# Applying Grid Computing to the Parameter Sweep of a Group Difference Pseudopotential

Wibke Sudholt[1 *], Kim K. Baldridge[1], David Abramson[2], Colin Enticott[2], and Slavisa Garic[2]

[1] Department of Chemistry & Biochemistry and San Diego Supercomputer Center (SDSC), University of California, San Diego (UCSD), 9500 Gilman Dr., La Jolla, CA 92093-0505, USA
{wibke, kimb}@sdsc.edu

[2] Center for Enterprise Distributed Systems (DSTC) and School of Computer Science and Software Engineering, Monash University, Clayton, Victoria, 3800 Australia
{davida, Colin.Enticott}@csse.monash.edu.au,
Slavisa.Garic@infotech.monash.edu.au

**Abstract.** Theoretical modeling of chemical and biological processes is a key to understand nature and to predict experiments. Unfortunately, this is very data and computation extensive. However, the worldwide computing grid can now provide the necessary resources. Here, we present a coupling of the GAMESS quantum chemical code to the Nimrod/G grid distribution tool, which is applied to the parameter scan of a group difference pseudopotential (GDP). This represents the initial step in parameterization of a capping atom for hybrid quantum mechanics-molecular mechanics (QM/MM) calculations. The results give hints to the physical forces of functional group distinctions and starting points for later parameter optimizations. The demonstrated technology significantly extends the manageability of accurate, but costly quantum chemical calculations and is valuable for many applications involving thousands of independent runs.

## 1 Introduction

Efforts in cyberinfrastructure, which offer new research avenues through high-performance grid and information technologies, enable a better coupling of the science and engineering communities. With grid computing, we see a paradigm shift away from large-scale hardware and compute-intensive use to that of end-to-end performance, coordinating software and interfaces, as well as on data, and remote access. This requires multidisciplinary expertise and a deeper level of collaboration. Grid technology promises novel modes of coupling scientific models and unique strategies of sharing data, which help bridging the gaps in our knowledge of natural complexity. Understanding structure/function relationships and molecular processes in biological systems can leverage computer-based information tools, and the level of content generated by high-throughput technologies pushes research developments to new heights.

[*] Now: Institute of Organic Chemistry, University of Zurich, Winterthurerstr. 190, CH-8057 Zurich, Switzerland, {wibke, kimb}@oci.unizh.ch

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3036, pp. 148–155, 2004.

Grid technology has already dramatically changed biomedical research, enabling a high rate of knowledge and application advancements. Theoretical studies mimic biological and chemical processes on a level of complexity such that their fundamental details can be extracted, which cannot be obtained otherwise. Experimental and clinical information then refines these empirical models, which in turn enhances their predictive power. Major components of this iterative procedure can be facilitated by grid resources, which enable the transparent interoperability of research advancements not solely in hardware and computational speed, but also in database/storage technologies, visualization/environment infrastructure, and computational algorithms, all under high-speed networks and remote access (portals). The ultimate goal of exploiting grid technologies for the life sciences is to facilitate knowledge acquisition by harnessing computational tools to create new algorithmic strategies, ease complex multi-step procedures, and organize, manage, and mine data, all in a seamless manner.

Computational modeling in the life sciences is still very challenging and much of the success has been despite the difficulties in integrating all the technologies. For example, many simulations still use a simplified physics/chemistry, or restrict the spatio-temporal dimension or resolution of the model systems. Grid technologies offer to create new paradigms for computing, enabling access to resources which could span the biological scale. In this work, we illustrate a conceptual approach for computational investigations that involve many steps of processing, bookkeeping, and repetitive computation over several variant parameters. Such investigations might be for the generation of algorithm pieces, or for the substantiation of a chemical hypothesis. The approach, which invokes new robust grid technologies, is illustrated for a particular case in the former category here, but a more general implementation is possible.

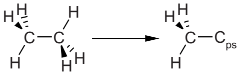
## 2 Motivation

### 2.1 Science Methodology

Biomedical research at every scale from molecules to organisms often involves numerical experimentation and hypothesis testing. When a parameter space is searched for optimal solutions, the computational requirements are amplified by several orders of magnitude. The simulation of extended molecular systems such as solutions, materials, and biomolecules is challenging by itself: The large number of atoms, dynamical sampling of the conformational space, and fine spatial and temporal resolution all require sophisticated techniques to correctly model physical and chemical behavior.

Hybrid quantum mechanics-molecular mechanics (QM/MM) methods, however, just describe a small, “active” region by accurate techniques based on the Schroedinger equation, while the surrounding larger, “inactive” region is treated with more approximate classical force fields. Unfortunately, these two physical concepts are so different that they cannot be easily coupled. In particular, when chemical bonds are cut between both parts, dangling bonds in the MM region can simply be eliminated, but the outermost atoms of the QM region would become radicals and behave completely different than required. One way to saturate these atoms includes the first atom of the MM part as a capping atom in the QM computation, parameterized such that it reflects the properties of the cut bond. This method does not require extended changes in the source code and does not lead to problems with artificial link atoms.

Zhang et al. recently developed such a “pseudobond” approach by adding an effective core potential (ECP) to a fluorine atom to model the methyl group in a carbon-carbon single bond [1]. Their parameterizations were done for the ethane molecule



(1)

($C_{ps}$ = F with pseudopotential), then tested on ethane derivatives, and later applied to enzymatic reactions. Unfortunately, we found serious instabilities in this pseudoatom ECP. Furthermore, it appears to be rather difficult to parameterize due to the diversity of the target properties and the multi-dimensionality of function and parameter space.

Therefore, we are developing a new effective pseudoatom potential [2], which only deals with the discrepancies between the isoelectronic $CH_3$ and F groups without exchanging the core. It is thus named “group difference potential” (GDP). This also provides direct information about substituent effects of functional groups, of interest from synthetic chemistry to drug development, for an important example case. In addition, the potential may be gradually switched on or off later, facilitating QM free energy difference determinations. A superposition of two Gaussian functions

$$U_{\text{eff}}(r) = A_1 \exp\left(-B_1 r^2\right) + A_2 \exp\left(-B_2 r^2\right) \tag{2}$$

turned out to be the most appropriate functional, a format already implemented in many quantum chemical program codes. This form corrects for differences in electron interaction and basis set through negative (attractive) and positive (repulsive) values for the amplitude coefficients $A_1$ and $A_2$. The positive exponential prefactors $B_1$ and $B_2$ specify the radial extent of each term around the fluorine atom; the smaller their value, the less compact the corresponding function, and vice versa.

Again, the prototype molecule ethane is examined here, but we plan to extend this concept to more complicated systems later. To analyze the results for each parameter set, a cost function is required that reduces the differences between the properties of ethane and pseudoethane to a single number. Bond lengths, bond angle, dissociation energy, Mulliken overlap populations and atomic charges were selected as independent properties. We applied the B3LYP/6-31G(d) level of theory with both spherical and cartesian basis set formats. This results in four calculations for every $A_1$, $A_2$, $B_1$, $B_2$ set, two on each pseudoethane and the corresponding pseudomethyl radical. All computations were done with the GAMESS program package [3]. The 32 properties identified (hydrogen values appear three times) have diverse units, accuracy, and importance. As such, the differences between actual and target properties $x_i$ and $X_i$ must be weighted appropriately by applying the normalized least squares expression

$$f\left(A_1, A_2, B_1, B_2\right) = \frac{1}{\sum_{i=1}^{32} w_i} \sum_{i=1}^{32} w_i \left( \frac{x_i - X_i}{u_i} \right)^2 . \tag{3}$$

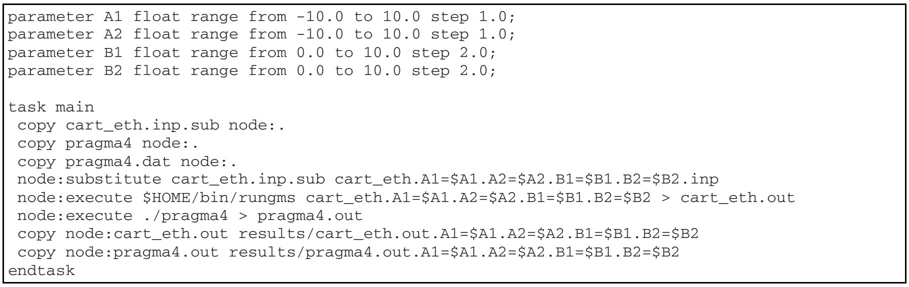
The “weighting” factors $w_i$ correct for the number of occurrences of each feature. The “unifying” factors $u_i$ reflect their individual accuracy and are chosen from chemical intuition. Equation (3) is evaluated after completion of each tuple of GAMESS jobs, so that the $w_i$ and $u_i$ values can be easily adjusted for later parameter optimizations.

To identify interesting low-cost regions and avoid trapping in local minima, we first scanned a portion of the parameter space in its entirety. This task consists of huge numbers of short, uncoupled QM calculations und hence is a perfect computing grid application. In the initial experiment on the 4th Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) workshop, the four variables were varied between –10 and 10 a.u. in steps of 1 a.u. for $A_1$ and $A_2$, and between 0 and 10 a.u. in steps of 2 a.u. for $B_1$ and $B_2$. Ignoring function symmetry for now, this leads to 15,876×4 jobs. On the Supercomputing (SC) conference 2003, we performed even larger parameter sweeps with 53,361×4 and 60,016×4 individual calculations.

## 2.2 Grid Methodology

The problem formulation above dictates a $A_1$, $A_2$, $B_1$, $B_2$ parameter set that minimizes the cost function value (3). This implies running GAMESS repetitively over a cross product of all values under consideration, resulting in tens of thousands of independent jobs. To perform this by hand on a single machine would be lengthy, and manually on a distributed computational grid, nearly unfeasible and error prone. Therefore, we invoked the Nimrod/G tool [2,4], which has been specifically designed to perform parameter sweeps using resources distributed across a computational grid [5]. Nimrod/G manages the experiment by finding suitable machines, sending input files to them, running a computation, and shipping the output files back to a central computer. The software also handles common events such as network and node failures.

Nimrod/G targets wide area networks as characterized by the global grid. At the system core is a database that stores all of the experiment details. Jobs are scheduled by considering constraints such as a soft real time deadline and the costs of various resources, and notionally allocating jobs to machines. They are actually executed by "agents", which themselves run on the various resources and request jobs. This architecture hides the latency for scheduling and invoking a remote computation. Nimrod/G is built on a variety of middleware layers, such as Condor, Legion, and Globus. Thus it only needs to interact with the uniform scheduling interface and security layer provided to the testbed resources regardless of their architecture, operating system, or configuration by, for example, the most widely deployed toolkit Globus.



```
parameter A1 float range from -10.0 to 10.0 step 1.0;
parameter A2 float range from -10.0 to 10.0 step 1.0;
parameter B1 float range from 0.0 to 10.0 step 2.0;
parameter B2 float range from 0.0 to 10.0 step 2.0;

task main
 copy cart_eth.inp.sub node:.
 copy pragma4 node:.
 copy pragma4.dat node:.
 node:substitute cart_eth.inp.sub cart_eth.A1=$A1.A2=$A2.B1=$B1.B2=$B2.inp
 node:execute $HOME/bin/rungms cart_eth.A1=$A1.A2=$A2.B1=$B1.B2=$B2 > cart_eth.out
 node:execute ./pragma4 > pragma4.out
 copy node:cart_eth.out results/cart_eth.out.A1=$A1.A2=$A2.B1=$B1.B2=$B2
 copy node:pragma4.out results/pragma4.out.A1=$A1.A2=$A2.B1=$B1.B2=$B2
endtask
```

**Fig. 1.** Shortened plan file used for the PRAGMA4 pseudopotential scan (one instead of four GAMESS jobs)

To create a computational experiment, the user builds a "plan" file like the one in Fig. 1. Plan files are fairly small and declarative in nature; thus new parameter sweeps can be set up very quickly. In the first part, they contain a definition of the parameters and their ranges. Parameters may be integers, floating point numbers or text. The second part is the “task” block. This set of commands is executed by the “agent” component for each parameter set; it includes the call of GAMESS here for example.

From the implementations of the Nimrod/G user interface we used the portal for this work. This web site enables the user to specify the available resources, as well as to setup and control an experiment through a conventional browser, without porting Nimrod/G to the client machine. Since GAMESS is available on various platforms, we were able to we built testbeds containing conventional workstations, clusters as well as vector supercomputers. These spanned a range of countries, organizations, administrative domains, queue managers, operating systems and architectures (for details, see Ref. [2]). In practice, some of the machines did not perform any computations, either due to their work load or software configuration problems. Fortunately, the dynamic nature of the grid allows deferring the decision about resource usage until execution time. Once all jobs have completed and the output files are returned to the user, the results need to be collapsed into an interpretable form. Here we used the scientific visualization package OpenDX. To explore the entire surface, we produced a sequence of visualizations, each showing isosurfaces of cost function value across three of the parameters, and a different frame for each value of the fourth parameter.

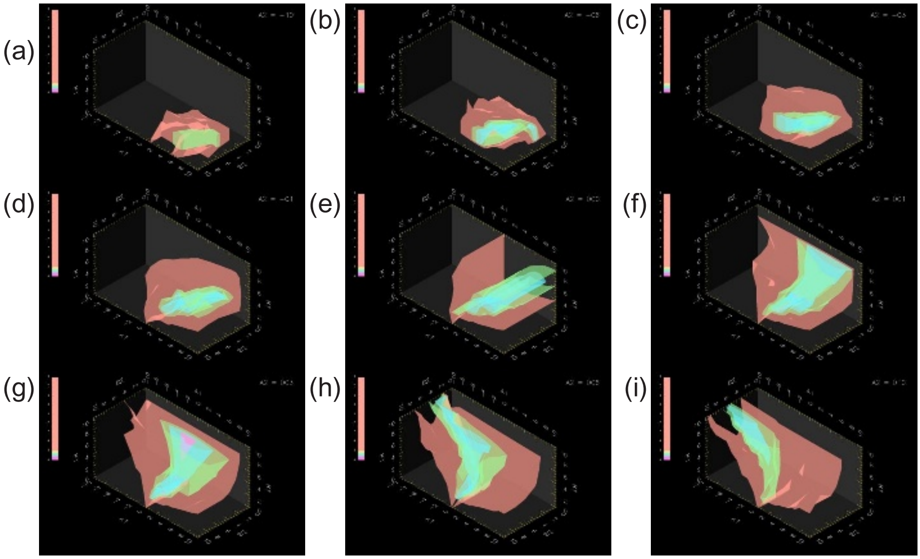


**Fig. 2.** Selected pictures of the GDP parameter space scanned in the PRAGMA4 experiment. $A_1$, $B_1$, and $B_2$ are displayed on the axes; $A_2$ = ±10, ±6, ±3, ±1, 0 evolves with the snapshots. Successively better isosurfaces of the cost function *f* are drawn (*red* = “zero” model)

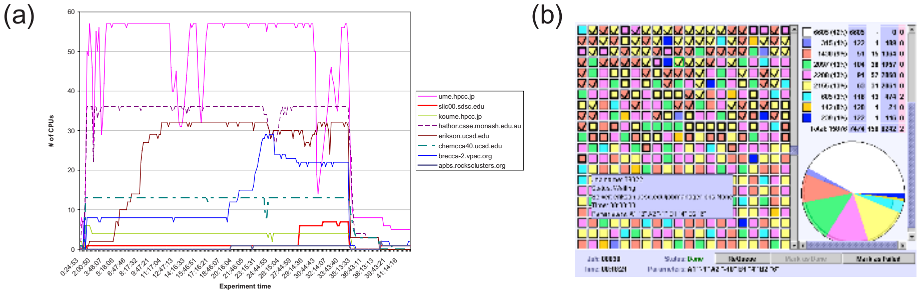


**Fig. 3.** (a) Distribution of CPU usage over the grid resources in the PRAGMA4 experiment. (b) Nimrod/G portal status display

## 3 Results and Discussion

### 3.1 Science Results

Images from the initial parameter scan at PRAGMA4 are displayed in Fig. 2. They demonstrate the complexity of the cost function hypersurface. To find starting points for subsequent optimizations, regions with values lower than the “zero” model (no or canceling GDP) are of special interest. Several such “local” minima are scattered over the parameter space with no apparent pattern, although further analysis suggests partial linear dependence. The most significant minimum here shows up in the middle of the scan around Fig. 2(g). However, all cost function values are still too high.

Therefore, in the SC2003 experiments we also performed sweeps with logarithmic point distribution and higher density in the most interesting region. The deepest minimum appears when a medium-size repulsive and a diffuse attractive Gaussian function are combined to build a maximum at the fluorine core and a shallow depression closer to the bound carbon atom. We tentatively attribute this to the larger size and smaller electron attraction of a methyl compared to the fluoro group. With these data collections in hand it also turned out that the significance of the currently best “global” minimum can be remarkably improved by reducing the bond angle unit factor, a fact that cannot be easily deducted from chemical reasons. This thus reveals the most promising parameter region and weighting for later GDP optimizations.

### 3.2 Grid Results

Fig. 3(a) visualizes the resource utilization during the PRAGMA4 experiment. Each curve represents a different machine, and shows the number of jobs running at any instant. The graph conveys the ability of the grid to dynamically adjust which resource provides a particular service to the available capacity. Nimrod/G leverages this by incorporating scheduling heuristics that allow moving load to meet soft deadlines. Most importantly, we were not able to accumulate the number of processors required to complete this work within 42 hours at any one of the sites. Overall, we executed over

200 days of processing. Due to the larger number of CPUs and the longer time frame, this amount was even multiplied during the later SC2003 experiments. Careful evaluation of the resource job statistics (see Ref. [2]), however, shows that although some machines had more executing jobs and provided more execution time, they did not produce more results, because others executed jobs faster and with fewer processors. Fig. 3(b) shows the portal status display in operation. Using this interface it is possible to see where individual jobs are running, and to diagnose any problems.

Although all experiments were very successful, we had some problems in setting up these large grid testbeds. Apart from network related and individual server issues, the biggest difficulty represented miss-configured Globus installations and bugs within Globus itself. We developed workarounds for these issues, described in Ref. [2].

## 4 Conclusions

The shift in science towards information-driven research enables computational studies coupled more tightly to experiment. The rapid growth of grid technologies facilitates the combination of software, data and analysis tools, and the development of grid-enabled chemistry and biology codes for complex problem solving. Linking together sophisticated methodologies as exemplified in this work facilitates new integration pathways to discovery, which can be automated and repetitively performed with variant input datasets. Additionally, end-to-end audit of the process is an implicit deliverable, i.e., the scientist has a record of every action performed on the data.

In collaboration with several international groups, the highlighted project illustrates access to global resources and application technologies via web interfaces. The goal is to develop computational capabilities which integrate our knowledge in (bio)chemistry, molecular modeling, experimental characterization, visualization and grid computing. Here, the GAMESS quantum chemistry software and the Nimrod/G grid distribution tool were coupled. Our purpose was the parameterization of a GDP pseudopotential, which describes the differences of a fluoro compared to a methyl group in the pseudoethane molecule for ultimate use as capping atom potential in QM/MM calculations. We developed a simple GDP formulation with four parameters and constructed a flexible cost function to measure their goodness. It was scanned on an array of points within a defined parameter space region. The resulting cost function hypersurface was further refined by parameter sweeps with different point distributions.

Subsequently, we plan to use the most suitable parameter combinations to start minimization runs. The Nimrod/O tool that performs automatic optimization [4], will be incorporated to search the parameter space and to find the final GDP. We already began to vary variables in the least squares procedure to generate a more funnel-like hypersurface. This will allow minimizations commencing from any remote place to travel towards the global minimum more directly. Overall, this procedure considerably reduces the development time of GDPs for further molecules and groups.

A second purpose of this study was to show how the middleware tool, Nimrod/G, significantly enhances scientific options. The up to 60,016×4 uncoupled QM calculations are systematically generated for a multidimensional grid of points, optimally distributed over several computing clusters, within a few days. The completion of such numbers of runs would not have been possible in a reasonable timeframe without such technology. The results allow a better conceptualization of the parameter optimizations, thereby providing more insight into the physics. Late failure of parameterizations can be improved and even the optimization procedure itself can be streamlined.

The described technology has been previously applied in other sciences, but is relatively new to quantum chemistry. One can imagine wider application, such as analysis of reactions, generation of algorithms, or cross-correlation of data. Related examples include parameterization of basis sets, force fields, and similar entities in computational chemistry. A classical case is the examination of potential energy surfaces. Similar approaches can also be used to scan large compound databases in high-throughput virtual screening. By integration into grid middleware architecture QM applications previously not feasible become doable. Furthermore, such infrastructure will help to tie computation with investigator intuition regardless of location, to facilitate scientific investigations by exploiting novel grid capabilities and teraflop hardware speeds, enabling direct user input and feedback. Such infrastructure will impact scientists that need such tools for interdisciplinary research. This will in turn foster development of new modeling, data, and computational science technologies.

**Acknowledgements.** We thank PRAGMA and all involved organizations, in particular the SDSC ROCKS group. We are grateful to J.P. Greenberg and K. Thompson, SDSC, for installing GAMESS and Globus. W.S. acknowledges support by J.A. McCammon, UCSD, and by a postdoc fellowship of the German Academic Exchange Service (DAAD). A part of the machines was sponsored by the National Biomedical Computational Resource (NBCR) and the W.M. Keck Foundation. K.B. acknowledges support from the NSF through DBI-0078296 and ANI-0223043 and from the NIH through NBCR-RR08605. We also thank D. Kurniawan from Monash University for the visualizations of the GAMESS output. The Nimrod project is supported by DSTC and GrangeNet, both of which are funded in part by the Australian Government. D.A. received financial support from the Australian Partnership for Advanced Computing (APAC) whilst on leave at UCSD.

## References

1. Zhang, Y., Lee, T.-S., Yang, W.: A Pseudobond Approach to Combining Quantum Mechanical and Molecular Mechanical Methods. J. Chem. Phys. **110** (1999) 46-54
2. Sudholt, W., Baldridge, K.K., Abramson, D., Enticott, C., Garic, S.: Parameter Scan of an Effective Group Difference Pseudopotential Using Grid Computing. New Generation Computing **22** (2004) 125-136
3. Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.H., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S.J., Windus, T.L., Dupuis, M., Montgomery, J.A.: General Atomic and Molecular Electronic-Structure System. J. Comput. Chem. **14** (1993) 1347-1363; http://www.msg.ameslab.gov/GAMESS/GAMESS.html
4. Abramson, D., Sosic, R., Giddy, J., Hall, B.: Nimrod: A Tool for Performing Parametised Simulations Using Distributed Workstations. The 4th IEEE Symposium on High Performance Distributed Computing, Virginia (August 1995); Abramson, D., Giddy, J., Kotler, L.: High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid? International Parallel and Distributed Processing Symposium (IPDPS), Cancun, Mexico (May 2000) 520- 528; http://www.csse.monash.edu.au/~davida/nimrod/
5. Foster, I., Kesselman, C. (eds.): The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers, USA (1999)