Considerations and Recent Advances in Molecular Dynamics

Simon Ye

Introduction

Molecular dynamics simulations were first proposed in the 1950-60s as a method to study the motions of atoms at the molecular level. Today, the most prominent use of molecular dynamics is the study of biological molecules and materials science. Molecular dynamics is the study of the motions of atoms and molecules by numerically simulating Newtonian dynamics. Since computers are discrete, they must simulate the continuous motions of atoms by dividing the trajectory into states where the velocities and positions of each atom are recorded. Forces and displacements are calculated for each time step, and the new state of the atoms computed, a process that continually repeats to simulate continuous motion. To guarantee that the discrete approximation does not deviate far from reality, the time step must be very small, making simulations very computationally complex.

The ability to study molecules in atomistic detail provides a number of advantages over experimental techniques. Modern day imaging techniques can only examine atoms with extreme difficulty, and even this has been a breakthrough of the last few years. The next difficulty is in studies of picosecond scale dynamics, which also became possible recently through ultrafast lasers such as 2D IR vibrational echo spectroscopy.[12] Despite not being able to account for excited electronic states, quantum effects, or the formation and breaking of chemical bonds, molecular dynamics can grant critical insight into many problems of biological interest, including protein-ligand binding, protein folding, RNA folding, and much more.[17]

Since molecular dynamics is a computationally difficult problem, simulations in the past were very short and were of limited size - less than 10 ps in total length and fewer than 1000 atoms. Today, it is possible to simulate systems of up to 10,000-100,000 atoms for time periods of up to 1 ms. This system size is enough to accommodate single proteins interacting with an environment of explicit waters or other ligands, and the time period is finally reaching the lengths necessary to exhibit protein folding/unfolding cycles or unguided ligand binding.

From a practical perspective, there are also now a number of highly optimized open source programs for MD with parallelizable code and/or GPU acceleration, including AMBER, GROMACS, CHARMM, Desmond, and NAMD, among many other packages. Although there are large scale projects either harnessing massive numbers of computers like Folding@Home, or utilizing extremely specialized hardware such as Anton, MD is also becoming a useful tool accessible to everyday chemists and biologists to guide their research. As these programs grow in features, they also grow in complexity, with numerous tweakable parameters for setting up simulations. Improper setup of a simulation may turn even the best programs into random number generators creating trajectories that have no basis in reality. Therefore, this review will cover some considerations in running molecular dynamics simulations, and provide a recent application that shows future promise.

Force Field

The force field is perhaps the most important component of a molecular dynamics program. In general terms, the force field computes the energy of system based on its conformation. A number of force fields are commonly used for biomolecular simulations, including AMBER,[2,5] OPLS,[11,20] and CHARMM.[7]

All of these force fields are considered additive all-atom force fields. Additive meaning that they can be separated into additive terms for bonded and nonbonded energies, and all-atom meaning that all atoms are modeled in the force field. Historically most of these force fields were united-atom, which meant that effects of hydrogen atoms were folded with the atoms they were bonded to. Nowadays these methods are dropped in favor of the more realistic all-atom approaches due to advances in computing power.[10] These "classical" force fields generally divide the total energy into four terms. The first is bond lengths, found via the sum of individual bonds treated as a harmonic potential. The second is bond angles, again found via a harmonic potential. These two terms are very strong, and are often constrained to save running time. The third term is torsions, found via a Fourier series for all unique sets of dihedral atoms. The last term is for pairwise non-bonded interactions, which is divided into a 6-12 Lennard Jones (LJ) potential and the Coulombic potential. The LJ potential represents dipole-dipole interactions and dispersion forces.

$$V(r^N) = V_{bonds} + V_{angles} + V_{torsions} + V_{nonbonded}$$

$$V_{bonds} = \sum_{bonds} \frac{1}{2} k_b (l - l_0)^2$$

$$V_{angles} = \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2$$

$$V_{torsions} = \sum_{torsions} \frac{1}{2} V_n [1 + cos(n\omega - \gamma)]$$

$$V_{nonbonded} = \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \{\epsilon_{i,j}[(\frac{r_{0ij}}{r_{ij}})^{12} - 2(\frac{r_{0ij}}{r_{ij}})^{6}] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\}$$

Each of these terms are parametrized based on a mixture of empirical data and quantum-level calculations. Bond and angle harmonic potentials are usually fitted to experimental vibrational data. Partial charges on atoms are derived from quantum calculations at the 6-31G* level of theory and restrained electrostatic potential (RESP) fitting. LJ parameters are derived somewhat differently between various force fields. OPLS is designed and fitted against condensed phase properties of organic liquids. In OPLS, these parameters are derived from Monte Carlo simulations of a large number of organic liquids to reproduce thermodynamic and bulk properties. AMBER similarly fits against various experimental parameters but is also based on some arbitrary determinations by the authors. There is also disagreement about how to scale down the LJ terms relative to Coulombic terms and differing atom types. Although the methods to derive LJ terms are not necessarily well-defined, they do match fairly well with experimental values. With all the other parameters fixed, torsional parameters can be fitted to quantum calculations at the RHF/6-31G* level for a variety of molecules and ions by comparing against MD simulations without torsional terms.

These force fields are similar on theoretical grounding but are fitted to very different sets of experimental values. When used to study model systems, such as short polypeptides, each force field occupies widely varying areas of the Ramachandran plot of ϕ / ψbackbone angles.[13,14,27] Therefore, each force field has undergone evolution as different research groups try to improve older parametrizations due to the increase in available experimental data and faster computers.[8] In recent years, AMBER and CHARMM have had reparametrizations improving dihedral potentials, resulting in improved force fields such as AMBER99SB-ILDN and CHARMM27/CMAP.[15] Regardless, force field performance is still system-dependent, so many variants are used in simulations today.

Future development in force fields will involve improving classical force fields, as well as developing new approaches such as polarizable and QM/MM approaches. Polarizable force fields allow atoms to hold polarizable charges instead of the static fixed charges of classical dynamics, and QM/MM hybrid approaches incorporate some quantum level calculations to parts of the system. These approaches have better accuracy, but at the expense of a much greater computational cost.

Bond Constraints

Bond constraints fix the bonds lengths and angles in the system, effectively removing rapid harmonic vibrations. The flip side is the ability to select a much larger time step for simulation. Typically without any bond constraints, simulations will be run around 1 fs per time step, but with bond constraints applied to all bonds, this can increase to 2-2.5 fs. Applying constraints to a limited subset such as hydrogen bonds will lead to an intermediate compromise.

A number of constraint algorithms have been developed over time, SHAKE, SETTLE, M-SHAKE and LINCS. All these methods work through the method of Lagrange multipliers applied to distance between atoms. This results in a nonlinear equation that is typically solved using Newton's method. The most time consuming step of these algorithms is the determination of the Jacobian J via Newton's method. Each algorithm implements this slightly differently.

SHAKE iteratively solves this via the Gauss-Seidel method, an iterative process that parallelizes poorly.[18] This drawback has led to development of a number of new methods. SETTLE solves the system analytically for groups of 3 constraints. This is very fast and works for explicit water models containing three bonds, but not for larger molecules.[23] M-SHAKE solves the Jacobian directly via LU decomposition, which is faster than SHAKE but grows with complexity $O(n^3)$ in the number of constraints due to the need for matrix inversion.[1] This makes it unsuitable for large molecules. LINCS estimates J^{-1} with a power series, which is faster and more parallelizable than SHAKE, but only works for sparse bond connectivity (sparse matrix), and works with constraining bond lengths only.[16] A newer method known as CCMA calculates J once at the beginning of the simulation and constructs approximations K to J that are fairly similar but easier to invert, leading to faster calculations. This algorithm does worse for molecules with greater flexibility.[9]

The development of newer algorithms for bond constraints has led to faster algorithms each with varying drawbacks. All the above methods solve for relative constraint tolerances to a specific point (usually 10^{-4}), leading to the same empirical result, but they will have different performance profiles depending on the system being studied and computer architecture used. Further research into efficient bond constraint algorithms is a promising field for the future, while users must consider different algorithms for maximizing simulation performance.

Solvent Model

Two models of solvent are used in practice - explicit and implicit solvent. For the most common case of water as the solvent, there are TIP3P, SPC, SPC/E, TIP4P, TIP5P as commonly used explicit water models. These model water as a rigid body with parameters for bond lengths, angles, dipole moments and other properties. For models with more than 3 sites like TIP4P and TIP5P, virtual dummy atoms are added to for better fitting with electrostatic properties. However, all of these explicit rigid-body water models suffer from not being able to accurately reproduce experimental bulk properties of water.

The most commonly used implicit water model is GBSA (Generalized Born / Solvent Accessible), which starts with the Generalized Born approximation to the Poisson-Boltzman equation describing the electrostatic potential of a solute in an ionic solvent. It then adds nonpolar contributions with are proportional to the solvent accessible surface area in the solute. The main advantage of using implicit solvent is the massive speed-up of not having to treat interactions of solute molecules.

Some studies have shown consistent properties between proteins in explicit and implicit water[26], while others have shown inconsistent results.[24,25] It is clear that the free energy landscape of implicit solvent is different from explicit solvent, but it is not clear that it is necessarily worse. In many cases, implicit solvent has been shown to reproduce experimental results with good accuracy. Considering the inherent problems of explicit water models, implicit solvent is commonly used for molecular dynamics, even though our intuition suggests that explicit water more closely models reality. This is an area of molecular dynamics that is very poorly understood where none of the models seem satisfactory. More research will be needed to better understand the role of solvation in molecular dynamics.

Ensembles

A number of different statistical ensembles are typically used for molecular dynamics. The microcanonical ensemble, or NVE, maintains constant the number of particles (N), volume (V), and energy (E) of the system. This ensemble maintains the correct properties of the canonical ensemble lending to easy calculation of statistical averages. Although this ensemble in principle should maintain a stable temperature, having low numbers of atoms in typical MD simulations, edge effects from periodic images, and computer approximation can cause drifts in temperature. Therefore it is the least widely used because usually there is a desire to maintain the system at a consistent temperature. An NVT ensemble controls the number of particles, volume, and temperature. This ensemble is usually used in production runs in place of NVE due to the ability to model the canonical ensemble while controlling temperature, but also requires a thermostat. An NPT ensemble controls the number of particles, volume, and temperature. This requires coupling to both a thermostat and a barostat.

Typical preprocessing steps in molecular dynamics starts with the coordinates of a protein with a large enough simulation box to accommodate explicit solvent and avoid interactions with periodic images. After energy minimization, an NVT phase is used to heat up the system containing protein and solvent. Using NPT directly may destabilize the system as low temperatures lead to inaccurate pressure estimations. This slowly increases the temperature to the desired level. An NPT phase proceeds after temperature has equilibrated, resizing box vectors to the correct size. For protein studies, this phase may also place position restraints on the protein to let the water “soak” around the protein without altering overall protein structure. Finally, the production run proceeds using an NVT ensemble to remove excess motions induced by pressure coupling while saving some calculations.

Thermostats and Barostats

A number of different thermostats are available: velocity rescaling, Berendsen, Nose-Hoover, and Langevin thermostats among others are commonly used.[4] Velocity rescaling simply slowly rescales atom velocities to the correct kinetic energy, which is simple but doesn’t allow natural fluctuations. The Berendsen thermostat couples the system to an external heat bath with a fixed temperature. Velocities are then rescaled according to a parameter τ that determines how tightly the system is coupled to the bath. The Berendsen thermostat is suitable for heating/cooling a system, but does not generate the correct canonical ensemble. Nose-Hoover adds an extra term to the Hamiltonian of the system representing an artificial mass and velocity for the heat bath. This method generates temperature oscillations consistent with the canonical ensemble, but care must be taken to ensure that the frequency of fluctuations is consistent with reality. Too tight coupling may cause unrealistically quick fluctuations, while too weak coupling may take too long to reach desired equilibration. Another problem is the translation of kinetic energy from one part of the system to another, as exhibited when the solvent heats up concomitantly with a cooling of the solute. To prevent this, protein and solvent are usually coupled to separate thermostats. The Langevin thermostat adds a frictional term and stochastic noise term to the computation of forces. This samples the canonical distribution but may cause drifts in energy over time.

Different barostats are also used in practice. The Berendsen barostat acts similarly to the thermostat of similar name but used to constrain pressure, with similar drawbacks of not generating the correct ensemble. The Nose-Hoover and Andersen barostats work similarly to the Nose-Hoover thermostat by addition an additional term to the Hamiltonian that can be thought of as representing a compressing piston. While the previous methods only allow a change in the overall size of the simulation box, an extension of these methods by Parrinello-Rahman allows the box vectors to change directions as well, a property that is not particularly useful in typical biomolecular simulations but useful in simulating crystal properties of metals in materials simulations. These methods coupling to a fictitious mass have similar to the thermostat drawbacks, making it possible to generate unrealistic oscillations of box vectors.

In practice, there is no widely agreed-upon theory of selecting and parametrizing the correct thermostat and barostat that guarantees a correct and reliable simulation. General rules of thumb, such as Berendsen thermostat/barostat for relaxation, and Nose-Hoover thermostat/ Parinello-Rahman barostat for production simulations, have been established as a reasonable best practice.

Enhanced Sampling

The energy landscape of molecular dynamics is extremely high dimensional. Dimensionality scales roughly $6N$ in the number of atoms. This means that the finding the global energy minimum can be extremely difficult in a landscape with innumerable local maxima with large energy barriers. To help with this problem, a number of methods for enhanced sampling have been developed to better explore the energy landscape at the expense of fully physical simulation.[21]

Replica-exchange molecular dynamics (REMD) is a common technique which runs a large number of replicas of a system at a range of temperatures.[22] Every few time steps, usually around 10 fs, the conformations of the replicas are randomly switched to improve the chance of overcoming energy barriers that would be difficult to surmout at lower temperatures. Metadynamics adds a history-dependent potential that disfavors visited past states, and can be thought of as filling local energy minima with gaussians to prevent wasteful revisiting.[28]

Umbrella sampling samples a number or replicas against specified reaction coordinate to determine the corresponding energy surface. One example of such a reaction coordinate may be one-dimensional center-of-mass pulling on the protein to approximate protein folding/ unfolding or the binding of ligands. Alchemical methods such as weighted histogram analysis method (WHAM) and thermodynamic integration applied to free energy perturbation may be used to calculate absolute and relative binding free energies for a protein-ligand complex without the need for a lengthy molecular dynamics trajectory. This is done by generating a series of intermediate nonphysical "alchemical" states and summing or integrating across these states to generate an optimal estimate of free energy differences.[6]

Application

A recent study used molecular dynamics to investigate drug binding to G-protein coupled receptors (GPCRs).[3,19] One third of all drugs target GPCRs, yet the pathway of binding from a completely disassociated state has remained relatively unknown. The receptor studied was the β_2-adrenergic receptor (β_2AR), which is targeted by beta blocker and beta agonist drugs treating a variety of conditions including hypertension, myocardial infarction, bradycardia, and angina pectoris. This study used unbiased all-atom MD simulations with three antagonists (propranolol, alprenolol, and dihydroalprenolol), and the agonist isoproterenol. Simulation conditions placed ligands at least 30 Å from the binding site and 12 Å from the receptor surface. The receptor was placed in an explicit lipid and water environment containing ~10,000 water atoms and ~60,000 total atoms. Simulations were performed with the CHARMM27 force field with all bond lengths constrained via M-SHAKE. After equilibration, production runs lasting 1-19 μs were run in the NPT ensemble.

The surprising result was that the largest energetic barrier to ligand binding was not anywhere close to the actual binding site. The overall mechanism of binding proceeds through one dominant pathway, in which the ligand (alprenolol in this case) first associates with a surface region termed the "extracellular vestibule", where it remains for hundreds of nanoseconds. Afterwards, alprenolol squeezes through a narrow passage into the binding pocket, where it immediately adopts the crystallographic pose. The unbinding process then follows the reverse of the binding process. The largest energetic barrier is from the binding of the alprenolol to the extracellular vestibule, not the squeezing of the drug into the active site. This provides new insight to the mechanism of drug binding, and opens up the possibility of allosteric regulation of β_2AR via new drug classes.

Conclusion

With recent advances in computing power and molecular dynamics algorithms, it is becoming feasible to simulate biological systems at time scales of interest. Many aspects of molecular dynamics are poorly understood and require intensive further research, but the field of biological simulation has matured over the last few decades with promising results. In the future, these simulations hold great promise towards biological understanding as they take center stage in the scientist's toolbox.

References

1. Kräutler, V., van Gunsteren, W.F. & Hünenberger, P.H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry* **22**, 501-508 (2001).
2. Cornell, W.D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179-5197 (1995).
3. Dror, R.O. *et al.* Activation mechanism of the β2-adrenergic receptor. *Proceedings of the National Academy of Sciences* (2011).doi:10.1073/pnas.1110499108
4. Hünenberger, P.H. Advanced Computer Simulation. **173**, 130 (2005).
5. Weiner, S.J., Kollman, P.A., Nguyen, D.T. & Case, D.A. An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry* **7**, 230-252 (1986).
6. Shirts, M., Mobley, D. & Chodera, J. Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? *Annual Reports in Computational Chemistry* **3**, 41-59 (2007).
7. Brooks, B.R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**, 187-217 (1983).
8. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**, 712-725 (2006).
9. Eastman, P. & Pande, V.S. Constant Constraint Matrix Approximation: A Robust, Parallelizable Constraint Method for Molecular Simulations. *Journal of Chemical Theory and Computation* **6**, 434-437 (2010).
10. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. & Case, D.A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **25**, 1157-1174 (2004).

11.
Jorgensen, W.L., Maxwell, D.S. & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **118**, 11225-11236 (1996).

12.
Chung, J.K., Thielges, M.C. & Fayer, M.D. Dynamics of the folded and unfolded villin headpiece (HP35) measured with ultrafast 2D IR vibrational echo spectroscopy. *Proceedings of the National Academy of Sciences* **108**, 3578-3583 (2011).

13.
Wu, X. & Wang, S. Helix Folding of an Alanine-Based Peptide in Explicit Water. *J. Phys. Chem. B* **105**, 2227-2235 (2001).

14.
Piana, S., Lindorff-Larsen, K. & Shaw, D.E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal* **100**, L47-L49 (2011).

15.
Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **78**, 1950-1958 (2010).

16.
Hess, B., Bekker, H., Berendsen, H.J.C. & Fraaije, J.G.E.M. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463-1472 (1997).

17.
Karplus, M. & McCammon, J.A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **9**, 646-652 (2002).

18.
Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J.. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327-341 (1977).

19.
Dror, R.O. *et al.* Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences* **108**, 13118 -13123 (2011).

20.
Jorgensen, W.L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 6665 -6670 (2005).

21.
Yang, W., Nymeyer, H., Zhou, H.-X., Berg, B. & Brüschweiler, R. Quantitative computer simulations of biomolecules: A snapshot. *Journal of Computational Chemistry* **29**, 668-672 (2008).

22.
Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**, 141-151 (1999).

23.
Miyamoto, S. & Kollman, P.A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **13**, 952-962 (1992).

24.
Rhee, Y.M. Simulations of the role of water in the protein-folding mechanism. *Proceedings of the National Academy of Sciences* **101**, 6456-6461 (2004).

25.
Chopra, G., Summa, C.M. & Levitt, M. Solvent dramatically affects protein structure refinement. *Proceedings of the National Academy of Sciences* **105**, 20239 -20244 (2008).

26.

Zhang, L.Y., Gallicchio, E., Friesner, R.A. & Levy, R.M. Solvent models for protein-ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. *Journal of Computational Chemistry* **22**, 591-607 (2001).

27.
Graf, J., Nguyen, P.H., Stock, G. & Schwalbe, H. Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. *Journal of the American Chemical Society* **129**, 1179-1189 (2007).

28.
Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100**, 020603 (2008).