Toward Teraflop Computing
and New Grand
Challenge Applications

ISBN-1-56072-247-9

MOLECULAR DYNAMICS ON PARALLEL COMPUTERS: APPLICATIONS FOR THEORETICAL BIOPHYSICS

Thomas Bishop, Helmut Heller and Klaus Schulten
Beckman Institute and Departments of Chemistry and Physics,
University of Illinois at Urbana-Champaign,
405 North Matthews, Urbana, Illinois 61801 USA

Abstract

Molecular dynamics is a valuable aid for refining the observed structure of biopolymers, for understanding the structure–function relationship of biopolymers, and for the rational design of drugs. In order to accurately characterize biopolymers, it is often necessary to explicitly include solvent or other environments, namely water, ions or cellular membranes resulting in system sizes of tens of thousands of atoms. The resulting large systems can only be simulated on massively parallel computers.

A molecular dynamics program, EGO, has been developed to run on parallel computers in order to simulate large systems. EGO, which was written in OCCAM II and runs on transputers, has been used to conduct simulations of biologically significant systems. Two transputer systems, based on the INMOS T805 processor, have been employed for molecular dynamics: a custom built 60-node system and a Parsytec GCel-64 with 64-nodes. EGO has recently been rewritten in C to run under the machine-independent coordination languages Charm and PVM, so that simulations can be conducted on a workstation cluster.

The program EGO has been applied to simulations of a protein-DNA complex in a water envelope resulting in a system of 13,500 atoms. The simulations have required dedicated computational resources of one month per simulation. Results from the simulations indicate that the so-called DNA binding domain of the glucocorticoid receptor protein induces a bend of approximately 35° in the DNA.

1 INTRODUCTION

X-ray crystallographic techniques have revealed the structure of numerous biological structures. These techniques have been sufficiently developed so that larger and more complex systems, including bonded heterogeneous molecular structures, are becoming available. In fact any recent issue of *NATURE* is likely to contain a newly derived crystallographic structure of biological importance. These static views provide insight into the structural behavior of the molecules revealing common structural motifs that nature has exploited. However, there still exist experimental difficulties in crystallizing or analyzing a particular structure so that the structure analyzed is not

necessarily the one of interest, and the crystalline environment may not be the native environment of the system being studied. This is definitely the case for biological systems in which membranes, water, and ions are abundant in the native environment. Molecular dynamics, which relies on an approximate initial conformation, is therefore a complementary tool for refining these systems and for investigating the dynamical degrees of freedom [7,27]. In particular by explicitly including lipids, water, and/or ions the molecular structure of interest may be refined in a natural environment, thus removing the effects of crystal packing. A major drawback of this approach is that the explicit inclusion of the environment significantly increases the number of particles for a given simulation, and, therefore, demands the most powerful computers available, namely parallel computers.

2 MOLECULAR DYNAMICS

2.1 Theory

Molecular dynamics is a classical approach to describe the time evolution of a system of atoms, which remains a valid approximation as long as no covalent chemical bonds are created or destroyed over the course of the simulation. (For a review of molecular dynamics principles see [1]) In this approach, Newton's equations of motion are numerically solved with each atom being represented as a point mass. The Verlet algorithm [39] is most often used for the numerical integration. The position, r_i , of atom i at time $t + \Delta t$ is determined by

$$\left.\begin{aligned} r_i(t+\Delta t) &= 2r_i(t) - r_i(t-\Delta t) + \mathrm{F}_i(t)\frac{(\Delta t)^2}{m_i} \\ \mathrm{F}_i(t) &= -\nabla_i \mathrm{E}(r_1(t), \ldots r_N(t)) \end{aligned}\right\} \quad (1)$$

where $\mathrm{F}_i(t)$ denotes the force acting on the i-th atom, m_i and r_i denote the mass and position of the i-th atom, and N denotes the total number of atoms. ∇_i is the differential operator $(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial y_i}, \frac{\partial}{\partial z_i})^T$. The accuracy of the Verlet algorithm is determined by the time step, Δt, which must be shorter than the fastest motion. The most rapid motion is due to the vibration of hydrogen atoms along covalent bonds which requires a time per integration step of 0.25 fs. This limit is many orders of magnitude below the time scales of relevant processes in biochemistry such that structural molecular dynamics simulations usually require about 10^6 integrations steps, and the resulting time scale accessible to molecular dynamics simulations is still short by a factor of $10^3 - 10^6$ for many other important processes.

The total energy function(2) typically employed to describe the interactions between atoms consists of several components:

$$\mathrm{E} = \underbrace{\mathrm{E}_{\mathrm{bond}} + \mathrm{E}_{\mathrm{angle}} + \mathrm{E}_{\mathrm{dihedral}} + \mathrm{E}_{\mathrm{improper}}}_{\mathrm{E}_{\mathrm{bonded}}} + \underbrace{\mathrm{E}_{\mathrm{elec}} + \mathrm{E}_{\mathrm{vdW}} + \mathrm{E}_{\mathrm{hbond}}}_{\mathrm{E}_{\mathrm{nonbonded}}} \quad (2)$$

The components of the total energy are broken down into two categories, the bonded and the nonbonded interactions. The bonded interactions describe interactions between atoms which are linked by covalent bonds and include: E_{bond} to describe high frequency vibrations along covalent chemical bonds, E_{angle} to describe bending motions between two adjacent bonds, $E_{dihedral}$ to describe torsional motion around a bond, and $E_{improper}$ to describe the planar orientation of one atom relative to three other atoms. The nonbonded interactions describe interactions between atoms which are not linked by covalent chemical bonds and include: E_{elec} the pair-wise Coulomb energy between charged atoms, E_{vdW} the pair-wise van der Waals energy, and E_{hbond} the hydrogen bond energy. E_{elec} is computationally the most expensive calculation since the long range nature involves all possible pairs of atoms in the system which results in the scaling of the computational requirements as $(N)^2$.

2.2 Implementation

The restriction to such short time steps and the large number of atoms typically involved in biological systems limits the total length of a simulation accessible to molecular dynamics. In fact molecular dynamics simulations conducted on high performance workstations are currently limited to a few thousand atoms and overall simulation periods of a few nanoseconds. These problems can be overcome by employing parallel computers and more advanced algorithms.

New algorithms have addressed the problem on two fronts. One front attemps to increase the time per integration step through implicit schemes, Langevin formulations, or constraining rapid degrees of freedom [28,30,38,40]. The other front explicitly addresses the evaluation of the long range Coulomb interactions since it is computationally the most expensive calculation. Standard molecular dynamics practice is to incorporate a distance cut-off beyond which no Coulomb interactions are calculated, thus reducing the number of interactions which must be calculated for any given atom. An appropriate switching function must be incorporated so that the forces do not terminate abrubtly at the cut-off boundry. However, this implementation presents errors which become serious for systems containing significantly charged groups [8,23], and as they exist in many biological systems, i.e., the phosphate groups of DNA or the charged head groups of membrane lipids. Hierarchical distance class methods [15,16,36] and multipole expansions [3,14] are alternatives to distance cut-off methods. These approaches preserve the long range nature of the Coulomb interactions while not demanding unacceptable computational resources.

Hierarchical distance class methods are based on a spherical subdivision of interatomic distances into several distance classes. The relative motion of neighboring atoms, which will be grouped in the innermost distance class, will have the most influence on the direction and magnitude of the electrostatic forces and accordingly these forces are calculated at every step. Non-neighboring atoms, grouped in outer distance classes, will have a lesser influence on the direction and magnitude of the

electrostatic forces so that these forces are updated less frequently. The reduced number of calculations for distant interactions results in a speed-up by a factor of six to ten without a significant loss of accuracy [16]. The *Fast Multipole Method* by Greengard and Rokhlin [14] is based on an approximate analytic solution of the Coulomb interactions that has also proven successful for molecular dynamics calculations conducted on high performance workstations [4], and is currently under development for parallel applications [13].

3 EGO

EGO [2, 18, 19] is a molecular dynamics package that is I/O compatible with the widely used molecular dynamics packages X-PLOR [9] and Charmm [6]. EGO was originally programmed in OCCAM II to utilize a system of transputers. Long range Coulomb interactions are accurately represented in EGO by implementing a distance class method [16] and parallelization is achieved by distributing the atoms among the computational nodes. Each node contains the required parameters for its "own" set of atoms, such as type, mass, charge and coordinates, while the atoms which reside on the other nodes are "external". The calculation of the pair-wise forces proceeds in two phases. In the first phase each node calculates in parallel with all the other nodes the pair-wise force between every atom of its "own" set. This phase requires no communication since all information about a node's "own" atoms is stored locally. In the second phase interactions between a node's "own" atoms and all "external" atoms are calculated. This phase requires communication between the nodes, and EGO has implemented a systolic double ring topology [17, 20, 31, 33] for the communication (see figure 1). The first ring consists of the host node and the network nodes, some of which may be exclusively devoted to hydrogen bonding calculations. The second ring contains only the network nodes which have not been set aside for hydrogen bonding calculations. For a transputer based machine in which every processor has four independent communication links the double ring topology is implemented as follows: the first ring is connected by two antiparallel unidirectional OCCAM II channels, while the second ring is connected using the two remaining links of each transputer. On other parallel platforms the hardware topology may not necessarily be configured as a ring, but the communication pattern remains essentially the same.

The first ring is used to send a copy of a node's "own" atoms to the neighboring node in a clockwise direction. When the "external" coordinates are received each node calculates a subset of the pair interactions between its "own" set and the "external" set of atoms. In the next communication step the "external" set of atoms is passed along the first ring in a clockwise direction to the next node. In this manner the individual sets of coordinates, one for each node, are simultaneously passed around the ring until each node receives back its "own" set. When each node receives back its "own" set of coordinates, every set has made a complete revolution around the ring, and therefore all pair-wise interactions have been computed.

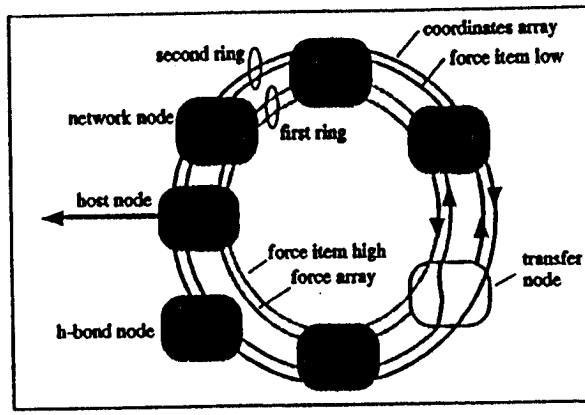

Figure 1:
Double ring topology of EGO. (Picture adapted from [19].)

Newton's third law of motion can be applied to cut the amount of computations in half but forces have to be communicated. Two kinds of force data are communicated, force items and force arrays. The force array is the result of the pair-wise interactions between "external" atoms and a node's "own" atoms. These forces are computed as soon as the "external" coordinates arrive. The force array is communicated entirely on the second ring and lags behind the coordinate array by one transfer step since forces can only be calculated after the coordinates are known. For a transputer implementation separate channels are used for the force array and the coordinate array since the channels can be operated in parallel.

Non-pair interactions involve three atoms(angle interactions) or four atoms(dihedral or improper interactions). The participating atoms may be located on several different nodes. In order to calculate these many-body interactions the necessary coordinates must be selected from the coordinates array as each set of "external" coordinates is passed around the ring and temporarily stored locally. When the complete set of coordinates necessary for a many-body interaction, up to four, has accumulated on the "owner's" node, the force is calculated. These forces are sent out one by one as force items to the nodes which "own" the atoms contributing to the interaction. The force item communication is done on two channels, which form topologically a spiral. One channel, the force item low channel, is on the first ring; the other channel, the force item high channel is on the second ring. New forces are entered on the force item low channel and each node listens on this channel to receive forces that it "owns". In order to avoid a possible deadlock situation in which all nodes are attempting to forward a force item but can not do so because an item is stuck in the channel, the force item high channel is introduced. The only operations on this channel are the forwarding of items and the removal of items belonging to the "owner". One and only

one node passes information from the force item low to the force item high channel, so that forces on the the first ring spiral into the second. Since the only operations on the force item high channel are removal and forwarding of information, it can never deadlock. This also implies that the one node on the force item low channel which passes information to the force item high channel can never deadlock and guarantees that the whole system is deadlock free [10].

The following table compares the average run time per integration step of EGO to X-PLOR for various size systems. The times for EGO were obtained on a Parsytec GCel containing 64 Inmos T805 30Mhz transputers. Two sets of times were obtained for X-PLOR using different high-performance workstations. For small systems and short cut-off distances the workstations out perform the transputer system; however, for larger systems the workstations run out of memory (denoted by err) and cannot match the performance of the transputer. The number of nodes exclusively devoted to the hydrogen bond calculation are also indicated. Performance analysis of EGO has demonstrated that it maintains a high degree of efficiency throughout the simulation [35].

number of atoms	HP-735 64Mb			SGI R4400-150Mhz 80Mb			Parsytec GCel 64 nodes 4Mb per node
cut-off	9Å	19Å	29Å	9Å	19Å	29Å	
1,175	0.48	1.29	1.86	1.63	6.01	8.39	4.98(5 h-bond nodes)
7,475	2.8	16.49	err	12.53	68.26	163.81	9.20(3 h-bond nodes)
13,566	8.52	err	err	28.67	179.62	err	25.13(3 h-bond nodes)

4 APPLICATION

EGO has been used to conduct molecular dynamics simulations of a specific domain of the glucocorticoid receptor interacting with DNA. The glucocorticoid receptor is a member of a family of proteins that regulate transcriptional activity within specific cells. The complete regulatory mechanism is not fully understood although it is known that the receptor must recognize and bind a particular sequence of DNA. For a review of this class of hormone receptors see [32]. The proteins in this family of receptors are called steroid hormone receptors and are responsible for the development and differentiation of a wide range of tissues which include: skin, bone, behavioral centers in the brain, and secondary reproductive tissues. The steroid hormones which activate the receptors include: glucocorticoids, estrogens, progestins, androgens, mineralcorticoids, ecdysteroids, vitamin D, RAR and thyroid hormone [12,22,26].

The X-ray crystallographic structure in [24] was used as the starting point for our molecular dynamics investigation. This structure was modified to remove additional base pairs from the DNA which had been added to enhance crystal formation. This modification restored the oligonucleotide to the naturally occurring sequence that the glucocorticoid receptor binds. Test simulations were then conducted on the system in

vacuum. Results indicated that the protein–DNA system in vacuum was not stable; therefore, water was added to the system. The instability of DNA in vacuum agrees with other simulations of DNA [5, 11, 37], and with the fact that the conformation of DNA is dependent on the environment [34]. The necessity of including explicit waters resulted in a rather large system size of approximately 13,500 atoms. Using the 64 nodes of a Parsytec GCel-64 required approximately one month of run time to complete a 90 ps simulation. The results indicate that the protein induces a bend in the DNA of approximately 35° which agrees with results obtained from gel-mobility shift assay studies of the estrogen receptor complex [29, 25]. The estrogen receptor is also a member of the steroid hormone receptor family and has a high degree of structure and sequence homology with the glucocorticoid receptor. In addition to the bending, several energetically favorable contacts between protein and DNA have been identified which were not observed in the crystallographic study [24].

5 DISCUSSION

Since parallel computers have proven to be an efficient and effective means of addressing the issues of large scale simulations, EGO has been rewritten in C to run under the machine-independent coordinate languages Charm and PVM. Charm is being developed by researchers at the University of Illinois to run on a variety of parallel platforms [21]. Once tested and verified for accuracy Charm-EGO and PVM-EGO will be available by anonymous ftp as a tool for investigating biological structures. This comes at a significant time since crystallographic structures of exciting biological systems are being determined regularly. The particular application mentioned in this article demonstrates the success and significance of molecular dynamics in bridging the gap between crystallographic data, which reveals a static all atom structure and macroscopic data, such as gel-mobility shift assay which reveals global structure without any atomic detail.

REFERENCES

[1] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids.* Oxford University Press, New York, 1987.

[2] Brad Banko and Helmut Heller. User manual for EGO — Release 1.1. Beckman Institute Technical Report TB-92-07, University of Illinois, 1991.

[3] J. E. Barnes and P. Hut. A hierarchical O(N logN) force calculation algorithm. *Nature*, 324:446, 1986.

[4] John A. Board, Jr., J. W. Causey, James F. Leathrum, Jr., Andreas Windemuth, and Klaus Schulten. Accelerated molecular dynamics simulation with the parallel fast multipole algorithm. *Chem. Phys. Lett.*, 198:89–94, 1992.

[5] Klaus Boehncke, Marco Nonella, and Klaus Schulten. Molecular dynamics investigation of the interaction between DNA and dystamycin. *Biochemistry*, 30:5465–5475, April 1991.

[6] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMm: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4(2):187–217, 1983.

[7] Charles L. Brooks III, M. Karplus, and B. M. Pettitt. *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics.* John Wiley & Sons, New York, 1988.

[8] Charles L. Brooks III, B. Montgomery Pettitt, and Martin Karplus. Structural and energetic effects of truncating long range interactions in ionic and polar fluids. *J. Chem. Phys.*, 83(11):5897–5908, December 1 1985.

[9] Axel T. Brünger. *X-PLOR.* The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, May 1988.

[10] William J. Dally and Charles L. Seitz. The torus routing chip. *Distributed Computing*, 1:187–196, 1986.

[11] Mats A. L. Eriksson and Aattoo Laaksonen. A molecular dynamics study of conformational changes and hydration of left-handed d(CGCGCGCGCGCG) in a nonsalt solution. *Biopolymers*, 32:1035–1059, 1992.

[12] R. M. Evans. The steroid and thyroid hormone receptor superfamily. *Science*, 240:889–895, 1988.

[13] L. Greengard and W. D. Gropp. A parallel version of the fast multipole method. Technical report, Rept. YALEU/DCS/RR-640, 1988.

[14] L. Greengard and V. Rohklin. A fast algorithm for particle simulation. *J. Comp. Phys.*, 73:325–348, 1987.

[15] Helmut Grubmüller. Dynamiksimulation sehr großer makromoleküle auf einem parallelrechner. Master's thesis, Physik-Dept. der Tech. Univ., Munich, Germany, 1989.

[16] Helmut Grubmüller, Helmut Heller, Andreas Windemuth, and Klaus Schulten. Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Molecular Simulation*, 6:121–142, 1991.

[17] Helmut Heller. *Simulation einer Lipidmembran auf einem Parallelrechner.* PhD thesis, Technical University of Munich, Germany, December 1993.

[18] Helmut Heller, Helmut Grubmüller, and Klaus Schulten. Molecular dynamics simulation on a parallel computer. *Molecular Simulation*, 5:133–165, 1990.

[19] Helmut Heller and Klaus Schulten. Parallel distributed computing for molecular dynamics: Simulation of large heterogeneous systems on a systolic ring of transputers. *Chemical Design Automation News (CDA News)*, 7(8):11–22, August 1992.

[20] W. D. Hillis and J. Barnes. Programming a highly parallel computer. *Nature*, 326:27, 1987.

[21] L. V. Kale. The chare kernel parallel programming language and system. In *Proc. of the International Conf. on Parallel Processing*, volume 2, pages 17–25, 1990.

[22] Vincent Laudet, Catherine Hänni, Jean Coll, Francois Catzeflis, and Dominique Stéhelin. Evolution of the nuclear receptor gene superfamily. *EMBO J.*, pages 1003–1013, 1992.

[23] Richard J. Loncharich and Bernhard R. Brooks. The effects of truncating long-range forces on protein dynamics. *Proteins: Structure, Function, and Genetics*, 6:32–45, 1989.

[24] B. F. Luisi, W. X. Xu, L. P. Freedman, K. R. Yamamoto, and P. B. Sigler. Crystallographic analysis of the interaction for the glucocorticoid receptor with DNA. *Nature*, 352:497–505, 1991.

[25] G. Redeuith M. Sabbah, S. Le Ricousse and E. Baulieu. Estrogen receptor-induced bending of the xenopus vitellogenin A2 gene hormone response element. *Biochemical and Biophysical Research Communications*, 185(3):944–952, 1992.

[26] Ernest Martinez, Francoise Givel, and Walter Wahli. A common ancestor DNA motif for invertebrate and vertebrate hormone response elements. *EMBO J.*, 10:263–268, 1991.

[27] J. A. McCammon and S. C. Harvey. *Dynamics of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, 1987.

[28] S. Miyamoto and P. A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comp. Chem.*, 13:952, 1992.

[29] Ann M. Nardulli and David J. Shapiro. Binding of the estrogen receptor DNA-binding domain to the estrogen response element induces DNA bending. *Molecular and Cellular Biology*, 12:2037–2042, 1992.

[30] A. M. Nyberg and T. Schlick. On increasing the time step in molecular dynamics. *Chem. Phys. Lett.*, 198:538, 1992.

[31] N. S. Ostlund and R. A. Whiteside. A machine architecture for molecular dynamics: the systolic loop. In B. Venkataraghavan and R. J. Feldman, editors, *Macromolecular Structure and Specificity: Computer-Assisted Modeling and Applications*, pages 195–208. Annals of the NY Acad. of Sciences 439, New York, 1985.

[32] M. G. Parker, editor. *Nuclear Hormone Receptors.* Academic Press, San Diego, CA, 1991.

[33] A. R. C. Raine, David Fincham, and W. Smith. Systolic loop methods for molecular dynamics simualtion using multiple transputers. *Comput. Phys. Commun.*, 55:13–30, 1989.

[34] Wolfram Saenger, editor. *Principles of Nucleic Acid Structure.* Springer–Verlag, New York, NY, 1984.

[35] Amitabh Sinha, Helmut Heller, and Klaus Schulten. Performance analysis of a parallel molecular dynamics program. *Comput. Phys. Commun.*, 1994. In press. [Beckman Institute Technical Report TB-92-13].

[36] W. B. Streett, D. J. Tildesley, and G. Saville. Multiple time-step methods in molecular dynamics. *Mol. Phys.*, 35(3):639–648, 1978.

[37] S. Swaminathan, G. Ravishanker, and D. L. Beveridge. Molecular dynamics of B–DNA including water and counterions: A 140–ps trajectory for d(CGCGAATTCGCG) based on the GROMOS force field. *J. Am. Chem. Soc.*, 113:5027–5040, 1992.

[38] W. F. van Gunsteren. Constrained dynamics of flexible molecules. *Mol. Phys.*, 40(4):1015–1019, 1980.

[39] Loup Verlet. Computer 'experiments' on classical fluids: I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, 159(1):98–103, July 1967.

[40] Guihua Zhang and Tarmar Schlick. LIN: A new algorithm to simulate the dynamics of biomolecules by combining implicit-integration and normal mode techniques. *J. Comp. Chem.*, 1993. In press.