Algorithms for timescale extension and loosely coupled parallelization applied to atomistic simulations of biomolecular systems

Erik Lindahl

Stockholm Bioinformatics Center, Stockholm University, 106 91 Stockholm, Sweden
lindahl@sbc.su.se,
WWW home page: http://www.sbc.su.se

Abstract. Classical molecular dynamics simulations of molecules such as proteins or DNA in water is one of the most common applications for high performance computing due to the extreme computational cost. This paper describes our recent work on extending the accessible temporal and spatial scales, both on commodity workstation hardware as well as massively parallel systems. Examples include new algorithms to remove the fastest degrees of freedom in atomistic simulations, virtual interaction sites, and coarse-grained models. We have further implemented new bandwidth-efficient parallelization techniques that do not sacrifice absolute single-node performance for scalability, shared-memory communication within multiprocessor nodes, low-bandwidth replica-exchange algorithms that make efficient use of slow networks. Finally, distributed computing approaches to protein folding with tens of thousands of clients worldwide are presented.

1 Introduction

Molecular dynamics simulations of biological macromolecules is a very powerful technique to understand properties such as diffusion, molecular interactions, atomic-level structure, and even predict structure or effects of perturbations through *in silico* protein folding. Simulations also hold promise for truly large-scale pharmaceutical applications such as virtual drug screening and accurate prediction of free energies of solvation. Over the last decade with have spend significant efforts both on developing and applying the molecular simulation package GROMACS[1, 2] for this type of problems.

A main bottleneck that needs to be addressed is the immense computational power necessary for useful simulations. Most interesting biological phenomena occur on timescales from microseconds to seconds, while individual simulation timesteps are usually in the femtosecond range. Atomic forces need to be calculated during each of these billions of steps, which is typically dominated by the evaluation of interactions between all pairs of atoms within a given cutoff R. In many cases the remaining long-range interactions are also included through lattice summation techniques such as particle-mesh Ewald (PME)[3].

To be able to simulate larger systems and slower processes we are focusing efforts in three different areas: removing the fastest motions in the system so the timesteps can be made longer, improving parallel scalability in general, and achieving better sampling through multiple simulations with different degrees of loose coupling (or even lack thereof).

2 Removing fast degrees of freedom

The fastest motions in a classical atomic system is the bond vibrations, in particular those involving hydrogens. To be able to integrate these motions accurately requires timesteps of 1 fs. However, the harmonic bond potentials is a pretty bad approximation of the quantum mechanical ground state, so GROMACS usually employs an algorithm called LINCS[4] to constrain all bond lengths and permit the use of at least 2 fs timesteps (doubling performance). In contrast to the standard SHAKE method[5], LINCS is based on a power expansion approximation to matrix inversion instead of iterative corrections. This makes it much more stable for long timesteps, as well as easier to parallelize. After bond vibrations, the next fastest motions are bond angles involving hydrogens. GROMACS has support for automatically removing these and replacing the hydrogens with virtual interaction sites that are generated from heavy atom coordinates. At the beginning of each step all hydrogen coordinates are reconstructed based on ideal geometry. Forces are calculated in the normal way but then projected back onto the constructing atoms by inverting the coordinate equations. Since *only heavy atom positions are updated by the equation of motion*, this enables energy conservation with timesteps as large as 5 fs. Both construction and back-projection is linear in the number of atoms, so the fraction of runtime spent on the virtual sites is below 1%.

The virtual interaction site approach is extremely powerful since nothing limits it to hydrogens. After bond angles, the next faster motions include bending of aromatic amino acid sidechain rings, but in an analogous way we can simply represent the entire ring with a handful of particles to retain overall rotational flexibility while turning the remaining atoms into virtual sites. Currently, we are working on true multiscale approaches where coarse-grained lipid force fields[7] are used for most interactions in the system, and atomic coordinates only constructed as virtual interaction sites where the lipids interact with small peptides.

3 Parallelization, Replica-Exchange and Distributed Computing

GROMACS was originally optimized for cheap low-bandwidth interconnects, but the architecture was recently redesigned from scratch to enable efficient three-dimensional domain decomposition. Our very high single-node performance forced us to focus on reducing the communication bandwidth requirements as much as possible. As recently discussed in the literature[8, 9], this is best accomplished through so-called *Neutral Territory* methods, where pairwise interactions in general are assigned to a node that is not the home of either particle. Even for a small systems with 3000 water molecules we are achieving close to 1 billion pairwise interactions per second over 16 nodes, which corresponds to over 100 nanoseconds of simulation per day with 2 fs timesteps.

However, even with extremely fast interconnects there are limits to how many nodes a given simulation can be effectively be scaled to. Currently, we believe this to occur somewhere in the range 100-200 CPUs for a protein system of 15,000 atoms, and much less for slow networks such as Ethernet. An interesting way to overcome this is to use loosely coupled independent simulations. Replica Exchange is one such approach to improve phase-space sampling, where a system is run at multiple temperatures and conformations periodically swapped[10]. This still requires nodes to communicate, but only when replica exchange decisions are taken every couple of thousand steps. Using this implementation, Seibert recently managed to not only fold short peptides[11], but also correctly predict which conformation is the native one.

Finally, another approach that has proven extremely useful is to simply assume first order exponential reaction kinetics and just run multiple *uncoupled* simulations to gather statistics. For a process with a reaction time in the microsecond range the probability of observation within 10 ns is extremely small, but not zero. If tens of thousands such simulations are performed the expectation value will however increase dramatically! Since no communication is required such computational power can be obtained from distributed computing, where simulations run as screensavers on home computers[12]. Interestingly, this approximation even works for systems as large as proteins, which made it possible for us to produce dozens of folding trajectories for the BBA5 25-residue protein in explicit water in only a couple of weeks[13].

References

1. Lindahl, E., Hess, B., van der Spoel, D.: GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol. Model **7** (2001) 306–317
2. van der Spoel, D, Lindahl, E., Hess, B., Groenhof, G., Mark A.E., Berendsen, H.J.C.: GROMACS: fast, flexible, and free. J Comp. Chem. **26** (2005) 1701–1718
3. Essman, U., Perera, L. Berkowitz, M.L., Darden, T., Lee, H., Pedersen, L.G.: A smooth particle mesh Ewald method. J. Chem. Phys. **103** (1995) 8577–8592.
4. Hess, B., Bekker, H., Berendsen, H.J.C., Fraaije, J.G.E.M.: LINCS: A Linear Constraint Solver for Molecular Simulations. J. Comp. Chem. **18** (1997) 1463–1472.
5. Ryckaert, J.P., Ciccotti, G., Berendsen, H.J.C.: Numerical Integration of the Cartesian Equations of Motion of a System with Constraints; Molecular Dynamics of n-Alkanes. J. Comp. Phys. **23** (1977) 327–341.
6. Feenstra, K.A., Hess, B., Berendsen, H.J.C.: Improving Efficiency of Large Timescale Molecular Dynamics Simulations of Hydrogen-rich Systems. J Comp. Chem. **20** (1999) 786–798
7. Marrink, S.J., de Vries, A.H., Mark, A.E.: Coarse Grained Model for Semiquantitative Lipid Simulations. J. Phys. Chem. B **108** (2004) 750–760

8. Snir, M.: A note on n-body computations with cutoffs. Theory Comp. Systems **37** (2004) 399–416
9. Shaw, D.E.: A fast, scalable method for the parallel evaluation of distance-limited pairwise particle interactions. J Comp. Chem. **26** (2005) 1318–1328
10. Hukushima, K., Nemoto, K.: Exchange Monte Carlo Method and application to spin glass simulations. J. Phys. Soc. Jpn. **65** (1996) 1604–1608
11. Seibert, M.M., Patriksson, A., Hess, B., van der Spoel, D.: Reproducible Polypeptide Folding and Structure Prediction using Molecular Dynamics Simulations. J. Mol. Biol. **354** (2005) 173–183
12. Shirts, M., Pande, V.S.: Screen Savers of the World Unite! Science **290** (2000) 1903–1904
13. Rhee, Y.-M., Sorin, E.J., Jayachandra, G., Lindahl, E., Pande, V.S.: Simulations of the role of water in the protein-folding mechanism. Proc. Natl. Acad. Sci. **101** (2004) 6456–6461