Mem. S.A.It. Suppl. Vol. 4, 75
© SAIt 2004

Memorie della
Supplementi

Molecular Dynamics simulation of biosystems: Perspectives and open problems

S. Melchionna

Dipartimento di Fisica and INFM, Università di Roma "La Sapienza", P.le A. Moro 5, 00185, Rome, Italy. e-mail: simone.melchionna@roma1.infn.it

Abstract. In this brief review I will expose some basic principles underlying the study of biological molecules via Molecular Dinamics and explain why this research domain poses new challenges for scientists interested in extended applications and development of novel methodologies.

Key words. Biological systems – Molecular Dynamics

The simulation of biological matter represents one of the grand challenges in computational science. Biomatter is composed of large macromolecular aggregates made of proteins, DNA, sugars, lipids, and so forth, typically embedded in a solvent such as water. These biological entities are either "soft" objects, such as membranes, or "hard" objects, such as proteins. These definitions carry some arbitrary meaning. For example, membranes are self-aggregating substances made of several lipids, without a specific conformation of the constituting molecules but with a global structural arrangement of the assembly. Proteins are long heteropolymers with a peculiar geometrical arrangement, the so-called folded state, such that the scaffolding confers an intrinsic rigidity to the whole structure. Often one wishes to simulate a system made of composite biological entities. Overall, the functional role of biomatter is deeply connected to the thermodynamic state, the structure and dynamics, of the heterogeneous environment.

Molecular Dynamics (MD) is a powerful simulation device to access the microscopic level of complex systems. For excellent books on Molecular Dynamics the reader should consult refs. [Allen & Tildesley (1987); Frenkel & Smit (1996)]. The button-up approach to simulation begins from the atomistic description once an accurate representation of interatomic forces is available. In many circumstances quantum effects can be safely neglected in favor of a classical and a statistical mechanics description, for systems in the condensed or in the gas phase. To this aim, a plethora of classical parametrizations, or force fields, is available to reproduce biological matter at atomic level. These force fields are usually fitted from experimental or ab-initio data of atoms and molecules in vacuo or in solvated conditions.

As we will soon see, MD is a powerful technique for complex systems since the atomic motion is followed along a quasi-continuous trajectory in phase space. The basic requirement for MD is that the interatomic potentials change smoothly in between subsequent time steps. Once this condition is met, atoms will move by following the interatomic forces, as prescribed by the equations of motion. This simple mechanism renders MD a more powerful technique than stochastic or probabilistic techniques, such as Langevin dynamics of Monte Carlo methods, in particular for systems at high density as encountered in condensed matter.

As compared to simpler systems, such as for liquids made of small molecules, the simulation of biosystems needs to address several specific problems [van Gunsteren et al. (1993)]. In fact, biological molecules present extended and articulated topological connectivity and differentiated interactions among atoms, ranging from excluded volume interactions, to electrostatic forces, and intramolecular forces associated to chemical bonding. As a result of the diversity in interactions, biomatter exhibits a wide spreading of dynamical modes with timescales ranging from femtoseconds to seconds, and metastabilities that need to be overcome in specific cases. Moreover, a proper description of biological matter needs to take into account wide portions of solvent, which represent a heavy computational cost. Despite these limitations, MD is widely employed and the simulators community is deeply involved in devising novel and efficient algorithms to exploit the technique to larger and more complex systems.

The central goal of MD is to generate a quasi-continuous trajectory in phase space starting from the classical equations of motion embodied by Newton's law ($m_i \ddot{r}_i = F_i$) or, more conveniently, Hamilton dynamics

$$\begin{aligned} \dot{r}_i &= \frac{p_i}{m_i} = \frac{\partial H}{\partial p_i} \\ \dot{p}_i &= F_i = -\frac{\partial H}{\partial r_i} \end{aligned} \qquad (1)$$

The equations of motion can be written in compact form once the Liouville operator is defined as

$$iL = \dot{r}\frac{\partial}{\partial r} + \dot{p}\frac{\partial}{\partial p} = \frac{p}{m}\frac{\partial}{\partial r} + F\frac{\partial}{\partial p} = [H, \ ] \qquad (2)$$

so that the equations of motion

$$\dot{\Gamma}(t) = iL_t\Gamma(t) \qquad (3)$$

have formal solution

$$\Gamma(t) = e^{itL_t}\Gamma(t_o)$$

where $\Gamma = \{r_i, p_i\}$ is the $6N$-dimensional vector spanning phase space. MD generates a discretized trajectory sampling the statistical distribution associated to the ensemble compatible with the boundary conditions (energy, temperature, volume, pressure, and so forth). From this brute force approach, one can extract the ensemble average of any phase space function $A(\Gamma)$. If $f(\Gamma)$ is the normalized distribution function, the average $< A >= \int d\Gamma A(\Gamma) f(\Gamma)$ is calculated in practical terms as

$$< A >\simeq \frac{1}{T}\sum_{n=1}^{T} A_{t_n} \qquad (4)$$

where A_{t_n} is the numerical value of the observable at time t_n. As the observation time T is long enough, the time average equals the ensemble average.

The goal of MD is to produce an accurate trajectory, so that the ensemble averages of any structural or dynamical quantity is approximated at best. However, one should always keep in mind that trajectories are intrinsically unstable. If Γ_{MD} indicates the sequence of points generated by the discretized MD propagator, it will diverge from the true trajectory according to the so-called Lyapunov instability,

$$|\Gamma_{MD}(t_n) - \Gamma(t_n)| \propto |\Gamma_{MD}(t_o) - \Gamma(t_o)| e^{(n-1)h/\tau} \qquad (5)$$

where τ is a characteristic time. Nevertheless, to sample accurately the long-time dynamics and the proper statistical distribution it is sufficient to require that the short-time dynamics is accurate enough. More specifically, this is achieved by satisfying the stability criterion that energy is nearly conserved at long times

$$\frac{|E(t_n) - E(t_o)|}{|E(t_o)|} < 10^{-5} \quad \text{for} \quad t_n - t_o \to \infty (6)$$

A more detailed treatment shows that a sufficient condition for the stability criterion is obtained by using a propagator that is time-reversible, as prescribed by Hamilton dynamics, and symplectic, having the effect of preserving measure, i.e. an arbitrary volume of points in phase space, over time [Tuckerman et al. (1990)].

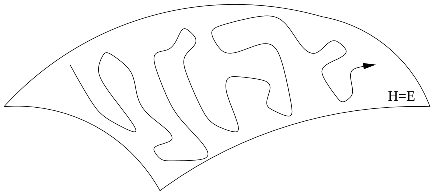

Fig. 1. A schematic trajectory on the constant energy surface in phase space.

The propagation of the MD trajectory is obtained by producing a sequence of points discretized over the time step h, so that the elementary propagated point is

$$\Gamma(h) = e^{ihL}\Gamma(0)$$

One of the most widely used propagators is provided by the "velocity Verlet" algorithm. The algorithm can be formally derived by exploiting a Trotter splitting of the propagator

$$\begin{aligned}\Gamma(h) &= e^{ih[\frac{p}{m}\partial/\partial r + F\partial/\partial p]}\Gamma(0)\\ &\simeq e^{i\frac{h}{2}F\partial/\partial p}e^{ih\frac{p}{m}\partial/\partial r}e^{i\frac{h}{2}F\partial/\partial p}\Gamma(0) + Err(h^3)\end{aligned} \tag{7}$$

Each of the three distinct exponential operators gives rise to three separate translations of positions and momenta, the algorithm resulting in the following explicit form

$$\begin{aligned}r(h) &= r(0) + \frac{h}{m}p(0) + \frac{h^2}{2m}F(0)\\ p(h) &= p(0) + \frac{h}{2}[F(0) + F(h)]\end{aligned} \tag{8}$$

Hamilton dynamics conserves energy and, since volume and the number of particles are kept constant over time, therefore the discretized trajectory samples the microcanonical ensemble in phase space (Fig. 1).

In many circumstances, however, the numerical simulation needs to obey to different "external" conditions, as for working at fixed temperature and/or pressure. Hamilton dynamics is not sufficient to simulate a system in the canonical or in the isothermal-isobaric ensembles, therefore alternative formulations must be employed. By restricting ourselves to the canonical case, there exist several choices for generating an isothermal dynamics. However, an isothermal dynamics does not necessarily samples the canonical distribution in both positions and momenta, i.e. $f(r, p) \propto e^{-\beta H(r,p)}$. Moreover, one would like to employ continuous equations of motion that exhibit a set of conserved quantity. This requirement has the practical advantage of helping the simulator in checking the quality of the dynamics and to fine-tuning the time step. The existence of a conserved quantity implicates that a symplectic, measure-conserving propagator would possibly produce long-time stability. All the above requirements are met by a class of non-Hamiltonian dynamics, the so-called Nosé-Hoover equations of motion, finding wide application in MD. The Nosé-Hoover class of dynamics can achieve canonical, isothermal-isobaric, isoenthalpic, or isostress sampling of the statistical distribution [Melchionna et al. (1993)].

The minimal size of the time step reflects the stiffness of the interatomic potentials acting within the system. Typically, the time step h is chosen to be at least two orders of magnitude smaller than the period associated to the fastest vibrations occurring in the system, $h \simeq \tau_{at}/100$, so that h depends on the stiffness k_{MAX} of the interatomic potentials, i.e. $h^{-1} \propto \sqrt{k_{MAX}}$. This condition implies that the hardest potentials acting in the system, usually associated to chemical bonding, impose a time step equal to one femtosecond or smaller. A powerful improvement is obtained by substituting interatomic bonding potentials with holonomic constraints, since unimportant degrees of freedom can be safely removed [Ciccotti & Ryckaert (1986)]. In such favorable cases, the time step can be raised by one order of magnitude. On the other hand, statistical mechanics exhibits that the distribution function associated to the constrained system differs from the unconstrained one by the presence of a configurational term [Melchionna (2000A)].

In order to build up a MD simulation one still needs to face other important details, i.e. setting up the boundary conditions for the system and computing the interatomic forces. For most of purposes, the biological system is in the condensed phase so that one needs to en-

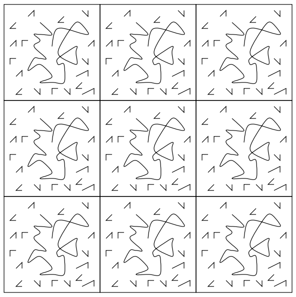

Fig. 2. Topological view of the Periodic Boundary Conditions.

sure that the simulation results are weakly dependent on the boundaries. If one is interested in studying a small system, as it is typically the case in MD, the use of a confining mechanism, such as a hard wall or a restraining potential, would induce spurious correlations in the system. The optimal choice is to employ periodic boundary conditions (PBC), a trick such that surface-induced effects are minimized in favor of the bulk behaviour of systems. The PBC topology can be schematized as made of a central MD cell containing the physical system surrounded by infinite replicas, or images, of the system in the x, y and z directions (Fig. 2). The shape of the central and image cells is usually parallelepiped and the overall system entirely tiles the cartesian space.

The computation of interatomic forces represents the heaviest cost in MD. In fact, the interaction among N atoms is usually expressed in terms of pairwise interactions and consequently the computation requires N^2 operations plus the interactions between the particles located in distinct PBC cells. On the other hand, there exist several smart algorithms that reduce the computational load to order N, typically costing 50-70% of the total simulation effort.

The optimal way to handle interactions is distinguished on the basis of the short-range or long-range nature of the forces. Short-range interactions are typically Van der Waals forces, representing excluded volume or chemical forces. The short-range nature of the potential allows to compute forces only for pairs of atoms with relative distance smaller than a cut-off. Firstly, in order to restrict the calculation to pairs where the first atom belongs to the central cell and the second atom belongs either to the central or to a single image cell, the cut-off must be smaller than half the size of the simulation cell. This condition is called the "minimum image condition" and can be met by systems with relatively small spatial extension (e.g. 20 Å). Secondly, the list of interacting pairs must be constructed prior to computing the interactions. The cost of scanning pairs is $\propto N^2$ in principle but by using a grid list, such as conveniently provided by the link cell method, the cost becomes $\propto N n_{neigh}$, where n_{neigh} is the average number of neighbors per atom (Fig. 3).

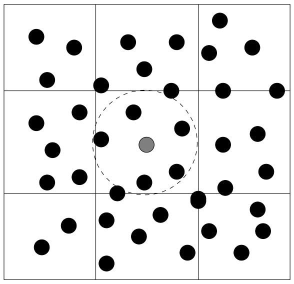

Fig. 3. The Link Cell Method. Scanning neighbors is done by dividing the MD cell in smaller cells, each with edge equal to the cut-off, and by filtering out neighbors belonging to second neighbor cells. The minimum image convention prescribes that the cutoff be smaller than half the MD cell size.

Long-range interactions are usually embodied by the electrostatics forces. The long-range form of the potential cannot leave aside from computing the interaction of one atom with all replicas of another atom, so that the minimum image condition cannot be employed. The calculation of electrostatics compatible with PBC is obtained by adopting the Ewald method [Hansen (1986)], a way to compute the Fourier-transformed long-range component of the Coulomb forces. The calculation of electrostatics via Ewald method represents a central task in biosimulations, since biological matter exhibit distributed charges. Unfortunately, the calculation of the Ewald interactions is an heavy part of MD, its cost growing as $N^{3/2}$. Several alternatives employing Fast Fourier Transforms or Real Space algorithms allow to considerably reduce the computational load, such as the particle-particle/particle mesh (PPPM) method of Eastwood and Hockney [Eastwood & Hockney (1974)], which scales as $N \log N$, or the fast multipole method of Greengard and Rokhlin [Greengard & Rokhlin (1987)], which scales as N. A popular alternative is provided by the Smooth Particle Mesh Ewald of Darden [Essmann et al. (1995)], a way to compute smooth electrostatic forces via Fast Fourier Transform, with a cost of order $N \log N$.

At this point, one might think that simulating an arbitrary collection of atoms can be done once one handles a good MD propagator and interatomic forces. The remaining question to be answered is how long should the MD trajectory be. The answer strictly depends on the type of system one is simulating, and in particular on the relaxation times exhibited by the system itself. In principle the length of the MD run should be greater than the largest relaxation time present in the system. This condition would ensure that sufficient statistics can be collected during the run and, more importantly, that time-dependent phenomena can be directly observed within the MD time window. Therefore, one of the crucial steps of biosimulation is to monitor whether if the macromolecules have sampled the available configurational space during the run. The simulation of proteins is particularly favorable. In fact, the phase space sampled by proteins is rather small, as compared to soft objects, as a consequence of their structural rigidity. However, in most cases proteins present relaxation times longer than the affordable trajectory length and many dynamical processes cannot be directly observed.

An important aspect is to choose a good set of quantities that allow to monitor carefully the sampling of macromolecules. In principle one could look at collective variables, such as the system potential energy or the overall volume, in the hope of observing either a stationary behavior, as an indication of the attainment of equilibration, or the dynamics of such extensive quantities, as an indication of the system characteristic times. Unfortunately, this route is often not viable since the slow structural variations are disguised by the large noise of these quantities and by error cancellations. A better choice is to consider the following "configurational distance"

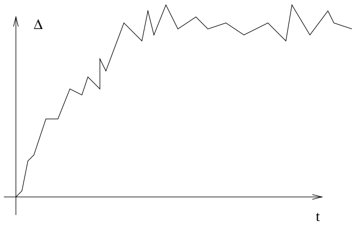

Fig. 4. The configurational distance quantifies the displacement of the macromolecule from a reference configuration. In this schematic plot, the reference configuration is taken as the one at initial time. At later times, the macromolecule diffuses away from the initial structure and reaches a plateau. The rate of convergence to the plateau measures the attainment of equilibration of the macromolecule.

$$\Delta(t) = \min_{T,R} \frac{1}{N} \sum_i \left| r_i(t) - r_i^{ref} \right|^2 \qquad (9)$$

where $\{r_i^{ref}\}$ is a set of reference coordinates, e.g. the coordinates of the folded structure, and the minimum is over irrelevant roto-translations of the whole molecule. $\Delta(t)$ is a measure of sampling in configurational space and is particularly well suited for folded proteins since it provides a sensible tool for monitoring both equilibration and sampling attitude on an atom-by-atom basis (Fig. 4).

A special topic in MD concerns the so-called rare events. When simulating complex materials it is likely that the simulation length is shorter than the duration of some event of interest, briefly indicated as $A \rightarrow B$, where A and B are the stable states of the transition. Two types of informations are relevant in the study of rare events, the shape of the reaction path and the associated free-energy profile, and the informations associated to the phase subspace orthogonal to the reaction path. The $A \rightarrow B$ transition can be inhibited by a number of dif-

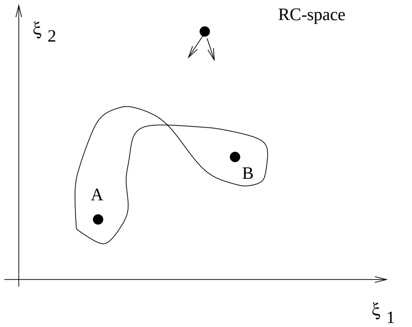

Fig. 5. A typical $A \to B$ transition takes place on a sub manifold of phase space, the so-called reaction space. The associated free energy landscape exhibits a saddle point where the path of minimal work will pass thru.

ferent reasons, typically because an intrinsic energetic or entropic barrier results in a dynamical bottleneck for the occurrence of the transition. In these circumstances, special techniques must be employed to force the occurrence of the rare event and to being capable of deriving the relevant informations for the "unforced" event In the following I will assume that the stable states A and B are known in advance, in the form of conformational states (Fig. 5).

There are several ways to handle rare events. At first, one should distinguish between two cases, when the reaction path is known or not known in advance. The case of known reaction path is by far the simplest one. In this circumstance, one can add some driving potential or constraint to walk along the reaction path via an artificial dynamics or a stepwise algorithm. The most convenient form of the biasing potential or constraint can be chosen by trial and error so to overcome the barrier. In the post-processing phase, a statistical mechanics treatment of the system at equilibrium allows to extract the thermodynamic and dynamical data related to the unbiased reaction, by means of appropriate procedures and reweighting formulae.

The most difficult case is when the reaction path is unknown and one has to employ some sort of blind-search procedure to both find the reaction path and retrieve the relevant structural and/or dynamical informations. Blind search approaches can be grouped into three main cathegories, global, local and thermodynamic methods. Global methods are those where an initial path is initially guessed and subsequently optimized. These methods are usually based on the principle of Least Action or modified actions [Olender & Elber (1996); Passerone & Parrinello (2001)] and they aim at obtaining the full trajectory in phase space rather in configurational space and its true dynamics. The drawback of global methods is that they need to sample a huge path space and require a good initial guess. The convergence capacity of global methods is not guaranteed in advance, since they rely on stationary conditions of the action rather than on a variational principle. Local methods are based on the construction of some sort of pseudo-dynamics with good local properties. As an example, one could find the path associated to the minimum energy path (MEP) along the reaction, corresponding to the zero-temperature path. Subsequently, the orthogonal subspace can be sampled by some dynamical or pseudo-dynamical method [Voter (1997); Melchionna (2000B)]. Unfortunately, in many circumstances, the finite-temperature path can be very different from the MEP one, so that the sampling would yield inconsistent results. Moreover, local methods do not usually provide the true dynamics, but they only sample the reaction path with a pseudo-dynamics. Thermodynamic methods are similar in spirit to the case of known reaction path. Here one guesses some driving potentials or constraints for the system at equilibrium or away from equilibrium. The way to refine the driving potentials works on a trial and error basis and can be fruitful if the reaction space has a rather simple topology. Other methods are finding wide interest, as the Path Sampling approach of Chandler and co-workers [Bolhuis et al. (1998)], in such they attempt to generate the reactive trajectories by stochastic or Monte Carlo dynamics in order to derive both the reaction path and the true dynamics. Before closing this brief review, I wish to underscore that Molecular Dynamics is a mature technique and its fundamentals are robust and well documented. Much research has yet to be invested to devise novel and smart methods for specific fields of applications, as for the case of biological simulations. On the practical side, a Molecular Dynamics programme can be rather articulated. In the case of biological matter, a large portion of the programme is dedicated to handling complex molecular topologies, potential functions and differentiated on-line and off-line tools for analysis. Moreover, the need for efficiency requires the use of parallel computing to simulate rather large systems ($10^4 - 10^5$ atoms) for a few tens of nanoseconds in a reasonable time. To this end, a few commercial or public domain softwares are available such as Amber [http://amber.scripps.edu/], Gromacs [http://www.gromacs.org/], Tinker [http://dasher.wustl.edu/tinker/], and others. Quite similarly, in the past years I have developed an independent MD package [Dlprotein (2001)]. Although these packages share several functionalities, for many purposes the choice of the MD engine and associated utilities can be quite crucial. In this respect, the philosophy underlying the software can be rather important as related to 1) providing a programme truly ”open” to other contributors and users, 2) with a broad range of functionalities and 3) with an algorithmic correctness satisfying the fundamental principles of Molecular Dynamics.

References

Allen M.P. &.Tildesley D.J, *Computer simulation of liquids* (Clarendon press, Oxford, 1987).

Bolhuis P. , Dellago C., & Chandler D., Faraday Discussions 1998, 110, 421.

Ciccotti G. & Ryckaert J.P., Comp.Phys.Rep. 1986, 4 , 345.

Melchionna S. & Cozzini S., “The DLPROTEIN User Manual”, 2001. Web site: http://www.sissa.it/cm/DLPROTEIN.

Eastwood J.W. &Hockney R.W., J.Comp.Phys. 1974, 16, 342.

Essmann U. *et al.*, J.Chem.Phys. 1995, 103, 8577.

Frenkel D. & Smit B., *Understanding Molecular Simulation* (Academic Press, London, 1996).

Greengard L. &Rokhlin V. , J.Comp.Phys. 1987, 73, 325.

Hansen J.-P., in *Molecular dynamics simulation of statistical mechanics systems*, edited by G.Ciccotti & W.G.Hoover (SIF, Como, 1986).

Melchionna S. , Ciccotti G., & B. L. Holian, Mol.Phys. 1993, 78, 533.

Melchionna S., Phys.Rev.E 2000, 61, 6165.

Melchionna S., Phys.Rev.E 2000, 62, 8762.

Olender R. & Elber R. , J. Chem. Phys. 1996, 105, 9299.

Passerone D. & Parrinello M., Phys.Rev.Lett. 2001, 87, 108302.

Stella L. & Melchionna S., J.Chem.Phys. 1998, 109, 10115.

Tuckerman M.E., Martyna G.J., &Berne B.J., J. Chem.Phys. 1992, 97, 1990.

van Gunsteren W.F. , Weiner P.K. &. Wilkinson Editors, *Computer Simulation of Biomolecular Systems A.J: Theoretical and Experimental Applications*, Vol.2, ESCOM, Leiden, 1993.

Voter A.F. , J. Chem. Phys. 1997, 106, 4665.