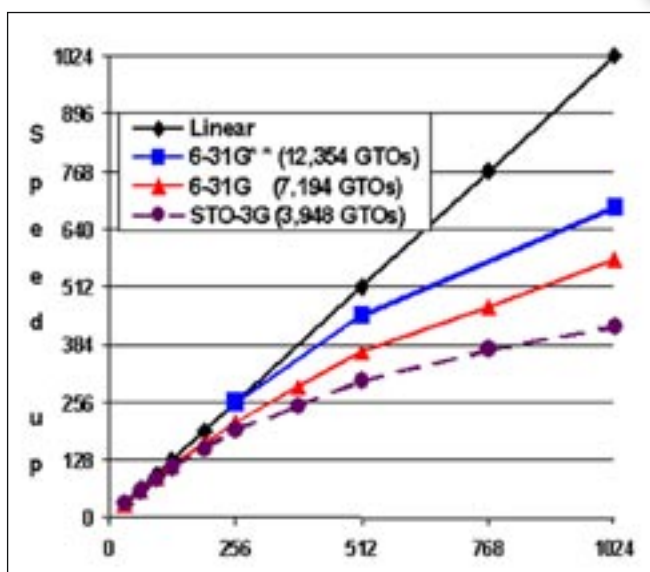
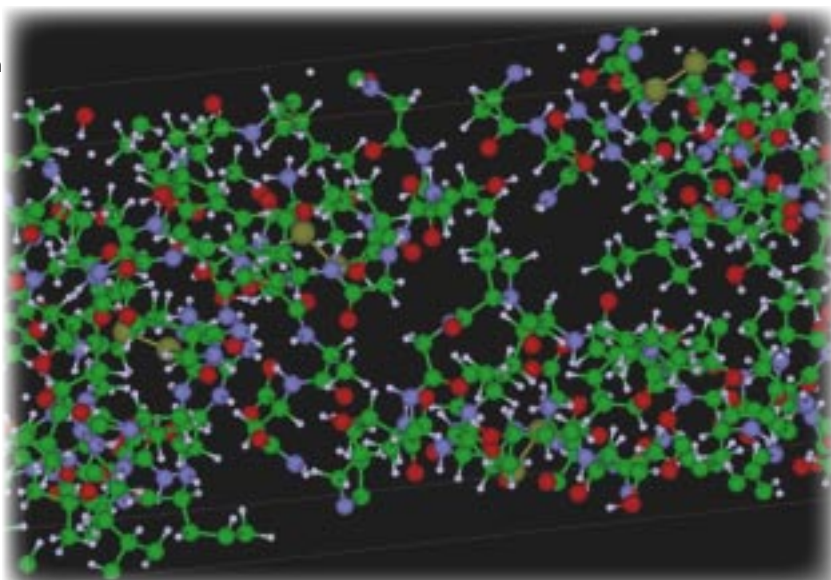


Figure 3: The scaling of CRYSTAL

The performance of CRYSTAL on HPCx

I shall illustrate the performance and scalability of CRYSTAL on HPCx by three calculations on crystalline Crambin. This is a small protein that can be extracted from *Crambe Abyssinica* (Abyssinian cabbage, see figure 1). The structure of this protein has been determined to extraordinarily high precision by X-ray diffraction studies (0.52Å resolution) [5,6]. Such a high resolution allows the experiment to resolve the hydrogen atoms in the structure, and so makes the experimentally determined crystal structure amenable to *ab initio* electronic structure calculations, such as that employed by CRYSTAL.

Figure 2 shows the conventional unit cell of Crambin. The symmetry is P2₁, so there are two chains per cell. Each chain contains 46 amino acid residues. In total there are 1284 atoms. This structure has been studied at the Hartree-Fock level of approximation using three different basis sets: STO-3G (3948 functions), 6-31G (7194) and 6-31G** (12354). It is believed that this last calculation is the largest ever Hartree-Fock calculation ever converged by an appreciable margin (compare [7]). Such sizes of system can only now be studied due to both the power of computers such as HPCx and, crucially, the improvement in software that allows such machines to be successfully exploited.

Figure 3 shows the scalability of CRYSTAL for the 3 basis sets. In each case it is not possible to run the calculation on a single processor, and so the displayed speed-up values have been generated by a fit of the measured times to Amdahl's law.

It can be seen that increasing the size of the basis set, and hence the quality of the calculation, increases the scalability. The fit to Amdahl's law suggests a speed-up of around 700 for the 6-31G** basis on 1024 processors - a gratifying result. The total time for the 6-31G** calculation is around 3 hours on 1024 processors.

Though the solution of the Hartree-Fock equations for Crambin is a challenging calculation for HPCx, it must be remembered that the solution of such problems is not the *raison d'être* of the service, but rather the science that such solutions can provide. Figure 4 shows some initial results from the 6-31G** calculation.

Figure 4 projects the electrostatic potential within the system onto an electron density isosurface which is at 0.1 electrons per unit cell volume, and has 0.1Å resolution. Regions of negative and positive potential are blue and red, respectively. Extremes in the potential are useful for identifying possibly active chemical groups, and in this case a carboxylate group can be seen near the upper left hand corner (the two large dark blue 'spheres') and a protonated arginine group near by.

In passing I note that the analysis of such results is non-trivial. To

Figure 4: The computed structure of Crambin showing a colour map of electrostatic potential projected onto an electron density isosurface.

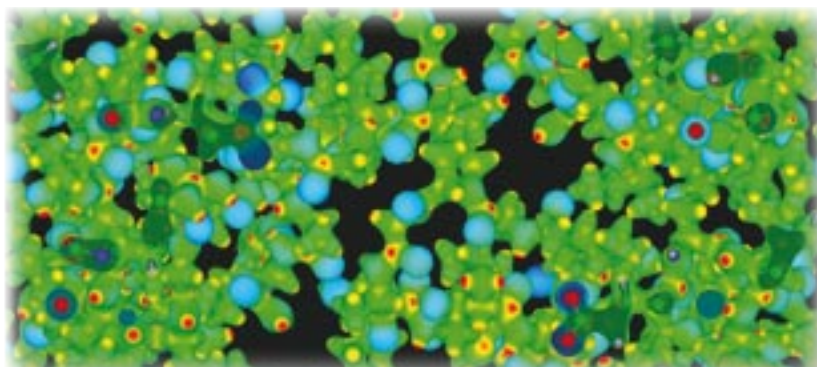
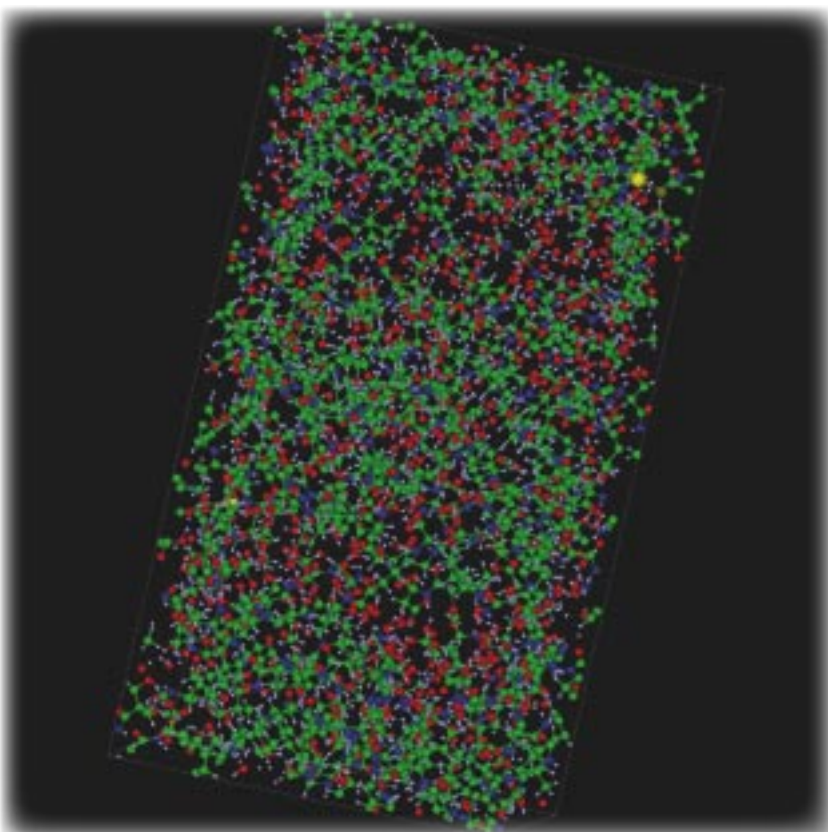


Figure 5: Rusticyanin



produce figure 4 it was necessary to parallelize the analysis codes themselves, such is the scale of the above calculation, and in fact the time to do the analysis alone is a sizeable fraction of that required to perform the Hartree-Fock calculation!

The capabilities of CRYSTAL on HPCx

Though the calculations on Crambin are challenging, the use of such a system as an example has a major drawback, namely that as a structural protein its biochemistry is of small interest as it gives little insight into major biochemical processes. So though Crambin is a very good test bed for the use of *ab initio* quantum mechanical methods to study proteins, the results of such calculations are of little interest to biochemists. Therefore, the next step in these calculations is to study a system of interest to that scientific area, and to use the results of such calculations to obtain a greater understanding of the processes occurring.

Such calculations are very challenging. Proteins that are of interest are typically much larger than Crambin, usually with hundreds to thousands of amino acid residues, but recently it has been demonstrated that such a system can be studied using HPCx and CRYSTAL.

The system is Rusticyanin, a blue copper protein, which is involved in biochemically important redox processes [8] through the copper atom in its core. Further, it is believed to be the ancestor of a large number of important enzymes [9]. It is illustrated in figure 5. The copper atom is the bright yellow sphere middle top right.

These calculations require the full power of HPCx. Rusticyanin has 155 residues in a chain, two chains per unit cell ($P2_1$ symmetry), 6284 atoms in the cell, and to perform calculations of a respectable quality over 33000 basis functions are required. However, such calculations have been shown to be possible on HPCx using CRYSTAL, calculations that are almost an order of magnitude larger than any published before (compare [7]). This exemplifies the scale of system that may be addressed by modern HPC systems such as HPCx, and shows that hitherto unexploited scientific areas may be investigated through HPC.

Acknowledgements

I would like to thank a large number of people for help with this work. Martyn Winn has been extremely helpful in the understanding of protein science. Barry Searle produced virtually all the pictures in this report, as the author of the software used (DLV), and further has been very helpful in calculating physical results from the results of my calculations. Samar Hasnain and Mike Hough provided the structure for Rusticyanin, and have helped extensively in its understanding. Vic Saunders has been extremely helpful with the basis sets, an absolutely crucial aspect to this work. I would also like to thank Adrian Wander and Nic Harrison for the scientific insight they have provided.

References

- [1] VR Saunders, R Dovesi, C Roetti, R Orlando, CM Zicovich-Wilson, NM Harrison, K Doll, B Civalleri, I Bush, Ph D'Arco, M Llunell *CRYSTAL 2003 User's Manual*, University of Torino, Torino, 2003
- [2] www.crystal.unito.it
- [3] www.cse.clrc.ac.uk/arc/bfg.shtml
- [4] www.netlib.org/scalapack/scalapack_home.html
- [5] C Jelsch, MM Teeter, V Lamzin, V Pichon-Lesme, B Blessing, C Lecomte, Proc. Nat. Acad. Sci. USA, 97 3171 (2000)
- [6] Entry 1EJG in www.ebi.ac.uk/msd
- [7] A Mitin, J. Mol. Struc (Theochem), 592 115 (2002)
- [8] R Walter, S Ealick, A Friedman, R Blake II, P Proctor, M Shoham, J. Mol. Biol., 263 730 (1996)
- [9] L Kanbi, S Antonyuk, M Hough, J Hall, F Dodd, S Hasnain, J. Mol. Biol., 320 263 (2002)