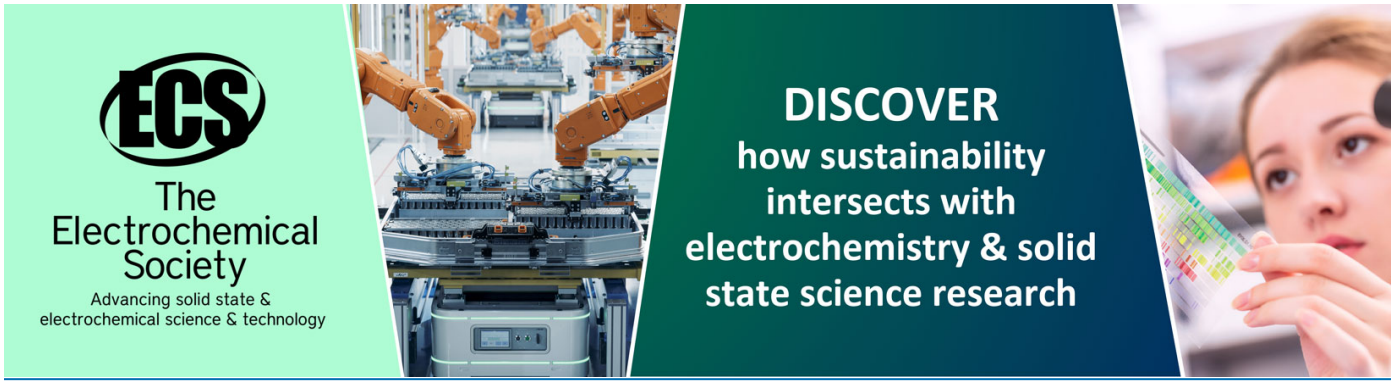
Biomolecular simulations on petascale: promises and challenges



Institute of Physics Publishing Journal of Physics: Conference Series **46** (2006) 327–333
doi:10.1088/1742-6596/46/1/046 SciDAC 2006

Biomolecular simulations on petascale: promises and challenges

Pratul K. Agarwal[*,a,b] **and Sadaf R. Alam**[a,c]

[a]Computer Science and Mathematics Division, [b]Computational Biology Institute,
[c]Future Technologies Group, Oak Ridge National Laboratory, Oak Ridge, Tennessee

*agarwalpk@ornl.gov

Abstract. Proteins work as highly efficient machines at the molecular level and are responsible for a variety of processes in all living cells. There is wide interest in understanding these machines for implications in biochemical/biotechnology industries as well as in health related fields. Over the last century, investigations of proteins based on a variety of experimental techniques have provided a wealth of information. More recently, theoretical and computational modeling using large scale simulations is providing novel insights into the functioning of these machines. The next generation supercomputers with petascale computing power, hold great promises as well as challenges for the biomolecular simulation scientists. We briefly discuss the progress being made in this area.

1. Multi-scale Modeling of Protein Structure, Dynamics and Function

Proteins participate in a variety of biochemical processes within the living cell. The structural organization of proteins consists of multiple levels: first the atoms of proteins are organized in repeating units called residues; second level organization consists arrangement of protein residues into local structures such as α-helices, β-sheets and loop regions; and at the third level, the entire protein is organized into characteristic three-dimensional shape (also known as *protein fold*) that is related to its function. The intrinsic dynamics and function of proteins span multiple scales of time. Intrinsic dynamics refers to the internal motions that occur within the protein at different time-scales, ranging from femtosecond to second and longer. The role of protein structure in protein function, such as enzyme catalysis, has been known for more than a century. As the well known "lock-and-key" hypothesis suggests, the direct interactions between enzyme and substrate have been investigated to understand the catalytic mechanisms. More recently, the internal motions of the proteins have also been implicated in the protein function such as enzyme catalysis. Experimental techniques, including X-ray crystallography, nuclear magnetic resonance (NMR), neutron scattering, biochemical and mutation studies, continue to provide some details at selective time-scales for protein dynamics and its link to protein function.

Theoretical and computational modeling of proteins based on large scale molecular dynamics (MD) simulations continue to provide novel insights, particularly in the link between structure, dynamics and function [1-5]. Further, the role of hydration-shell and bulk solvent as well as temperature effects in enzyme mechanisms are now being understood [3]. Figure 1 depicts the wide range of time-scales for activity of several enzymes. The fastest enzyme performs its function* over a billion times per second, while slower enzymes can take seconds or longer to complete one cycle. It is

* substrate turnover step, excluding the binding and release of substrate/product and cofactor

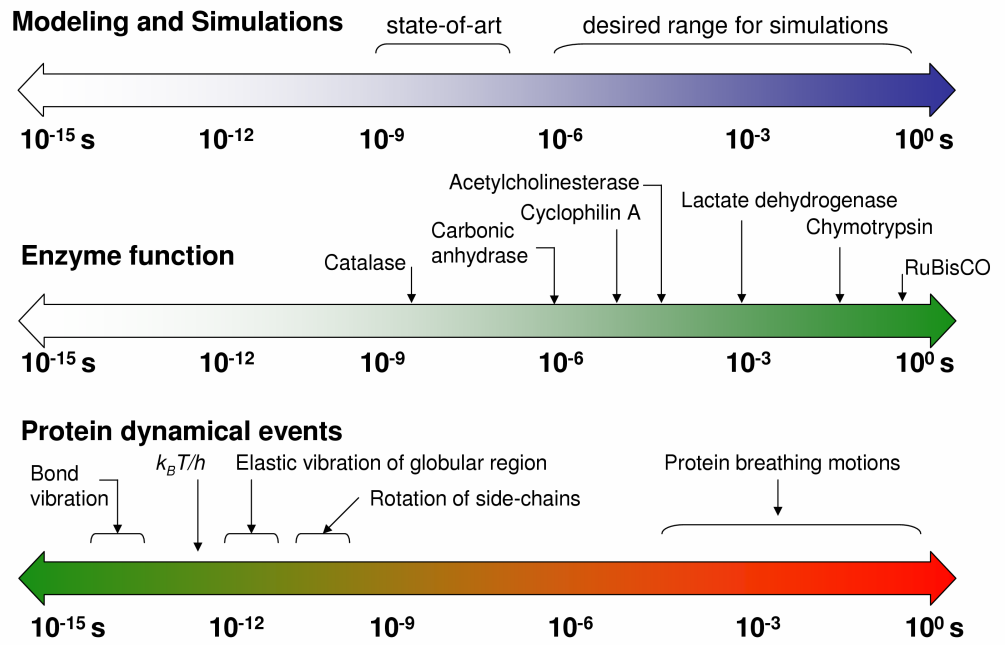


Figure 1: Multiscale modeling and simulations in biology. Computational biochemists/biophysicists are now regularly using molecular simulations to investigate enzyme complexes. The structure, dynamics and function of enzyme complexes spans multiple scales of time and length. Enzymes catalyze biochemical reactions as fast as billions of times per second on side of the range, while on the side they can take seconds or longer for one catalytic cycle. The wide range of internal protein motions occur on 10^{-15} to $>10^{0}$ seconds, which are linked to a variety of protein functions (such as enzyme catalysis) on similar time-scales. However, the current simulations fall short by several orders of magnitude. The typical state-of-art simulations can only reach 10^{-8} seconds at best for a real biological system, while the desired time-scale is 10^{-6} to 10^{0} seconds or higher.

interesting to note that a wide range of internal protein motions also occur on similar time-scales as the enzyme function, therefore, raising the interesting question whether enzyme dynamics and function are interrelated or not. Currently, computational biochemists and biophysicists can only simulate a fraction of the biologically relevant time-scales for most enzymes. Most common MD simulations on a single workstation or using a small PC-cluster explore nanosecond (10^{-9} s) time-scale for a medium size protein in aqueous environment consisting of 5,000-20,000 atoms. Supercomputers can simulate around 100 nanoseconds for multi-million atom systems. These simulations continue to provide novel insights into enzymes.

Computational modeling of enzyme cyclophilin A has led to an interesting discovery that the internal protein dynamics of this enzyme is linked to its peptidyl-prolyl *cis/trans* isomerization catalytic activity [1-3]. These modeling studies have identified protein vibrational modes that occur at the time-scale of the reaction and play a role in promoting catalysis. Detailed analysis has led to the discovery of a network of protein vibrations promoting enzyme catalysis in cyclophilin A (see Figure 2). The network is formed by protein residues connecting the surface regions of the enzyme, all the way to the active-site and passing through internal regions. The vibrations impact catalysis by picking up energy required for the reaction, from the thermo-dynamical energy of the solvent and transferring it to the active-site. In the transition state theory framework, these vibrations play a role in decreasing the activation energy barrier as well as promoting more reaction trajectories to successfully cross the transition state. The presence of these vibrational modes as well as the network has been recently

verified by NMR studies [6]. Moreover, genomic and structural analysis indicates that the network of protein vibrations is a conserved part of the overall shape of enzymes (protein fold) catalyzing the peptidyl-prolyl *cis/trans* isomerization reaction. Computational modeling of another enzyme dihydrofolate reductase, which catalyzes transfer of hydride, has also led to the discovery of a network of coupled protein motions promoting the catalytic step [5]. Similar to the enzyme cyclophilin A, the network is formed by residues starting at the surface and reaching into the active-site. These studies lead to identification of residues that are dynamical contributors to catalysis. The role of protein residue isoleucine14 (a part of the discovered network of vibrations) in the catalytic step, has been confirmed by biochemical and mutation studies [7]. The genomic and structural of dihydrofolate reductase has also shown that the dynamically contributing residues are conserved across various species ranging from bacteria to human.

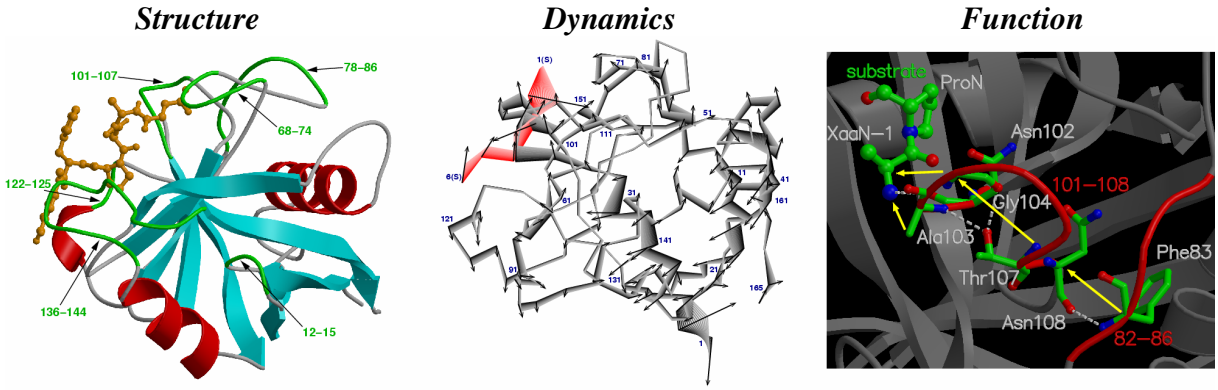


Figure 2: An integrated view of protein structure, dynamics and function: Multi-scale theoretical and computational modeling of enzyme cyclophilin A has provided into insights into understanding the role of protein structure and internal dynamics in enzyme catalysis. These studies have lead to the discovery of a network of protein vibrations promoting enzyme catalysis [1-4].

Multi-scale modeling of proteins is required to understand protein structure that spans multiple scales of length and organization as well as dynamics and function, which spans multiple scales of time. Another aspect of proteins, which has generated considerable interest from computational molecular biologists, is the determination of 3-dimensional protein structure from the primary sequence and investigating protein folding, the process by which proteins fold into their native or functional shape. The process of protein folding also involves wide range of time-scales, with faster events of local structure folding occurring at picosecond-nanosecond time-scales while the overall process taking between milliseconds to seconds. As Figure 1 indicates, the commonly investigated time-scale for MD simulations is nanosecond, which falls 4-6 orders of magnitude short of the desired time-scale of biological activity. It has been suggested that the computational requirements for multi-scale modeling of a medium size protein can be as high as 1 petaFLOP/s[†] for an entire year [8].

2. Getting ready for Petascale

Atomistic biomolecular simulations, based on Newtonian mechanics (also referred to as classical or molecular mechanics), are currently the most common computational method that are used for protein/enzyme modeling. Simulations that investigate behavior of system with evolution of time, through time-integration techniques, are referred to as molecular dynamics or MD simulations. For many years the lack of sufficient computing power has been suggested as a limiting factor for simulation of biologically relevant time-scales. In the coming years, the computing power is expected to grow by several folds as the arrival of petascale machines is around the corner. These new

[†] 1 petaFLOP/s = 10^{15} floating point operations per second

supercomputing machines hold a great promise for the biomolecular simulations. However, it is becoming clear that the design of MD software packages has inherent imitations and these codes are not able to scale beyond a few hundred processors. The petascale machines are expected to have thousands of processors, therefore, the lack of proper scalability will continue to hinder the progress in this area. The ability to utilize petascale computing power to simulate biologically relevant time-scales will require significant advances in simulation code design, efficiency, scalability, stability and portability on different hardware platforms.

To investigate the scalability of biomolecular simulation codes on the petascale machines, we are currently studying the scaling behavior of popular MD applications on two contemporary massively parallel systems: IBM Blue Gene/L and Cray XT3 [9, 10]. Although the two systems have similar distributed-memory architecture, the processor and memory configurations of the two systems are substantially different. We have recently demonstrated that the scaling limits stem from the underlying implementation and algorithmic characteristics of the applications [9, 10]. For instance, AMBER's [11] particle-mesh Ewald (PME) method [12] based simulations performed using the sander module do not scale beyond 128 processors on the two systems. We have also observed that the XT3 system provides an order of 3-6 times improved simulation performance as compared to the Blue Gene/L system with similar processor count. At the same time however, another MD code LAMMPS [13] that uses domain decomposition technique, shows better scaling behavior. We have successfully used LAMMPS on over 16,000 processors of IBM Blue Gene/L to simulate a biomolecular system with over 290,000 atoms. On Oak Ridge National Laboratory's Cray XT3, the simulation of same biomolecular system yielded over 6 nanoseconds/day simulation performance on 4096 processors (see Figure 3), which is a significant performance milestone. Note that state-of-art simulations on previous systems gave a performance metric of 0.5-1.0 nanoseconds/day for the same biomolecular system. The use of 4096 Cray XT3 processors (~20 teraFLOP/s) allows simulation time-scales that are an order of magnitude better than previously achieved. The petascale machines in the near future will provide >1000 teraFLOP/s of computing power, which is 50 times more computing power; therefore, hold great promise for biomolecular simulations.

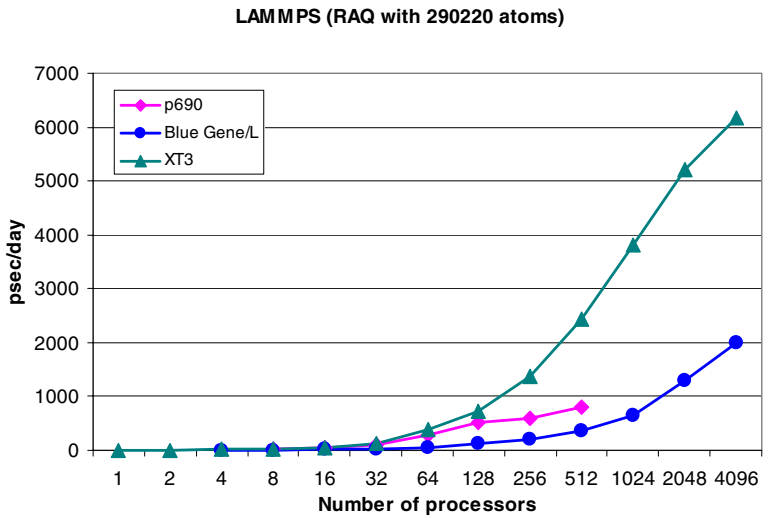


Figure 3: Biomolecular simulations of an enzyme system using LAMMPS. The model of enzyme RuBisCO consisted of 290220 atoms including explicit solvent. The scaling performance was investigated on 3 different platforms.

3. Improving Biomolecular Simulations

The breakthrough in biological investigations based on large scale simulations can come from several avenues. Availability of MD software that scales on massively parallel processing (MPP) systems will have the most impact. The scalable software will enable simulation of the biologically relevant time-scales, microsecond and longer, on a regular basis. Note, a number of simulation trajectories are typically required to get information that is representative of the native ensemble. The popular MD software packages such as CHARMM [14] and AMBER provide a large number of functionalities that allow investigations of various aspects of biomolecules. However, the underlying code in these

packages has been developed over several decades and the programming model on the state-of-art machines have evolved considerably. Unfortunately, the design of the software packages has not kept up with the design of the MPP hardware. Therefore, it is not surprising that these software frameworks do not scale beyond a few hundred processors. There is an urgent need to improve the scaling of the codes so that the users can continue to benefit not only from the wide functionality provided in these packages but also the petascale computing resources.

Novel theoretical and computational techniques that allow characterization of biomolecular structure, dynamics and function over multiple scales of time and length will particularly make a significant impact. Knowledge at multiple scales is crucial for better understanding of biochemical processes. Increase in computing power that allows calculations at a faster rate holds the key to enable longer time-scale simulations, while the improved memory capabilities are essential to accommodate length scales. At the same time however, one cannot overlook the communication overheads that result from mapping and distribution of workload on MPP resources. AMBER framework for instance has an underlying workload and distribution scheme in its popular program sander that inhibits scaling of computation and physical memory requirements beyond 128 processors [9, 15]. Algorithms that overcome these limitations will be critical in overcoming these scaling bottlenecks.

In addition to the scaling limitation, a very large number of processors pose system-level challenges such as stability, availability and fault-tolerance. Failures at the hardware and software levels are expected every few hours on petascale systems with over 100,000 processors. Note, the exact figure for mean-time-between-failures is still a topic of debate by the engineers. Crucial time is lost in detecting failed simulations due to hardware or system level software failure, and restarting the application. One of the realities of using supercomputing resources is that these machines are shared between other applications. Therefore, failure will also lead to time lost in the queue. These failures will have drastic impact on performance of simulations or time-to-solution and therefore the scientific outcomes. Strategies are needed to detect and handle failures at runtime (fault-tolerance) within the simulations to minimize their impact, and to improve dynamic reconfiguration capabilities. Currently, we are investigating in-built fault-tolerance within MD simulation codes and approaches that will allow codes to achieve high scalability, such as: transaction based approach; optimization based on network topology approach; and global and local master based approach. These approaches aim at not only targeting the scaling barrier of biomolecular simulations but also creating a coherent mechanism to carry out multiple trajectory simulations on very large-scale systems.

4. Multi-paradigm hardware

It is becoming evident from the latest technological advances in the microprocessor industry as well as the increasing heat generation and power requirements of off-the-shelf cluster systems that the future high-end computing systems will no longer continue to enjoy the benefits from increasing clock speeds of the microprocessors. Instead, emerging technologies such as multi-core systems, unconventional processing devices (e.g. the Cell system) and accelerators like Field Programmable Gate Arrays (FPGAs) will provide the much needed computing capabilities. For instance, the 50 teraFLOP/s capability at Oak Ridge National Laboratory's Cray XT3 system in August 2006 will be achieved from dual-core Opteron processors. We are in process of investigating the impact of shared resources in dual and multi-core systems on popular MD applications including AMBER, CHARMM and LAMMPS [16]. The preliminary results show that the shared resources between multi-core systems have a limited affect on the performance on MD applications. Typically 80-90% increase in performance is achievable for each additional core. For larger simulations systems, memory requirements become a limiting factor. Nonetheless, we anticipate the performance loss can be minimized by exploring and subsequently incorporating appropriate memory and processor affinity techniques.

FPGA and other accelerator devices including widespread Graphical Processing Units (GPUs) provide high performance processing capabilities in an extremely cost and power efficient manner. However, there are challenges in exploiting these devices for scientific computation such as the ones which are at the core of MD simulations. The programming paradigms and languages that are widely used for programming these devices are domain-specific; only a few high-level programming

interfaces are available—most of these software stacks are either vendor or target system specific. Another issue is the limited availability and support for double-precision floating-point operations. Presently one vendor, SRC Computers Inc., provides support for Fortran programming. Fortran is widely used in MD simulation packages. In order to exploit the additional computing power and to investigate the feasibility of FPGA devices for large-scale MD applications, we have ported the PME method of AMBER 8.0 on the SRC FPGA devices [17]. Our initial results provide an order of magnitude speedup on the microprocessor based solutions with single-precision floating-point calculations. Moreover, we observed that the inherent parallelism in FPGA devices result in sustained performance for biological systems simulations with large number of atoms. On the microprocessor based systems, the computation cost and subsequently the simulation times increase almost exponentially by increasing system sizes. Since the capabilities and performance on FPGA devices are increasing at a much faster rate than the mainstream microprocessor systems, we anticipate that the future FPGA devices are capable of breaking the scaling barriers for the biomolecular simulations. Note, an additional benefit of these devices over petascale systems is that these devices can provide unprecedented processing capabilities to simulation scientists that have access to limited resources such as the desktop systems or small clusters.

Summary

The tremendous growth of available computing power continues to make impact on the biomolecular simulations, enabling vital scientific breakthroughs in the area. It is expected that the available computing power in the coming decade will be orders of magnitude more than what is currently available. This is expected to come not only from the traditional supercomputer but also new developments in the computer hardware such as the development of FPGA and dual/quad core processors. Unfortunately, the popular MD packages that are widely in use will not be able to efficiently utilize these resources. The need of the hour is development of new highly scalable biomolecular simulations codes. Also the community will benefit from closer interaction between biologists, simulation scientists and computer scientists.

References

[1] P. K. Agarwal, A. Geist, A. Gorin (2004), "Protein Dynamics and Enzymatic Catalysis: Investigating the Peptidyl-Prolyl cis/trans Isomerization Activity of Cyclophilin A", *Biochemistry*, **43**, 10605-10618.

[2] P. K. Agarwal (2004), "Computational studies of the mechanism of cis/trans isomerization in HIV-1 catalyzed by cyclophilin A", *Proteins*, **56**, 449-463.

[3] P. K. Agarwal (2005), "Role of Protein Dynamics in Reaction Rate Enhancement by Enzymes", *J. Am. Chem. Soc.,* **127**, 15248-15246.

[4] P. K. Agarwal (2006), "Enzymes: An integrated view of structure, dynamics and function", *Microbial Cell Factories*, **5**:2.

[5] Agarwal PK, Billeter SR, Rajagopalan PTR, Benkovic SJ, Hammes-Schiffer S. (2002) "Network of coupled promoting motions in enzyme catalysis", *P. Natl. Acad. Sci. USA,* **99**, 2794-2799.

[6] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay and D. Kern (2005), "Intrinsic dynamics of an enzyme underlies catalysis", *Nature*, **438**, 117-121.

[7] Schnell JR, Dyson HJ, Wright PE. (2004), "Effect of cofactor binding and loop conformation on side chain methyl dynamics in dihydrofolate reductase", *Biochemistry,* **43**, 374-383.

[8] F. Allen, G. Almasi, W. Andreoni, D. Beece, B. J. Berne, A. Bright, J. Brunheroto, C. Cascaval, J. Castanos, P. Coteus, P. Crumley, A. Curioni, M. Denneau, W. Donath, M. Eleftheriou, B. Fitch, B. Fleischer, C. J. Georgiou, R. Germain, M. Giampapa, D. Gresh, M. Gupta, R. Haring, H. Ho, P. Hochschild, S. Hummel, T. Jonas, D. Lieber, G. Martyna, K. Maturu, J. Moreira, D. Newns, M. Newton, R. Philhower, T. Picunko, J. Pitera, M. Pitman, R. Rand, A. Royyuru, V. Salapura, A. Sanomiya, R. Shah, Y. Sham, S. Singh, M. Snir, F. Suits, R.

Swetz, W. C. Swope, N. Vishnumurthy, T. J. C. Ward, H. Warren, R. Zhou, and I. B. G. Team (2001), "Blue Gene: A vision for protein science using petaflop supercomputer", *IBM Syst. J.,* **40,** 310-327.

[9] S. R. Alam, P. K. Agarwal, Al Geist J. S. Vetter (2006), "Performance Characterization of Molecular Dynamics Techniques for Biomolecular Simulations," *Proc. ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP).*

[10] C. P. Sosa, P. K. Agarwal, S.R. Alam, R. Smith, D.A. Case, M. Crowley, "Molecular Dynamics Performance Analysis on the Massively Parallel Supercomputer Blue Gene/L: AMBER8", *Under review.*

[11] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman (1995), "AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules", *Comput. Phys. Commun.,* **91**, 1-41.

[12] T. Darden, D. York and L. Pedersen (1993), "Particle mesh Ewald--an Nlog(N) method for Ewald sums in large systems", *J. Chem. Phys*. **98**, 10089-10092.

[13] S. J. Plimpton (1995), "Fast Parallel Algorithms for Short-Range Molecular Dynamics", *J. Comp. Phys.,* 117, 1-19; http://www.cs.sandia.gov/~sjplimp/lammps.html

[14] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus (1983) "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations", *J. Comput. Chem.*, **4**, 187-217.

[15] M. F. Crowley, T. A. Darden, T. E. Cheatham, and D. W. Deerfield, "Adventures in improving scaling and accuracy of a parallel molecular dynamics program" (1997), *J. Supercomput.,* **11** 255-278.

[16] S. R. Alam, R. F. Barrett, J. A. Kuehn, P. C. Roth and J. S. Vetter, "Characterization of Scientific Workloads on Systems with Multi-core Processors," *Submitted to 2006 IEEE Symposium on Workload Characterization.*

[17] S. R. Alam, P. K. Agarwal, D. Caliga, M. C. Smith and J. S. Vetter, "Acceleration of Bio-molecular Simulations on Field Programmable Gate Arrays using High-level Languages," IEEE Computer special issue on high performance reconfigurable computing and applications. *Manuscript under preparation.*