1 | Theory

1.1 Molecular dynamics

Molecular dynamics (MD) is a technique for studying the time evolution of N particles in a volume V. The aim of MD simulations is to compute the equilibrium and transport properties of a classical many-body system. The term *classical* means that the interaction between particles and their movement is studied over a given time t through a classical statistical mechanical treatment of the constituent particles. This is an appropriate approximation particularly for studying the dynamical properties systems where quantum effects are not considered; such that $hv \ll k_BT$. Indeed, many biomolecular systems can be reasonably well described using this classical-mechanical assumption. However, when relatively light particles such as electrons or light nuclei are involved, quantum effects must not be ignored. Such quantum effects can be modeled using a variety of t In a standard MD simulation, interaction forces between constituent particles are defined by a set of empirically-derived functions and parameters called *force fields* (see section ...). These functions are solved at discrete time steps through a numerical integration of Newton's equations of motion. In a conventional MD simulation, the total energy E and the total linear momentum ρ are constants of motion and so measure time averages in an ensemble very similar to the microcanonical ensemble (constant $NVE - \rho$). It is however possible to perform MD simulations in other statistical ensembles where temperature, stress or pressure can be kept as constant (see section 1.2). MD simulations provide a powerful predictive tool studying the dynamical behavior of highly complex molecular systems, with high spatial and temporal resolution. Consequently, such approaches have become a commonly used tool to understand various chemical and physical phenomena of organic molecules.

An MD simulation can be split into three steps: preparation of the system, equilibration and production run. In preparing the system, the initial coordinates of the molecule(s) are selected and placed into a periodic box. Solvent atoms are then added to the molecule-containing box. Finally, the desired force field parameters are defined as well as the desired statistical ensemble for the system, the integration method, the integration time step and the total simulation time.

How the system is equilibrated depends on the chosen statistical ensemble. If the microcanonical ensemble is used, an initial part of the production MD simulation can taken as the equilibration. If the canonical or isothermal-isobaric ensembles are used, however, it is necessary to choose a proper thermostat and a proper pressure coupling regime. In addition, a pre-run is required for converging the system to the desired temperature and/or pressure.

Providing the equilibration run does not produce significant fluctuations in the thermodynamic properties of the system, the production run portion of the simulation can be started. The ideal simulation length depends primarily on the time-scale of the phenomena of interest. Taking the intrinsic dynamical motion of protein atoms as an example, different atomic and molecular motions are dispersed across a wide range of time scales. The positions of amino acid side chains may fluctuate with a relatively with relatively high frequency (on the *ps* time scale), whereas large tertiary structural rearrangements of the entire protein may occur on much longer time scales (approaching the μs to ms time scales). In reality, the time scales of these motions largely varies from protein to protein and may depend on the physical and chemical properties of the polypeptide chain. However, slow structural dynamics and hence long time scales, still represent a significant computation cost. Apart from using a 'brute-force' approach, this can be overcome by decreasing the number of degrees of freedom of the system, or by biasing the energetic landscape such that the protein is able to traverse free energy barriers and explore conformational space otherwise inaccessible.

1.2 The statistical mechanical basis of molecular dynamics

1.3 Molecular dynamics in various ensembles

Considering a system with N particles in a volume V, the total energy E is a constant of motion. Assuming that time averages are equivalent to ensemble averages, the time averages obtained in a conventional MD simulation are equivalent to ensemble averages obtained from an NVE microcanonical ensemble. However, it is frequently desirable to simulate systems in other ensembles where, similar to experimental set-ups, pressure and temperature are kept constant. Physical and mathematical considerations for such statistical ensembles will be discussed bellow.

1.3.1 Constant temperature (NVT)

A constant temperature can be imposed by bringing the system in thermal contact with a large heat bath. This allows for the study of different molecular systems at different temperatures and sampling of the canonical statistical ensemble. Additionally, constant temperature simulations can help in avoiding steady energy drifts caused by the accumulation of numerical errors during an MD simulation. Under these conditions, the probability of the system populating a given energy state is given by the Maxwell-Boltzmann velocity distribution

$$P(p) = (\frac{\beta}{2\pi m})^{3/2} \cdot exp[\frac{-\beta p^2}{2m}] \tag{1.1}$$

We then obtain a simplified relation between the imposed temperature T and the kinetic energy for each particle in the system

$$k_B T = m\langle v_\alpha^2 \rangle \tag{1.2}$$

were k_B is the Boltzmann constant, m is the mass of the particle and $\langle v_\alpha^2 \rangle$ is the time averaged αth component of its velocity. This equivalence is often used to measure the temperature of a microcanonical system. Given that the temperature of a system is proportional (though not directly) to the average kinetic energy of the particles, the temperature can be controlled by scaling the velocities. If the temperature at time t is $T(t)$ and the velocities are multiplied by a factor λ, then the system temperature can be calculated by

$$\Delta T = \frac{1}{2}\sum_{i=1} 2\frac{m_i(\lambda v_i)^2}{N_{df} k_B} - \frac{1}{2}\sum_{i=1} 2\frac{m_i v_i^2}{N_{df} k_B} \tag{1.3}$$

$$\Delta T = (\lambda^2 - 1)T(t) \tag{1.4}$$

$$\lambda = \sqrt{\frac{T_0}{T(t)}} \tag{1.5}$$

multiplying the velocities at each time step by a factor $\lambda = \sqrt{T_0/T(t)}$, where $T(t)$ is the current temperature calculated from the kinetic energy and T_0 is the desired temperature provides the simplest way of keeping a constant system temperature. A problem with this approach is that it does not permit temperature fluctuations, which are present in the canonical ensemble.

A simpler formulation of velocity scaling is given by the Berendsen temperature coupling algorithm, in which the velocities are scaled at each step such that the rate of change of the temperature is proportional to the difference in temperature

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_0 - T(t)) \tag{1.6}$$

where τ is a parameter which determines how tightly the system is coupled to the thermal bath. The Berendsen temperature coupling algorithm gives an exponential decay of $T(t)$ towards T_0. The coupling parameter τ is given by

$$\lambda^2 = 1 + \frac{\delta t}{\tau} \cdot [\frac{T_0}{T(t - \frac{\delta t}{2})} - 1] \tag{1.7}$$

In practice the coupling parameter τ is empirical and the choice of its value alters the strength of coupling between the system and the thermal bath and should be chosen within a reasonable range. When $\tau \longrightarrow \infty$ the thermostat is inactive leading to the MD equation of motion, which samples a microcanonical ensemble. Conversely too small values of τ will cause unrealistically low temperature fluctuation. Values of $\tau \simeq 0.1ps$ are typically chosen for condensed-phase systems. Even when τ is properly chosen the system samples a *weak-coupling* ensemble, which is neither canonical nor microcanonical and systematically underestimates temperature fluctuations. This arises from the neglect of the stochastic contribution to temperature fluctuations on the microscopic timescale. Despite the empirical nature of the Berendsen coupling algorithm, it is very efficient for converging systems towards a desired temperature and consequently is commonly used in the equilibration step of an MD simulation.

To accurately probe the canonical ensemble the Extended System approach was first proposed simultaneously by Nosé and Hoover. The premise of the Nosé-Hoover algorithm is to use the extended Lagrangian to consider the thermal bath as part of the system by adding a dynamic variable $\tilde{s}$, which has a non-zero mass and a velocity $\dot{\tilde{s}}$. In the extended system the atomic coordinates are identical to the non-coupled system, however the time scale is stretched by the factor $\tilde{s}$ so that

$$dt = \tilde{s}dt \tag{1.8}$$

The Lagrangian for the extended system is given as

$$L_{Nose} = \sum_{i=1}^{N} \frac{m_i}{2} s^2 \dot{r}_i^2 - U(r^2) + \frac{Q}{2} \dot{s}^2 - g k_B T_0 ln\tilde{s} \tag{1.9}$$

where Q is the mass of $\tilde{s}$, g is the number of degrees of freedom of the system. The first two terms of the extended Lagrangian represent the potential energy subtracted from the kinetic energy of the real system. The third and forth terms represent the kinetic energy minus the potential energy assigned to $\tilde{s}$. The Nosé-Hoover thermostat allows the system to sample the canonical ensemble. The energy of the real system will fluctuate about a mean and accompanying the fluctuations of $\tilde{s}$, heat transfers occur between the system and the thermal bath, which regulate the system temperature.