1. Introduction

One of the principal tools in the theoretical study of biological molecules is the method of molecular dynamics simulations (MD). This computational method calculates the time dependent behavior of a molecular system. MD simulations have provided detailed information on the fluctuations and conformational changes of proteins and nucleic acids. These methods are now routinely used to investigate the structure, dynamics and thermodynamics of biological molecules and their complexes. They are also used in the determination of structures from x-ray crystallography and from NMR experiments.

Biological molecules exhibit a wide range of time scales over which specific processes occur; for example

- Local Motions (0.01 to 5 Å, 10^{-15} to 10^{-1} s)
  - Atomic fluctuations
  - Sidechain Motions
  - Loop Motions
- Rigid Body Motions (1 to 10Å, 10^{-9} to 1s)
  - Helix Motions
  - Domain Motions (hinge bending)
  - Subunit motions
- Large-Scale Motions (> 5Å, 10^{-7} to 10^{4} s)
  - Helix coil transitions
  - Dissociation/Association
  - Folding and Unfolding

The goal of this course is to provide an overview of the theoretical foundations of classical molecular dynamics simulations, to discuss some practical aspects of the method and to provide several specific applications within the framework of the CHARMM program. Although the applications will be presented in the framework of the CHARMM program, the concepts are general and applied by a number of different molecular dynamics simulation programs. The CHARMM program is a research program developed at Harvard University for the energy minimization and dynamics simulation of proteins, nucleic acids and lipids in vacuum, solution or crystal environments (Harvard CHARMM Web Page http://yuri.harvard.edu/).

Section I of this course will focus on the fundamental theory followed by a brief discussion of classical mechanics. In section II, the potential energy function and some related topics will be presented. Section III will discuss some practical aspects of molecular dynamics simulations and some basic analysis. The remaining sections will present the CHARMM program and provide some tutorials to introduce the user to the program. This course will concentrate on the classical simulation methods (i.e., the most common) that have contributed significantly to our understanding of biological systems.

Molecular dynamics simulations permit the study of complex, dynamic processes that occur in biological systems. These include, for example,

- Protein stability
- Conformational changes
- Protein folding
- Molecular recognition: proteins, DNA, membranes, complexes
- Ion transport in biological systems

and provide the mean to carry out the following studies,

- Drug Design
- Structure determination: X-ray and NMR

---

2. Historical Background

The molecular dynamics method was first introduced by Alder and Wainwright in the late 1950's (Alder and Wainwright, 1957,1959) to study the interactions of hard spheres. Many important insights concerning the behavior of simple liquids emerged from their studies. The next major advance was in 1964, when Rahman carried out the first simulation using a realistic potential for liquid argon (Rahman, 1964). The first molecular dynamics simulation of a realistic system was done by Rahman and Stillinger in their simulation of liquid water in 1974 (Stillinger and Rahman, 1974). The first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI) (McCammon, *et al*, 1977). Today in the literature, one routinely finds molecular dynamics simulations of solvated proteins, protein-DNA complexes as well as lipid systems addressing a variety of issues including the thermodynamics of ligand binding and the folding of small proteins. The number of simulation techniques has greatly expanded; there exist now many specialized techniques for particular problems, including mixed quantum mechanical - classical simulations, that are being employed to study enzymatic reactions in the context of the full protein. Molecular dynamics simulation techniques are widely used in experimental procedures such as X-ray crystallography and NMR structure determination.

References

Alder, B. J. and Wainwright, T. E. *J. Chem. Phys.* **27**, 1208 (1957)

Alder, B. J. and Wainwright, T. E. *J. Chem. Phys.* **31**, 459 (1959)

Rahman, A. *Phys. Rev.* **A136**, 405 (1964)

Stillinger, F. H. and Rahman, A. *J. Chem. Phys.* **60**, 1545 (1974)

McCammon, J. A., Gelin, B. R., and Karplus, M. *Nature (Lond.)* **267**, 585 (1977)

3. Statistical Mechanics

Molecular dynamics simulations generate information at the microscopic level, including atomic positions and velocities. The conversion of this microscopic information to macroscopic observables such as pressure, energy, heat capacities, etc., requires statistical mechanics. Statistical mechanics is fundamental to the study of biological systems by molecular dynamics simulation. In this section, we provide a brief overview of some main topics. For more detailed information, refer to the numerous excellent books available on the subject.

Introduction to Statistical Mechanics:

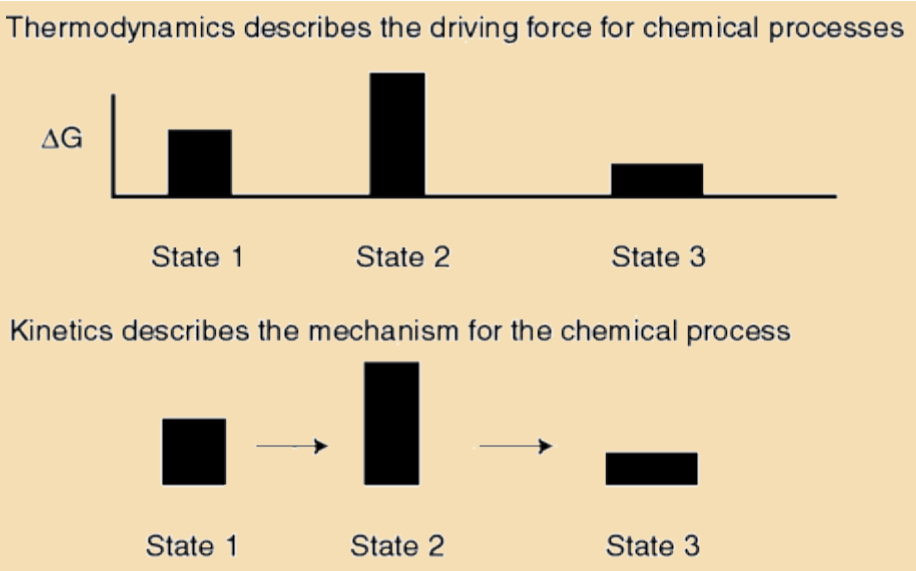
In a molecular dynamics simulation, one often wishes to explore the macroscopic properties of a system through microscopic simulations, for example, to calculate changes in the binding free energy of a particular drug candidate, or to examine the energetics and mechanisms of conformational change. The connection between microscopic simulations and macroscopic properties is made via *statistical mechanics* which provides the rigorous mathematical expressions that relate macroscopic properties to the distribution and motion of the atoms and molecules of the N-body system; molecular dynamics simulations provide the means to solve the equation of motion of the particles and evaluate these mathematical formulas. With molecular dynamics simulations, one can study both thermodynamic properties and/or time dependent (kinetic) phenomenon.

Reference Textbooks on Statistical Mechanics

D. McQuarrie, Statistical Mechanics (Harper & Row, New York, 1976)

D. Chandler, Introduction to Modern Statistical Mechanics (Oxford University Press, New York, 1987)

R. E. Wilde and S. Singh, Statistical Mechanics, Fundamentals and Modern Applications (John Wiley & Sons, Inc, New York, 1998)



Statistical mechanics is the branch of physical sciences that studies macroscopic systems from a molecular point of view. The goal is to understand and to predict macroscopic phenomena from the properties of individual molecules making up the system. The system could range from a collection of solvent molecules to a solvated protein-DNA complex. In order to connect the macroscopic system to the microscopic system, time independent statistical averages are often introduced. We start this discussion by introducing a few definitions.

Definitions

The *thermodynamic state* of a system is usually defined by a small set of parameters, for example, the temperature, T, the pressure, P, and the number of particles, N. Other thermodynamic properties may be derived from the equations of state and other fundamental thermodynamic equations.

The *mechanical* or *microscopic state* of a system is defined by the atomic positions, **q**, and momenta, **p**; these can also be considered as coordinates in a multidimensional space called **phase space**. For a system of N particles, this space has 6N dimensions. A single point in phase space, denoted by **Γ**, describes the state of the system. An*ensemble* is a collection of points in phase space satisfying the conditions of a particular thermodynamic state. A molecular dynamics simulations generates a sequence of points in phase space as a function of time; these points belong to the same ensemble, and they correspond to the different conformations of the system and their respective momenta. Several different ensembles are described below.

An ensemble is a collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state.

There exist different ensembles with different characteristics.

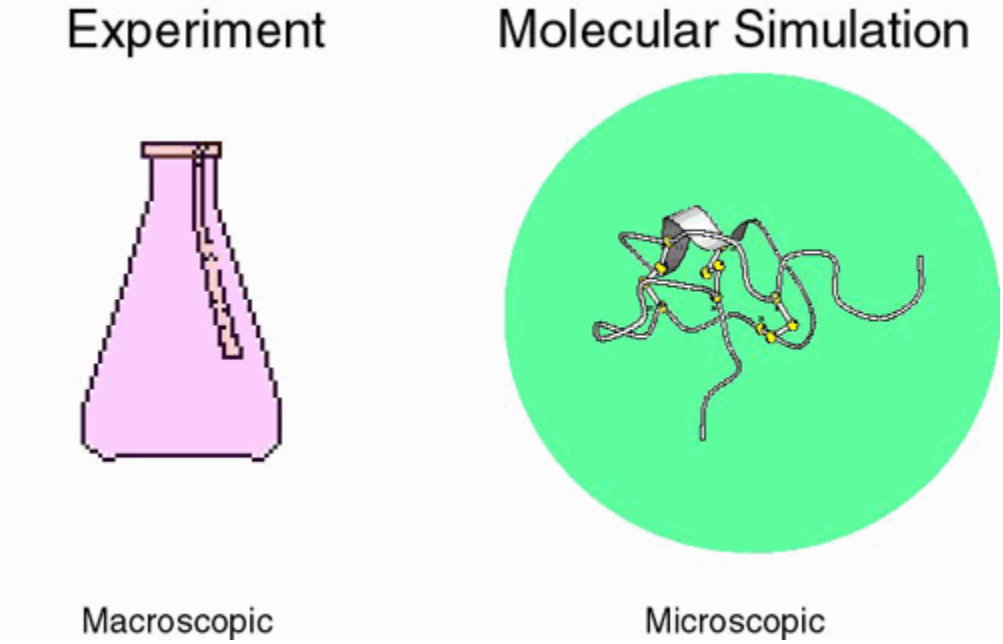
- Microcanonical ensemble (NVE) : The thermodynamic state characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed energy, E. This corresponds to an isolated system.
- Canonical Ensemble (NVT): This is a collection of all systems whose thermodynamic state is characterized by a fixed number of atoms, N, a fixed

volume, V, and a fixed temperature, T.

- Isobaric-Isothermal Ensemble (NPT): This ensemble is characterized by a fixed number of atoms, N, a fixed pressure, P, and a fixed temperature, T.
- Grand canonical Ensemble (μVT): The thermodynamic state for this ensemble is characterized by a fixed chemical potential, μ, a fixed volume, V, and a fixed temperature, T.

Calculating Averages from a Molecular Dynamics Simulation

An experiment is usually made on a macroscopic sample that contains an extremely large number of atoms or molecules sampling an enormous number of conformations. In statistical mechanics, averages corresponding to experimental observables are defined in terms of ensemble averages; one justification for this is that there has been good agreement with experiment. An ensemble average is average taken over a large number of replicas of the system considered simultaneously.



In statistical mechanics, average values are defined as ensemble averages.

The ensemble average is given by

$$\langle A \rangle_{ensemble} = \iint dp^N dr^N A\left(p^N, r^N\right)\rho\left(p^N, r^N\right)$$

where

$$A\left(p^N, r^N\right)$$

is the observable of interest and it is expressed as a function of the momenta, *p*, and the positions, *r*, of the system. The integration is over all possible variables of *r* and *p*.

The probability density of the ensemble is given by

$$\rho\left(p^N, r^N\right) = \frac{1}{Q}\exp\left[-H\left(p^N, r^N\right) / k_B T\right]$$

where *H* is the Hamiltonian, *T* is the temperature, *kB* is Boltzmann's constant and *Q* is the partition function

$$Q = \iint dp^N dr^N \exp\left[-H\left(p^N, r^N\right) / k_B T\right]$$

This integral is generally *extremely* difficult to calculate because one must calculate all possible states of the system. In a molecular dynamics simulation, the points in the ensemble are calculated sequentially in time, so to calculate an ensemble average, the molecular dynamics simulations must pass through all possible states corresponding to the particular thermodynamic constraints.

Another way, as done in an MD simulation, is to determine a time average of A, which is expressed as

$$\langle A\rangle_{time} = \lim_{\tau\to\infty}\frac{1}{\tau}\int_{t=0}^{\tau} A\left(p^N(t), r^N(t)\right)dt \approx \frac{1}{M}\sum_{t=1}^{M} A\left(p^N, r^N\right)$$

where τ is the simulation time, M is the number of time steps in the simulation and *A(pN,rN)* is the instantaneous value of *A*.

The dilemma appears to be that one can calculate time averages by molecular dynamics simulation, but the experimental observables are assumed to be ensemble averages. Resolving this leads us to one of the most fundamental axioms of statistical mechanics, the **ergodic hypothesis,** which states that the time average equals the ensemble average.

The **Ergodic hypothesis** states

$$\langle A\rangle_{ensemble} = \langle A\rangle_{time}$$

Ensemble average = Time average

The basic idea is that if one allows the system to evolve in time indefinitely, that system will eventually pass through all possible states. One goal, therefore, of a molecular dynamics simulation is to generate enough representative conformations such that this equality is satisfied. If this is the case, experimentally relevant information concerning structural, dynamic and thermodynamic properties may then be calculated using a feasible amount of computer resources. Because the simulations are of fixed duration, one must be certain to sample a sufficient amount of phase space.

Some examples of time averages:

Average potential energy

$$V = \langle V \rangle = \frac{1}{M} \sum_{i=1}^{M} V_i$$

where *M* is the number of configurations in the molecular dynamics trajectory and *Vi* is the potential energy of each configuration.

Average kinetic energy

$$K = \langle K \rangle = \frac{1}{M} \sum_{j=1}^{M} \left\{ \sum_{i=1}^{N} \frac{m_i}{2} v_i \cdot v_i \right\}_j$$

where *M* is the number of configurations in the simulation, *N* is the number of atoms in the system, *mi* is the mass of the particle *i* and *vi* is the velocity of particle *i*.

A molecular dynamics simulation must be sufficiently long so that enough representative conformations have been sampled.

---

4. Classical Mechanics

The molecular dynamics simulation method is based on Newton's second law or the equation of motion, **F=ma**, where **F** is the force exerted on the particle, **m** is its mass and **a** is its acceleration. From a knowledge of the force on each atom, it is possible to determine the acceleration of each atom in the system. Integration of the equations of motion then yields a trajectory that describes the positions, velocities and accelerations of the particles as they vary with time. From this trajectory, the average values of properties can be determined. The method is deterministic; once the positions and velocities of each atom are known, the state of the system can be predicted at any time in the future or the past. Molecular dynamics simulations can be time consuming and computationally expensive. However, computers are getting faster and cheaper. Simulations of solvated proteins are calculated up to the nanosecond time scale, however, simulations into the millisecond regime have been reported.

Newton's equation of motion is given by

$$F_i = m_i a_i$$

where F_i is the force exerted on particle *i*, m_i is the mass of particle *i* and a_i is the acceleration of particle *i*. The force can also be expressed as the gradient of the potential energy,

$$F_i = -\nabla_i V$$

Combining these two equations yields

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2}$$

where *V* is the potential energy of the system. **Newton's** equation of motion can then relate the derivative of the potential energy to the changes in position as a function of time.

Newton's Second Law of motion: a simple application

$$F = m \cdot a = m \cdot \frac{dv}{dt} = m \cdot \frac{d^2 x}{dt^2}$$

Taking the simple case where the acceleration is constant,

$$a = \frac{dv}{dt}$$

we obtain an expression for the velocity after integration

$$v = at + v_0$$

and since

$$v = \frac{dx}{dt}$$

we can once again integrate to obtain

$$x = v \cdot t + x_0$$

Combining this equation with the expression for the velocity, we obtain the following relation which gives the value of x at time t as a function of the acceleration, a, the initial position, x_0 , and the initial velocity, v_0..

$$x = a \cdot t^2 + v_0 \cdot t + x_0$$

The acceleration is given as the derivative of the potential energy with respect to the position, *r*,

$$a = -\frac{1}{m}\frac{dE}{dr}$$

Therefore, to calculate a trajectory, one only needs the initial positions of the atoms, an initial distribution of velocities and the acceleration, which is determined by the gradient of the potential energy function. The equations of motion are deterministic, e.g., the positions and the velocities at time zero determine the positions and velocities at all other times, t. The initial positions can be obtained from experimental structures, such as the x-ray crystal structure of the protein or the solution structure determined by NMR spectroscopy.

The initial distribution of velocities are usually determined from a random distribution with the magnitudes conforming to the required temperature and corrected so there is no overall momentum, i.e.,

$$P = \sum_{i=1}^{N} m_i v_i = 0$$

The velocities, *vi,* are often chosen randomly from a Maxwell-Boltzmann or Gaussian distribution at a given temperature, which gives the probability that an atom i has a velocity v_x in the x direction at a temperature T.

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{m_i v_{ix}^2}{k_B T}\right]$$

The temperature can be calculated from the velocities using the relation

$$T = \frac{1}{(3N)}\sum_{i=1}^{N}\frac{|p_i|}{2m_i}$$

where N is the number of atoms in the system.

Integration Algorithms

The potential energy is a function of the atomic positions (3N) of all the atoms in the system. Due to the complicated nature of this function, there is *no analytical solution* to the equations of motion; they must be solved numerically.

Numerous numerical algorithms have been developed for integrating the equations of motion. We list several here.

- Verlet algorithm
- Leap-frog algorithm
- Velocity Verlet
- Beeman's algorithm

Important: In choosing which algorithm to use, one should consider the following criteria:

- The algorithm should conserve energy and momentum.
- It should be computationally efficient
- It should permit a long time step for integration.

Integration Algorithms

All the integration algorithms assume the positions, velocities and accelerations can be approximated by a Taylor series expansion:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \ldots$$

$$v(t+\delta t) = v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \ldots$$

$$a(t+\delta t) = a(t) + b(t)\delta t + \ldots$$

Where ***r*** is the position, ***v*** is the velocity (the first derivative with respect to time), ***a*** is the acceleration (the second derivative with respect to time), etc.

To derive the **Verlet** algorithm one can write

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2$$

$$r(t-\delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2$$

Summing these two equations, one obtains

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2$$

The Verlet algorithm uses positions and accelerations at time t and the positions from time t-δt to calculate new positions at time t+δt. The Verlet algorithm uses no explicit velocities. The advantages of the Verlet algorithm are, *i)* it is straightforward, and *ii)* the storage requirements are modest . The disadvantage is that the algorithm is of moderate precision.

The Leap-frog algorithm

$$r(t+\delta t) = r(t) + v\left(t+\frac{1}{2}\delta t\right)\delta t$$

$$v\left(t+\frac{1}{2}\delta t\right) = v\left(t-\frac{1}{2}\delta t\right) + a(t)\delta t$$

In this algorithm, the velocities are first calculated at time t+$1/2\delta t$; these are used to calculate the positions, r, at time t+δt. In this way, the velocities *leap* over the positions, then the positions *leap* over the velocities. The advantage of this algorithm is that the velocities are explicitly calculated, however, the disadvantage is that they are not calculated at the same time as the positions. The velocities at time t can be approximated by the relationship:

$$v(t) = \frac{1}{2}\left[v\left(t-\frac{1}{2}\delta t\right) + v\left(t+\frac{1}{2}\delta t\right)\right]$$

The Velocity Verlet algorithm

This algorithm yields positions, velocities and accelerations at time t. There is no compromise on precision.

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2$$

$$v(t+\delta t) = v(t) + \frac{1}{2}\left[a(t) + a(t+\delta t)\right]\delta t$$

Beeman's algorithm

This algorithm is closely related to the Verlet algorithm

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{2}{3}a(t)\delta t^2 - \frac{1}{6}a(t-\delta t)\delta t^2$$

$$v(t+\delta t) = v(t) + v(t)\delta t + \frac{1}{3}a(t)\delta t + \frac{5}{6}a(t)\delta t - \frac{1}{6}a(t-\delta t)\delta t$$

The advantage of this algorithm is that it provides a more accurate expression for the velocities and better energy conservation. The disadvantage is that the more complex expressions make the calculation more expensive.