International Tables for Crystallography (2006). Vol. F, Section 20.2.3, pp. 489–491.

20.2. Molecular-dynamics simulations of biological macromolecules

BY C. B. POST AND V. M. DADARLAT

20.2.1. Introduction

Molecular dynamics (MD) is the simulation of motion for a system of particles. Advances in the theory of atomic interactions and the increasing availability of high-power computers have led to rapid development of this field and greater understanding of macromolecular motions. In the earliest molecular-dynamics simulations of protein molecules (McCammon *et al.*, 1977; McCammon & Harvey, 1987), the systems were greatly simplified in order to fit within the computing capabilities of that time. Simplifications included the exclusion of water molecules and even of explicit hydrogen atoms; the effect of hydrogen atoms was built into the heavy-atom properties using so-called extended-atom parameters. Simulation time periods were limited to tens of picoseconds for systems of less than 10^3 atoms. Modern simulations, by contrast, are based on improved force fields (MacKerell *et al.*, 1998) and benefit from considerable development in algorithms. In addition, the possible size and time period of simulations have increased by orders of magnitude; large systems of the order of 10^4 atoms (including explicit solvent molecules) and nanosecond time periods are accessible. With dedicated computer time, the microsecond regime is possible (Duan & Kollman, 1998). Interestingly, the first 100 ps simulation of an enzyme complex was of hen egg-white lysozyme (Post *et al.*, 1986), the first enzyme whose structure was solved by X-ray crystallography. Then the simulation required several months of dedicated time on a Cray supercomputer, but now it can be accomplished in less than a week on a common workstation.

A consequence of this enormous growth in computing power has been the particularly successful application of molecular dynamics of biological molecules to three-dimensional structure determination and refinement. It is now practical to use molecular dynamics, in combination with crystallographic and NMR data, to search the large conformational space of proteins and nucleic acids to find structures consistent with the data and to improve the agreement with the data. The advantages of molecular dynamics over manual rebuilding and least-squares refinement are the abilities to overcome the local minimum problem in an automated fashion and to search the complex conformational space of a macromolecule more extensively (Brünger *et al.*, 1987).

20.2.2. The simulation method

Molecular mechanics, whereby the energy of the system is expressed in classical terms as a function of atomic coordinates, is well established as a useful approach for describing atomic interactions (Brooks *et al.*, 1988; Goodfellow & Levy, 1998). Owing to the size of proteins and nucleic acids, the potential-energy function for large biomolecules is empirically based rather than derived from quantum-mechanical calculations. The total force on each atom, $\mathbf{F}_i$, is calculated from the gradient, or the first derivative of this potential energy, with respect to the atomic coordinates. The motion of the atom resulting from the net force is described by Newton's equation of motion,

$$\mathbf{F}_i = m_i\mathbf{a}_i, \qquad (20.2.2.1)$$

where m_i is the mass of atom i and $\mathbf{a}_i$ is the acceleration. Integration of equation (20.2.2.1) gives

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \mathbf{F}_i(t)(\Delta t)^2/(2m_i), \qquad (20.2.2.2)$$

where Δt is the time step in the integration, $\mathbf{r}_i(t+\Delta t)$ is the atomic position at time $(t+\Delta t)$, given the position $\mathbf{r}_i(t)$ at time t, and $\mathbf{v}_i(t)$ is the velocity. The forces on the particles change continuously so that a numerical solution of the equation is required. The Verlet algorithm (Verlet, 1967) or a variation, Leapfrog, is commonly used.

20.2.3. Potential-energy function

20.2.3.1. *Empirical energy*

The central element of simulations is the interaction potential between atoms as a function of atomic position, **r**. The success of simulations in describing the average structure of proteins and other biological features suggests that such relatively simple potential functions adequately represent proteins and nucleic acids. The empirically based components of the energy function, $E_{\rm empir}$, include geometric terms for bond lengths, bond angles and torsion angles, and non-bonding terms for steric van der Waals interactions and electrostatic interactions. A commonly used energy function is

$$E_{\rm empir} = E_{\rm geom} + E_{\rm nonb}, \qquad (20.2.3.1)$$

$$E_{\rm geom} = \sum_{\rm bonds} (1/2)k_b(b-b_{\rm eq})^2 + \sum_{\rm angles} (1/2)k_\theta(\theta-\theta_{\rm eq})^2 + \sum_{\rm torsions} (1/2)k_\varphi[1+\cos(n\varphi-\delta)], \qquad (20.2.3.2)$$

$$E_{\rm nonb} = \sum_{\rm nbpairs} \left(q_iq_j/Dr_{ij}\right) + 4\varepsilon_{ij}\left[\left(\sigma_{ij}/r_{ij}\right)^{12} - \left(\sigma_{ij}/r_{ij}\right)^{6}\right], \qquad (20.2.3.3)$$

where k_b, k_θ and k_φ are force constants, $b_{\rm eq}$ and $\theta_{\rm eq}$ are equilibrium values for bond lengths, b, and angles, θ, respectively, and φ is the torsion angle of periodicity n and phase δ. The non-bonded terms depend on the interatomic distance r_{ij}, the dielectric constant D, the partial atomic charge q_i, and the van der Waals parameters ε_{ij} and σ_{ij}. The bond-stretching and angle-bending contributions are represented by harmonic potentials, while the energy associated with rotation about a bond, the torsional potential, is modelled by a cosine function [equation (20.2.3.2)]. The electrostatic component of the non-bonded interactions [first term of equation (20.2.3.3)] follows Coulomb's Law, and a Lennard–Jones 6–12 potential function [second term of equation (20.2.3.3)] is used to model steric repulsion and attractive dispersion interactions. $E_{\rm nonb}$, as a sum over pairs of atoms not involved in either a bond or bond angle, requires the use of a pairwise list between atoms. The small contribution from pairs separated by a large distance allows the use of cutoff limits for this list, but at some cost in accuracy.

Initial values for the atomic coordinates and velocities are required to begin the molecular-dynamics simulation. While initial coordinates are obtained from the model built into the electron-density map, it is necessary to generate initial velocities computationally. The most common approach is to assign random values for each atom, i, consistent with the temperature chosen for the system: $3kT/2 = (1/2)\sum m_iv_i^2$.

Integration of the equation of motion [equation (20.2.2.1)] also requires specification of the time step Δt. In the case of structure determination, this choice is limited only by the numerical stability of the calculation. Too large a value for Δt results in errors in the integration, manifested by a rapid and unacceptable increase in energy. Whereas a value for Δt of 1 to 2 fs is required for accurate trajectories and strict conservation of the energy, structure-determination protocols can employ larger values and are limited only by the need for numerical stability.

Both the temperature and Δt influence the sampling rate of conformational space. Enhanced sampling increases the rate of

convergence to the structure solution. Issues related to sampling are detailed in Chapter 18.2.

20.2.3.2. *Particle mesh Ewald*

In an MD simulation, the accurate and rapid calculation of long-range electrostatic interactions is a central issue for the correct physical representation of the system. The Coulombic potential [first term of equation (20.2.3.3)] has been used in most cases, but the Coulombic interactions must be limited for practical reasons to a subset of pair interactions so that long-distance interactions are truncated. More recently, the Ewald method has been implemented to avoid the need for truncating the non-bonded pair list while maintaining computational efficiency. The electrostatic energy is calculated using periodic boundary conditions* and requires a double summation over all atoms in the central unit cell, as well as their infinite number of periodic images.

The total electrostatic energy of a periodic system with a neutral unit cell containing N point charges $q_1, q_2, \ldots, q_n$ located at positions $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ is

$$E_{\rm elec} = \tfrac{1}{2}{\sum_{n}}' \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{q_i q_j}{|\mathbf{r}_{ij}+\mathbf{n}|}.$$

The sum over $|\mathbf{n}|$ takes into account all the periodic unit cells, and $\mathbf{n}$ reflects the shape of the unit cell. For a simple periodic cubic lattice, $\mathbf{n} = (n_x L, n_y L, n_z L)$, where L is the length of the unit cell, and n_x, n_y and n_z are integers. The prime indicates that in the central unit cell the terms $i = j$ are not included. The sum is convergent for very large $|\mathbf{n}|$.

An efficient way of evaluating the triple sum above is represented by the Ewald method. In this method, each point charge is surrounded by a Gaussian charge distribution of opposite sign and same integrated magnitude as the original charge:

$$\rho_i(\mathbf{r}) = -q_i \beta^3 \exp(-\beta^2 \mathbf{r}^2)/\pi^{3/2}.$$

β is inversely proportional to the width of the charge distribution and $\mathbf{r}$ is the position relative to the centre of the distribution. The result of adding this charge distribution is a screening of the electrostatic interactions, so that the interaction between neighbouring charges is effectively short-ranged. A total screened potential is calculated by summation in the real-space lattice over the initial point-charge distribution together with the screening-charge distribution. To end up with the original point-charge distribution, a cancelling distribution having the opposite sign and same functional form as the Gaussian screening distribution must be added. Together, these Gaussian distributions form a smooth varying function of $\mathbf{r}$, which can be approximated by a superposition of continuous functions. The contribution of the cancelling distribution is calculated by adding the Fourier transforms of the distributions in reciprocal space at each point charge and transforming the total back into real space.

The final Ewald expression for the total electrostatic energy contains a real-space sum plus a reciprocal-space sum, minus a self-term introduced by the interaction of the cancelling distribution with itself:

$$E_{\rm elec} = \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\sum_{n} q_i q_i \frac{\operatorname{erfc}(\beta|\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|} + q_i q_j \Phi_{\rm rec}(\beta \mathbf{r}_{ij})\right] - \frac{\beta}{\pi^{1/2}}\sum_{i=1}^{N} q_i^{\,2} + J(\mathbf{D}, \mathbf{P}, e').$$

$J(\mathbf{D}, \mathbf{P}, e')$ (de Leeuw *et al.*, 1980) depends on the total dipole moment of the unit cell, the shape of the macroscopic crystal lattice and the dielectric constant of the surrounding medium. The term $\operatorname{erfc}(\beta|\mathbf{r}_{ij}+\mathbf{n}|)$ is the complementary error function which falls to zero as $(\beta|\mathbf{r}_{ij}+\mathbf{n}|)$ increases. The term $\Phi_{\rm rec}(\beta \mathbf{r}_{ij})$ is the reciprocal Ewald sum:

$$\Phi_{\rm rec}(\beta \mathbf{r}_{ij}) = \frac{1}{\pi V}\sum_{m\neq 0} \frac{\exp(-\pi^2 \mathbf{m}^2/\beta^2)}{\mathbf{m}^2} \exp(2\pi i \mathbf{m}\mathbf{r}).$$

V is the volume and $\mathbf{m}$ is a vector in the reciprocal space. The Ewald sum in reciprocal space is the solution to Poisson's equation, with Gaussian charge densities as sources and with periodic boundary conditions. For large β, the only contribution to the sum in the real space comes from the minimum image terms. When β is large, the charge distributions are sharp and the summation in reciprocal space must be done over a large number of reciprocal-space vectors to achieve convergence.

Calculation of the Ewald sum in the original form is slow for large macromolecular solution simulations. The usual implementation of the Ewald summation for calculating electrostatic interaction is an order N^2 algorithm, or at best, $N^{3/2}$ (Christiansen *et al.*, 1993) by adjusting β to optimize computational effort. The particle mesh Ewald (PME) method (Essmann *et al.*, 1995; Darden *et al.*, 1993) is an approximation of the Ewald sum that allows an order N log N algorithm for the calculation of the total electrostatic energy. This reduction in the number of steps is accomplished by choosing β large enough that atom pairs separated by distances larger than a specified cutoff (9–12 Å) make a negligible contribution to the direct-space sum, which thus becomes an order N computation. Moreover, the reciprocal-space sum is expressed in terms of the electrostatic structure factors. Each atomic charge distribution is approximated by a gridded charge distribution. The resulting approximate structure factors are calculated by a three-dimensional fast Fourier transform applied to the grid. Using an optimized β and grid density for each simulated system, the PME method can compute the electrostatic energy for a periodic boundary system at the same level of accuracy as the classical Ewald summation in a much shorter time (Darden *et al.*, 1993).

20.2.3.3. *Experimental restraints in the energy function*

For the purpose of structure determination, the potential-energy function used for molecular-dynamics calculation incorporates the information from experimental data in the form of non-physical restraint terms. These restraint terms, introduced to bias the conformational sampling toward structures consistent with the experiment, are used in addition to the total potential function and, in some sense, can fulfil the requirements of a physical term in equation (20.2.2.2) (see below). The experimentally based restraint terms are added to the potential-energy function to give a total effective potential, $E_{\rm tot} = E_{\rm empir} + E_{\rm rest}$. Whereas structure-determination protocols based on NMR data employ a number of types of restraint terms, data from X-ray crystallography provide a single restraint term, $E_{\rm rest} = wE_{\rm Xray}$, the residual between the observed and

* The periodic boundary conditions method allows for the simulation of the atoms in a central unit cell in the field of image unit cells generated by symmetry operations.

calculated structure-factor amplitudes; where

$$wE_{\rm Xray} = w\sum_{hkl}(|F_o| - k|F_c|)^2,$$

where w and k are scale factors.

20.2.4. Empirical parameterization of the force field

Considerable effort has gone into the development of a number of force fields for use in molecular-dynamics simulations of biomolecules (Jorgensen & Tirado-Rives, 1988; MacKerell *et al.*, 1995, 1998; van Gunsteren *et al.*, 1996). The parameters described here are those of the *CHARMM*22 force field, the force field used in the *X-PLOR* program. Estimation of the force constants, equilibrium values and non-bonding parameters in equations (20.2.3.2) and (20.2.3.3) involves a self-consistent approach that balances the bonding and non-bonding interaction terms among the macromolecule and solvent molecules (MacKerell *et al.*, 1998). A wide range of data are taken into account during an interactive process of optimization in order to adequately account for the extensive and correlated nature of the parameters in a consistent fashion. Small-molecule model compounds representative of proteins or nucleic acids are considered in detail, and a hierarchical approach is applied to extend the parameters to larger molecules with minimal adjustment at the points of connection.

The empirical basis of the parameters is broad. Gas-phase geometries and crystal structures are used to determine equilibrium bond lengths, bond angles, and dihedral phase and periodicity. Vibrational spectra, primarily from gas-phase infrared and Raman spectroscopy, are used to fit values for the force constants. Torsion-angle terms are estimated from relative energies of different conformers of model compounds, such as 4-ethylimidazole and ethylbenzene, based on gas-phase data. In cases where no satisfactory experimental data are available, *ab initio* calculations are used to obtain the required energy surfaces. Adjustments are made to describe the energy barriers and positions of saddle points, as well as the minimum-energy structures.

Optimization of the non-bonded parameters includes fitting the van der Waals and electrostatic terms of equation (20.2.3.3), while maintaining a balance among the protein–protein, water–water and protein–water interactions. The parameterization of the *CHARMM*22 force field is based on the water model and water–water interactions of the *TIP*3*P* model (Jorgensen *et al.*, 1983). As such, use of this parameter set with another water model will lead to inconsistencies in the balance of intermolecular interactions. Data from dipole moments, heats and free energies of vaporization, solvation and sublimation, and molecular volumes, as well as *ab initio* calculations of interaction energies and geometries are used to optimize intermolecular interactions. Partial charges of atoms are determined by fitting *ab initio* interaction energies and geometries of small-molecule compounds that model the peptide backbone and amino-acid side chains. Magnitudes and directions of dipole-moment values are also used to optimize partial charges. Experimental gas-phase dipole-moment values are used when available, while *ab initio* calculated values are adopted otherwise. The van der Waals parameters are then refined by comparing results of condensed-phase simulations on pure solvents with heats of vaporization and molecular volumes.

The crystallographic restraint term in the potential-energy function, $E_{\rm rest}$, must also be parameterized to optimize the agreement with the experimental structure-factor amplitudes while simultaneously retaining good geometry and non-bonding interactions. Optimization of $E_{\rm rest} = wE_{\rm Xray}$ involves only the estimation of w. Unlike the parameters in $E_{\rm empir}$, w has no physical basis and is usually chosen to make the force due to $E_{\rm rest}$ balance the total force contributed by all terms in $E_{\rm empir}$. As refinement of the structure progresses, these forces, and hence w, necessarily change since the quality, in terms of geometry and non-bonding interactions, of the structure improves and the crystallographic residual is reduced.

20.2.5. Modifications in the force field for structure determination

Simulated-annealing protocols require modification of the parameters to maintain the correct geometry and local structural integrity of the molecule in order to allow heating to very high, non-physical temperatures for several thousand integration steps. Such modifications are acceptable in the case of structure determination since the primary goal is to define the optimum equilibrium structure in best agreement with the crystallographic or NMR data. Simulations intended to reproduce the fluctuations or dynamic properties of the system must employ carefully defined parameters without such modifications. These modifications include substantial increases in the force constants for bond lengths and angles, *e.g.*, factors of two to ten are used in the parameters specified in the *X-PLOR* file parallhdg.pro. A number of improper torsional terms are added to maintain proper chirality.

The specific terms in $E_{\rm nonb}$ are also modified for the purpose of structure determination. In this methodology, the goal is to converge efficiently to a model that satisfies the experimental data, rather than to obtain an accurate description of the conformational surface, such as estimating fluctuations and equilibrium distributions. Alterations in $E_{\rm nonb}$ include the replacement of the computationally expensive $E_{\rm vdW}$ by a quartic or harmonic repulsive term, which prevents steric conflict among atoms, but ignores dispersive attraction. The electrostatic term, $E_{\rm elec}$, is frequently excluded altogether, since the $1/\mathbf{r}$ dependence of the Coulombic potential allows charge interactions to dominate the interatomic forces far from the global minimum in a fashion that hinders movement toward the global minimum. Exclusion of this important physical property of biological systems is possible, because the crystallographic structure factors contain sufficient information to reflect adequately the imprint of electrostatics on the average structure.

20.2.6. Internal dynamics and average structures

It is most often the goal of the structural biologist to define a single average structure of a macromolecule. The well recognized internal motions arising from thermal fluctuations of a macromolecule may be necessary for function, but, nonetheless, the methods of structure determination generally aim to model a single average structure. Internal motions range from the high frequency, small amplitude motions (*i.e.* those modelled by crystallographic B values) to low frequency, larger amplitude motions of loops and whole domains. Some studies (Kuriyan *et al.*, 1986; Post, 1992) have examined the validity of the assumptions about fast timescale motions made by the methods of structure determination. It is reasonable that some of the differences between the structure solutions of a protein obtained by NMR spectroscopy and X-ray crystallography are due to differences in the effects of internal motions. The application of molecular-dynamics algorithms for structure determination has allowed the use of protocols that account for effects of internal motions by employing time-averaged restraints (Schiffer *et al.*, 1995).

references