Molecular Modeling of Nucleic Acid Structure: Energy and Sampling

Building on the introduction to molecular modeling of nucleic acid structure presented in *UNIT 7.5*, this unit discusses computational and theoretical methods aimed at giving greater insight into the structure and energy of nucleic acids. This includes describing means for representing the energy in model nucleic acid structures and methods for sampling relevant conformational states.

THE ENERGY REPRESENTATION

To aid a modeler in judging the reliability of a given model structure, it is useful to include some representation of the energy. The presumption is that structures with lower energies—such as those with less steric overlap, less distortion, and more favorable interactions—will be more representative. As discussed (*UNIT 7.5*), adding a representation of the energy to a physical model is difficult; however, some implementation of the energy for a given molecular model can be easily programmed on a computer. In common use, the level of theory applied spans the range from highly accurate ab initio quantum mechanical (QM) to simpler, empirical molecular mechanical energy treatments. As the accuracy of the energy increases, so does the computational cost. The increase in computational cost is tremendous and limits the level of theory that may be applied. For calculating the energy of a single model structure, very accurate methods can be applied; however, if one is interested in dynamics or investigating the energy of many configurations of a given model, less accurate energy representations are generally necessary.

Quantum Mechanical (QM) Treatments

To provide an accurate and complete theoretical description of the energy (as a function of the atomic and electronic configuration or structure) of a molecular system, ab initio QM treatments can be applied. Standard codes for performing QM calculations include Gaussian, Jaguar, Q-Chem, and GAMESS (see Internet Resources). Performing calculations with these programs is fairly straightforward, even for those without a broad theoretical background in the methods. For reviews of QM methods, see the books by Szabo and Ostlund (1989) or Levine (1991). A serious drawback of these methods is that they are extremely computationally demanding, which typically limits accurate QM calculations to small model systems (<100 atoms). Recent improvements in the methods, coupled with the availability of greater computational power, have led to more involved and highly accurate QM calculations. Accuracy is improved with the use of larger and higher level basis sets and with the inclusion of electron correlation. More recently there has also been an increase in the use of density functional methods, which allow the investigation of larger systems (and implicitly include some electron correlation); however, even with these improvements, tractable model systems are limited to individual base pairs or stacked nucleotides. The effect of the solvent, if included at all, can only be included implicitly (via a mean field or continuum treatment) or through the inclusion of a very small number of explicit waters around the molecule in the QM calculation. Prior to using these methods to investigate nucleic acid models, it would be wise to consult the detailed literature for similar examples. Some recent examples of QM treatments of nucleic acids include the investigation of base-pairing energetics, base stacking, the interaction of metal ions with base pairs, and even base solvation in simulations of individual bases surrounded by a small amount of explicit water (for review see Hobza and Sponer, 1999). Very few QM calculations have been applied directly in macromolecu-

Contributed by Thomas E. Cheatham, III, Bernard Brooks, and Peter A. Kollman
Current Protocols in Nucleic Acid Chemistry (2001) 7.8.1-7.8.20
Copyright © 2001 by John Wiley & Sons, Inc.

lar modeling applications. An exception to this is the critical role of QM methods in the development, evaluation, and critique of empirical potential functions (as are introduced later in this section); this probably represents the largest use of QM methods in nucleic acid modeling.

As mentioned in *UNIT 7.5*, although the QM treatments have the potential to provide a very high level of accuracy, this level of accuracy is not always required, and faster less-accurate techniques may be appropriate; however, the commonly applied approximations come at a cost. Without a QM treatment, it is generally not possible to accurately represent processes that involve chemical changes (i.e., chemical reaction, bond breaking, or bond forming), excited states (i.e., electronic transitions), or electron transfer. In practice this is not a major limitation, since with nucleic acids the investigator is most often simply interested in the structure, dynamics, and relative importance of a given model.

The middle ground between pure ab initio QM techniques and the faster empirical potentials (discussed in the next section; see Molecular Mechanics: Empirical Potential Energy Functions) are semi-empirical treatments. These apply a quantum mechanical formalism where significant approximations are applied in the calculation to decrease the computational cost, while accuracy lost through the approximations is offset by the addition of empirical parameters. The semi-empirical methods allow investigation of ~2500 atoms for geometry optimization or ~100 atoms for dynamics. Their use in biomolecular simulations has been limited in part due to the difficulty in properly representing hydrogen-bonded systems; however, the future holds promise with the development of semi-empirical parameterizations for biomolecular systems, better methods for treating large systems (Thiel, 1997), and semi-empirical codes designed to run on massively parallel computers (Dixon and Merz, 1997). Standard semi-empirical codes include MOPOC, MNDO4, AMSOL, Argus, and ZINDO (see Internet Resources) among others, and utilize various parameterizations including AM1, PM3, and MNDO. Reviews on semi-empirical methods discuss the codes and parameterizations in greater detail (Stewart, 1990; Zerner, 1991).

Despite the higher levels of accuracy possible with the quantum mechanical and semi-empirical treatments, the most commonly applied treatment of energy uses simplified molecular mechanical (MM) potential energy functions. Given an appropriate parameterization, these MM potential functions are able to accurately reproduce nucleic acid structure at the atomic level in a very computationally efficient manner. This allows rapid evaluation of not only energy, but forces (or the first derivative of the energies with respect to the positions), thereby allowing analysis of many configurations in a reasonable time frame.

Before leaving this discussion of ab initio and semi-empirical methods, it is important to mention that there has been significant progress in the development of hybrid methods. These treat part of the system quantum mechanically (the part undergoing chemical change) and the remainder with a significantly faster molecular mechanical (empirical) potential. The hybrid QM/MM treatments allow representation of larger systems (such as enzymes) with explicit representation of the environment, while still treating the region of chemical interest (such as the active site) quantum mechanically. Drawbacks of these methods are that they are extremely computationally demanding (depending in large part on the size of the QM region) and that there are a number of open research questions, such as how best to merge the QM and MM regions, how best to decide what part of the system should be represented quantum mechanically, and what level of treatment to apply in the QM region (i.e., ab initio versus specifically parameterized semi-empirical) or MM regions (Field et al., 1990; Stanton et al., 1995; Gao, 1996; Cummins and Gready, 1997; Chatfield et al., 1998). Although most of the QM/MM applications have been limited to enzyme systems or small molecules in solution, a recent application involved model

reactions for ribozymes (Lahiri and Nilsson, 1997). Code for performing QM/MM is in many of the standard molecular dynamics programs, including AMBER (Pearlman et al., 1995), CHARMM (Brooks et al., 1983), and GROMOS (van Gunsteren and Berendsen, 1987).

Molecular Mechanics: Empirical Potential Energy Functions

The most commonly applied methods for describing energy use an empirically derived or molecular mechanics (MM) potential function. This involves the application of a simplified potential function that has been parameterized to properly model the structures of interest. The specific parameterization, or force field, needs to represent not only the *intra*molecular interactions (based on the covalent structure of the molecule) but the *inter*molecular interactions between all the atoms and molecules. Most of the commonly applied empirical derived force fields describe the intermolecular interactions in a similar manner. The complete and true atomic energy representation, $U(r_1, r_2, \ldots, r_N)$, involves the interactions of all N atoms (at positions r_1 to r_N) in the system. Following Allen and Tildesley (1987), this energy representation can be decomposed into a sum of pairs, triples, quadruples, and higher interactions between atom centers:

$$U(r_1,r_2,\ldots,r_N) = \sum_{i=1}^{N} U(r_i) + \sum_{i=1}^{N}\sum_{j>i}^{N} U(r_i,r_j) + \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{k>j>i}^{N} U(r_i,r_j,r_k) + \ldots$$

Equation 7.8.1

Approximations are then applied to simplify the representation. Given that the individual interactions for the terms from the quadrupolar interaction and higher-order interactions are usually very small, these are often neglected. The first-order term, $U(r_i)$, involves interactions with external fields and, therefore, is also generally not included. This leaves the dominant terms, specifically the pair interactions (a function of the distance between the atoms; r_{ij}) and the three-body (atom-dipolar) interactions. The three-body interactions are less often included explicitly in biomolecular simulation at present due to the increased cost of calculation and the nonadditivity of the energy; however, these interactions are not completely neglected since their effect is *implicitly* included in the molecular mechanical force field during the explicit parameterization of the pairwise interactions. To summarize, the most commonly applied empirical force fields use an additive pairwise potential (with the nonadditivity and three-body terms omitted).

Intramolecular Interactions

The intramolecular interactions describe the covalent structure of the molecule. This includes the bonds, angles, dihedrals, and overall connectivity and flexibility of the model. A reasonable MM representation can be obtained either with full *atomic* or *internal coordinate* representations. Atomic representations imply that each atom center is allowed to move independently and explicit bond, angle, and dihedral terms are added to represent the connectivity. Two common representations are applied: all-atom force fields treat each atom as independent, whereas united-atom force fields fold multiple atomic centers into a single particle, such as treating a methylene group ($-CH_2-$) as a single center.

Internal coordinate representations, on the other hand, attempt to represent the inherent motions and flexibility of the system via rotations about naturally rigid groups of atoms. These can be useful since there are effectively fewer degrees of freedom, with flexibility only included where necessary to represent the natural motion of the molecule. For example, with nucleic acids the bases can often be treated as rigid units, as can most of the backbone (except for ε and ζ), with rotations and translations set up to represent deviations from a common helical axis system, such as rotations involving inclination,

tip and twist, and translations (e.g., *x*- and *y*-displacement from the helical axis and rise between base pairs). To include the flexibility of the sugar pucker, the bonds at the C4′-O4′ atoms can be broken, leading to two torsions and three angles to describe the ring. This internal coordinate representation is used in the program JUMNA (junction minimization of nucleic acids; Lavery et al., 1995). Alternative treatments may use different rotations (e.g., twist, roll, and tilt; Gorin et al., 1990), more complicated treatments (Nesterova et al., 1997), or could (as is done in protein simulations) consider the bonds and angles as fixed while allowing free rotation around all the torsions. Ideally, a good internal coordinate representation will attempt to minimize the number of degrees of freedom while still retaining good structure and dynamics. Reduction of the number of degrees of freedom is desirable since this has clear benefits when trying to sample the possible conformations (as will become more apparent later).

Although the internal coordinate treatments lead to fewer degrees of freedom, using this kind of representation in molecular dynamics simulations is difficult because accurate treatment requires inverting the moment of inertia tensor (or mass matrix) or the application of computationally demanding holonomic constraints (that maintain the fixed structure within rigid groups and properly equalize or propagate the forces). In all-atom or united-atom treatments, because each atom center is free to move, the mass matrix is diagonal and thus, trivially invertable. Even though there are approximations that effectively treat the moment of inertia tensor when it is not diagonal with order N rather than N^3, accuracy is lost. Based on these efficiency issues, internal coordinate treatments are generally only used with minimization or Monte Carlo simulation (as described more fully in the next section). A further potential difficulty with the internal coordinate representation is that it requires specification of the rigid units; if the unit is rigid, it cannot distort structurally, in contrast to what might be the expected behavior under certain conditions. Thus, care has to be taken not to rigidify a part of the molecule that may not be rigid in practice. Moreover, rigid rotation about a given bond effectively leads to higher rotational barriers since there is no coupling to other modes (i.e., the bonds or angles cannot open up to facilitate rotation). For example, the *gauche, gauche* rotational barrier for butane is roughly twice as large when the bonds and angles are held rigid. Of course, this artifact is not a problem in practice since the force fields are parameterized to compensate; this just points out that it is not possible to directly mix intramolecular force fields designed for use with internal coordinates with those designed for use in all-atom treatments.

The alternative to using an internal coordinate treatment is to not treat any of the internal coordinates as rigid, so that each atom is free to move. In this case, to represent the covalent structure and intramolecular energetics, explicit energetic representations for the bonds, angles, and dihedrals need to be added. In the simplest form, this is typically done with harmonic potentials to maintain bond lengths and angles, and Fourier terms for the torsion angles to represent rotation about bonds (Figure 7.8.1). Either harmonic or Fourier terms are also commonly added to maintain planarity or prevent rotation about double bonds. Higher-order terms can also be added as necessary for more detailed representation. A common form for the intramolecular part of the mechanical potential function is as follows:

$$U_{\text{intramolecular}} = \sum_{\text{bonds}} k_b(r - r_{eq})^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \sum_{\eta} \frac{V_\eta}{2}[1 + \cos(\eta\phi - \gamma)] + \ldots$$

Equation 7.8.2

The parameterization, or force field, refers to the specific equilibrium geometry values for the various terms (such the equilibrium bond length, r_{eq}) and the force constants representing the energy (or vibrational frequency, k_b) of distortion away from the equi-

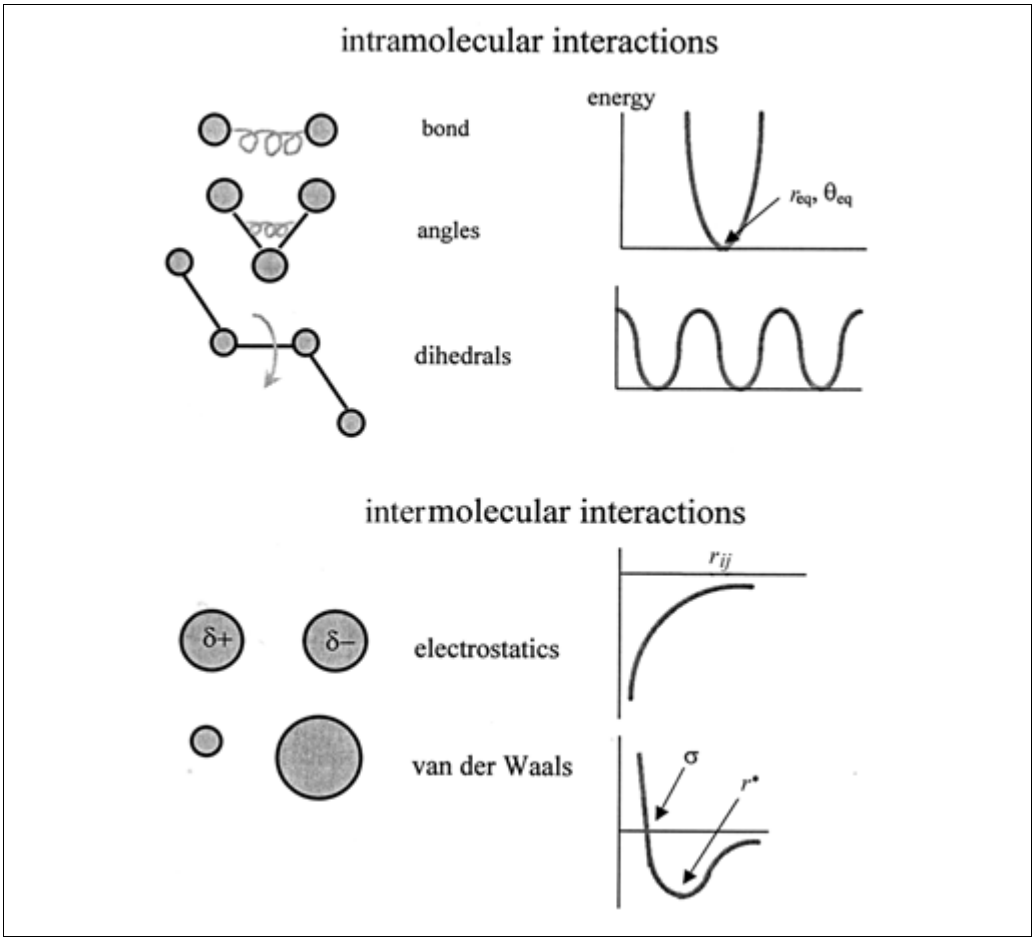

Figure 7.8.1 Schematic of the interactions in a pairwise additive molecular mechanics force field.

librium geometry. The parameterization is typically performed on molecular fragments with the implicit assumption that the force field parameters are transferable to other fragments. In other words, a carbon-carbon bond in propane is the same as a carbon-carbon bond in pentane. Bond lengths and bond angles are determined via experiment (crystallography or other spectroscopic techniques) with force constants for the vibration inferred from microwave, infrared (IR), and other spectroscopic data. In the parameterization, the least well-determined parameters (based on experimental information) are the dihedral terms, since these terms effectively include not only the equilibrium torsion value but 1-4 atom interactions and other implicit interactions explicitly omitted. Given this, the dihedral part of the force field is usually the last part to be parameterized (based on QM and empirical data) and modification of these parameters can be used to fix up deficiencies from the other intramolecular and intermolecular interactions.

More complex molecular mechanical representations are also possible, such as those including cubic and quartic terms for bond stretching, cubic angle terms for anharmonic bending, bond-torsion, angle-torsion, bend-bend, and other terms. Although these higher-order terms can aid in parameterization efforts to better represent vibrational spectra, structure, and heats of formation in a diverse set of molecules (including strained molecules), these are not typically included in biomolecular force fields. This is because biomolecular force fields are primarily parameterized to represent structure and secondarily relative conformational energetics in as simple and transferable a means as possible. An adequate representation is obtained without the added complexity, as will become apparent. For more information about the more strongly parameterized class 2 and class

3 force fields that include the more complicated intramolecular energy representation, see discussions of the following force fields: MMFF94 (Halgren, 1996), MM3 and MM4 (Allinger et al., 1989, 1996), and QMFF (Maple et al., 1994).

Intermolecular Interactions

The standard form for the intermolecular pair energy involves a Lennard-Jones potential to represent the electron cloud repulsion (r_{ij}^{-12}), dispersion attraction (r_{ij}^{-6}) interactions, and a Coulombic term (with atom point charges q_i and q_j, and dielectric constant ε) representing the electrostatic interactions between all the atom pairs (where r_{ij} is the distance between atoms i and j).

$$U_{\text{intermolecular}} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}$$

Equation 7.8.3

Note that the dielectric constant (ε) is shown in a simplified form that implicitly includes the $4\pi\varepsilon_0$ leading factor, where ε_0 is the permittivity of free space. Also note that self-interactions (1-1), interactions between bonded atoms (1-2), and interactions between the atoms involved in angles (1-3) are most often omitted, since their interaction has already been included in the intramolecular part of the force field. Interactions between terminal atoms involved in a dihedral angle (1-4 interactions) are sometimes scaled. Polarization effects are also often included implicitly in the parameterization, although explicit polarization (as is discussed later) can be included at additional cost. The specific parameters (A_{ij}, B_{ij}, q_i, q_j) in large part determine the reliability of the intermolecular potential. There are many philosophies regarding how to "best" derive these parameters, ranging from total reliance on high-level QM treatments of fragments to parameterization based entirely on empirical data, or some combination of each. A recent trend has been to derive the Lennard-Jones parameters for particular atom types from simulations of neat liquids (Jorgensen et al., 1996). Like with the bond, angle, and dihedral parameters, there is an implicit assumption that (in general) the Lennard-Jones parameters are transferable. It should be noted that the parameters A_{ij} and B_{ij} for the Lennard-Jones part of the equation above represent *mixed* van der Waal parameters for atoms i and j. Since the literature is sometimes confusing with presentation of the mixed parameters as A_{ij} and B_{ij} (with or without pre-exponentiation) or in terms of r^* (the minimum of the potential well in Å) or σ (the zero of the potential in Å), and since this is further complicated by application of different combining rules, it is worth a brief digression. The two forms of the van der Waals energy in terms of r^* and σ are as follows:

$$E_{\text{LJ}} = 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] = \varepsilon\left[\left(\frac{r^*}{r_{ij}}\right)^{12} - 2\left(\frac{r^*}{r_{ij}}\right)^{6}\right]$$

Equation 7.8.4

This implies that:

$$r^* = \sigma\sqrt[6]{2}$$

Equation 7.8.5

Given an r^* value (representing the van der Waal radius) for two atoms i and j that are not the same, it is necessary to define a mixed van der Waal radius. There are two methods in common usage, arithmetic or Lorentz-Bertholet combining rules where:

$$r_{ij}^{*} = \frac{r_i^{*} + r_j^{*}}{2}$$

Equation 7.8.6

(as applied in AMBER and CHARMM) and geometric mean combining rules where:

$$r_{ij}^{*} = \sqrt{r_i^{*} r_j^{*}}$$

Equation 7.8.7

(as applied in GROMOS). In both cases, a geometric mean is used for the well depth. These differences again point to the critical need to be careful when trying to adapt force fields applied in one program for use in another or, alternatively, in mixing parameters from different force fields.

Other forms for the intermolecular potential can also be employed to represent the intermolecular interactions, such as replacing the repulsive (12) potential by an exponential, as in MM2 (Allinger, 1977), or replacing the Lennard-Jones potential by a buffered 7-14 potential for the van der Waal interactions and shifting the electrostatics to prevent infinite attractive electrostatics from dominating the finite van der Waal interactions at short range, as employed in MMFF94 (Halgren, 1996). Although these more complicated functional forms arguably work much better for the varied small molecules of interest to pharmaceutical companies, these force fields have not seen significant use in biomolecular simulation applied to nucleic acids.

For the electrostatic interactions, most MM methods assume fixed atomic point charges. Polarization is only included implicitly via the construction of point charge values that lead to effectively larger dipoles. The point charges are one part of the MM model that is not very transferable since a given atomic charge depends critically on its environment; therefore, charges are typically calculated for individual fragments (such as each individual nucleic acid or amino acid) rather than for specific atom types. With the lack of explicit polarization, a major limitation of the additive pairwise force fields is in the treatment of transition metals or multivalent ions (which may not be treated properly with standard additive empirical force fields); in this case, accurate treatment may require the inclusion of explicit polarization effects or even some QM treatment.

Although polarization (induced dipole) effects have often been neglected to date, these effects can be included in the MM potential energy function. This can be done by including inducible dipoles on each center (μ_i) reacting to the electrostatic field (E_i^0) on center i arising from all other fixed charges and representing the polarization energy (U_{pol}) as follows:

$$U_{\text{pol}} = -\frac{1}{2} \sum_i \mu_i E_i^0$$

Equation 7.8.8

This is evaluated self-consistently, solving for the inducible dipole based on the polarizability of atom i (α_i) and the total electrostatic field (E_i) at the polarizable center, where $\mu_i = \alpha_i \times E_i$, noting that the total electrostatic field:

$$E_i = E_i^0 + \sum_{i \neq j} T_{ij} \mu_j$$

Equation 7.8.9

and T_{ij} is the dipole tensor:

$$T_{ij} = \frac{1}{r_{ij}^3}\left(3\vec{r}_{ij}\frac{\vec{r}_{ij}}{r_{ij}^2} - 1\right)$$

Equation 7.8.10

Additional terms can be added to represent three-body exchange repulsions (Caldwell et al., 1990). Alternative treatments of polarization include the fluctuating charge model that treats the charges as dynamic variables in the simulation (Rick et al., 1994). In the published literature, the inclusion of explicit polarization effects in molecular mechanics treatments has generally only included simulations of small polarizable molecules or ions in polarizable water. Polarization effects were added to better represent molecule association or free energies of solvation. A fully polarizable treatment has not been applied to large biomolecules in solution, in part because adding explicit polarization tremendously increases the cost of the calculations, but also because a consistent polarizable force field for nucleic acids or proteins has not been developed. This will likely change in the near future due to the availability of faster and better methods to include explicit polarization, which in turn will lead to the development of more reliable nucleic acid force fields that are parameterized for use with explicit polarization.

The Total Energy

Added together, the two energy terms ($U_{intra} + U_{inter}$) represent the total potential energy of the system. An important, but often overlooked, point is that it is often meaningless—despite the use of common units for energy such as kcal/mol—to compare absolute molecular mechanical energies between different molecules (or force fields) due to the lack of a common zero point energy. Moreover, many of the commonly employed force fields were not parameterized to very accurately estimate heats of formation, so even with a complete specification of a reaction linking two molecules together, it is likely that the relative energy between two different molecules will not be accurate. Despite this warning, it is possible to compare the relative energies of different conformations of the *same* molecule (under the same conditions, e.g., same force field, same number of explicit waters), remembering that this energy difference is not a free energy but only a relative energy or enthalpy. Although low energy structures are likely more representative, it is important to remember that this is not always the case (at normal temperatures) due to possibly large differences in entropy. For the simulation of nucleic acid systems, the most reasonable representation of the structure comes from the force fields specifically parameterized to represent nucleic acids. Current all-atom force fields for nucleic acids that perform reasonably well include the Cornell et al. (1995) and recent variants (Cheatham et al., 1999; Wang et al., 2000), BMS (Langley, 1998), and CHARMM all-27 nucleic acid (Foloppe and MacKerell, 2000) force fields. In addition to all-atom force fields, the JUMNA internal coordinate force field and others previously mentioned also perform reasonably well.

BEYOND ENERGY EVALUATION

Evaluating the energy of a given model does not suggest anything about the relative stability or appropriateness of that model; however, differences in the relative energy between two conformations of the same model structure can suggest which structure is more enthalpically reasonable. Coupled with some representation of the relative entropy, this can give insight into the relative importance of each conformation. In general, however, it is not easy to estimate the entropy for a given conformation. Although there are some approximate methods for estimating relative conformational free energies (Kollman et al., 2000), these require knowing a priori what the representative structures are.

Ideally, the investigator would like to know the conformation that represents the lowest energy (enthalpy) and ultimately the lowest free energy structures, since these structures are more representative of the true conformation of the molecule. A simple way in principle to find low-energy structures is to find the set of coordinates that minimizes the potential energy. In practice, this is limited due to the complexity of the potential energy hypersurface and the large number of degrees of freedom. Without exhaustive sampling of all possible conformations of a molecule, in general, it is impossible to determine if a given low-energy structure represents the true "global" minimum structure or whether it is simply a "local" minimum of the potential energy (where minimum energy structure means a structure that is at the bottom of one of many possible wells in the energy representation). Without knowing what the global minimum is, it is impossible to determine if a given model structure is at all representative of what might be expected at room temperature. Even with a knowledge of the global minimum energy structure, a molecule may have a number of other low-energy structures nearby that may be populated at room temperature.

To find representative conformations, it may be desirable to apply methods that sample according to the expected probability of observing a given conformation (at a given temperature). Examples of methods that do this are molecular dynamics (MD) or Monte Carlo (MC) methods. This is sampling according to the Boltzmann distribution and, in the limit of infinite sampling, this gives a complete representation of all the possible conformations and their relative probabilities of occurrence, p_i, where ε_i is the total energy of the i^{th} state, T is the temperature, and k_B is the Boltzmann constant.

$$p_i = \frac{e^{-\varepsilon_i / k_B T}}{\int e^{-\varepsilon_i / k_B T}}$$

Equation 7.8.11

The term in the denominator is the partition function and specifies an integral over all phase space or the complete set of possible coordinates and momenta. Of course, in practice, sampling is limited and the sheer complexity of the accessible conformational space may prevent finding all the low-energy conformations and, therefore, full determination of the partition function. The complexity of the potential energy hypersurface, and the exponential explosion of the number of possible conformations as the number of degrees of freedom increases, limit exhaustive search of this space to systems that possess only a few degrees of freedom. In 1990, none of a variety of systematic and random conformational search methods in both internal (torsion) and all-atom coordinate frames was independently able to find all the relevant low-energy conformers (within 3 kcal/mol) of the cyclically constrained cycloheptadecane molecule, which formally has 147 degrees of freedom (all atom; Saunders et al., 1990). Although computer power has increased tremendously in this time, order-of-magnitude computational speed advances do not significantly improve the effective conformational sampling due to the exponential explosion.

A simple way to estimate the effective complexity is to assume that the number of possible minima or low-energy conformations relates to the simplistic set of three low-energy rotations about a given single bond (i.e., the *trans*, *gauche*$^+$, and *gauche*$^-$ states) or $3^n - 1$, where n is the number of rotatable bonds. Although this is typically less than the total number of degrees of freedom, it is still large! With fewer degrees of freedom, there are less minima and sampling is easier. Moreover, minimization algorithms are less likely to become stuck in less representative local minima; however, the complexity quickly grows; even with the internal coordinate treatments for nucleic acids described, there are still ~20 degrees of freedom per nucleotide (representing hundreds of possible conformations

for a given nucleotide assuming three low-energy states for each of approximately five rotatable bonds). This places exhaustive sampling out of reach for any system with more than a few nucleotides. This difficulty in finding the global minimum (or set of coordinates that leads to the lowest energy) is often termed the multiple minima problem. In the context of the effective amount of sampling attainable during MD simulation, this problem is often termed the conformational sampling problem and relates to the improbability of overcoming large energy barriers.

The multiple minima or conformational sampling problem is why it is desirable to choose reasonable initial structures (i.e., experimentally derived structures or valid model structures) when modeling. With reasonable structures, likely MD or MC (see below) simulation will sample reasonably well near the initial structure; however, large conformational changes, such as B-DNA to Z-DNA transitions or RNA folding will not likely be seen in a reasonable time. Of course, these limits in sampling imply that even with unreasonable structures, such as the imaginary and perhaps metastable B-RNA structure, MD or MC will likely sample reasonably well near the initial model structure. This is indeed the case, as shown in state-of-the-art all-atom solvated MD simulations where B-RNA is stable for >10 nsec and B-RNA to A-RNA transitions are not observed unless artificial means are applied to force the conformational transition (Cheatham and Kollman, 1997). In the next sections, minimization, Monte Carlo, and molecular dynamics methods will be discussed in more detail; more detailed treatments can be found elsewhere (Valleau and Whittington, 1977; Allen and Tildesley, 1987; McCammon and Harvey, 1987; Brooks et al., 1988; van Gunsteren and Berendsen, 1990; Leach, 1997.)

Minimization

Minimizing the potential energy corresponds to instantaneously "freezing" the system or dropping to the bottom of the nearest potential energy well (as shown schematically in Figure 7.8.2A). Minimization is a standard optimization problem that can be approached using tools of various complexity ranging from simple and inefficient zero order methods, such as grid search, which only require evaluation of the energy for a particular conformation, to more efficient but complicated n^{th}-order methods, which use information about all the derivatives of the energy function up to the n^{th} order. Thus, to perform minimization with an n^{th}-order method requires analytic derivatives of the potential energy function up through the $(n-1)^{th}$ derivative (since the n^{th} can be approximated by finite difference methods). In practice, a variety of first- and second-order methods are typically applied since these provide a reasonable balance between functionality (i.e., finding a minimum) and efficiency. Finding the set of coordinates that minimizes the potential energy does not guarantee that the conformation found represents the lowest energy structure due to the presence of multiple minima. As discussed, the molecule can get trapped into a local minimum, and this local minimum may not be representative of the "true" conformation.

In spite of the local minimum problem, minimization is still a useful tool for modeling, since after making a particular modification to a given model, say the replacement of the phosphodiester backbone of DNA by a poly-amide (PNA) backbone, the conformation can be minimized. This will remove gross steric overlap and relax strained bonds and angles. While this will not say much about the relative stability of this backbone modification, it can suggest whether the backbone replacement is at least sterically feasible for the given nucleic acid conformation. Coupled with a good chemical intuition, this may provide sufficient information to the modeler to suggest whether this backbone modification is potentially useful or not. For example, consider related backbone modifications to PNA that include one more or one less methyl group along the backbone. Simple minimization may suggest that the shorter backbone will strain the nucleic acid, leading to a lower rise between base pairs, and the longer backbone will increase the

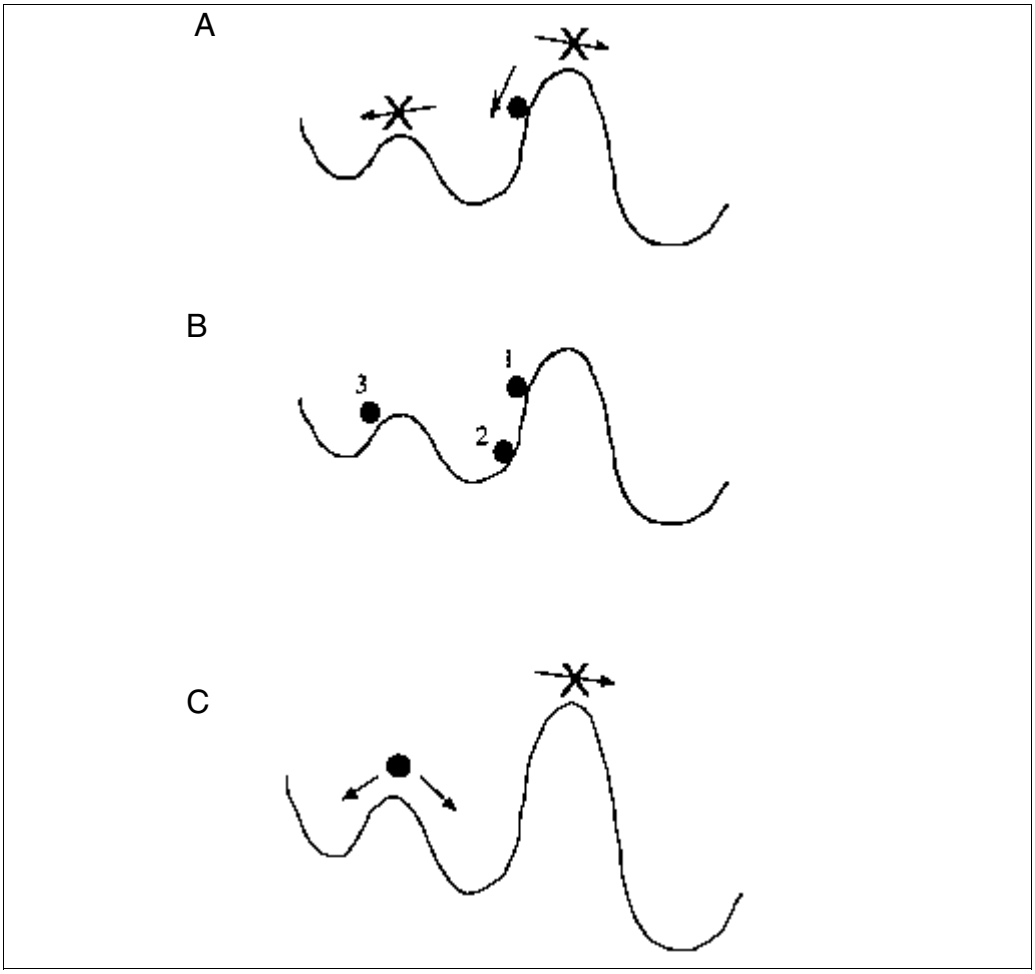

Figure 7.8.2 Schematic representations of the sampling of various methods. These plots represent the energy of the system along an arbitrary reaction coordinate. The wells represent energy minima in the phase space. The state of the system is depicted by the location of the ball. (**A**) Minimization. The system moves to the bottom of the nearest well and barriers are not overcome. (**B**) Monte Carlo. Each configuration of the system is represented by a number and barrier crossing relates to the move set and total number of moves. (**C**) Molecular dynamics. The state of the system evolves due to force according to Newton's equations of motion. In short simulations, large barriers will not be surmounted.

separation between base pairs, which might suggest that these backbone modifications are less reasonable. While some insight can be gained, it is important to remember that entropic effects are not included (such as the likely greater configurational entropy loss on binding for the larger PNA) and that the conformation sampled by the minimization is only a local minimum. In addition to being a useful tool for modeling, minimization is necessary prior to running molecular dynamics simulations to relieve any large steric overlap or remove strained bonds or angles that might otherwise lead to initially large forces. During the dynamics, the large forces lead to large displacements, which in turn may lead to more overlaps and more large displacements; this cascade of events can lead to local hot spots or failure of the integrator during MD.

There are two common first-order minimization methods in common usage: the steepest descent (Wiberg, 1965) and conjugate gradient (Fletcher and Reeves, 1964) methods. These use the first derivative to give information about the slope of the potential (but not the curvature). In steepest descent, movement is made parallel to the net force. Given reasonable step sizes, this method will not cross barriers and will readily traverse down

to the bottom of the nearest potential energy well. Although this method is not very efficient, it is very stable and is therefore often used to initially minimize the structure when there are large energies. Often this is used as the first step in modeling, particularly when there are potentially large van der Waals overlaps to initially relax the most drastic energetic penalties. The conjugate gradient method improves upon the steepest descent by using the gradients from previous steps to further guide the minimization. This method is more efficient than steepest descent and is appropriate to apply after initial minimization (e.g., to remove the largest steric overlaps). Near the bottom of the harmonic well, or after some initial minimization, use of second order methods (which assume an approximately quadratic relationship to the energy) such as Newton-Raphson (TNPACK; Schlick and Overton, 1987) can be used to speed up convergence at the expense of greater memory usage and the need for second derivatives (either by finite difference or analytical). These methods are typically more expensive since they require inverting the second derivative matrix. The expense can be significantly reduced by limiting the space sampled to regions where there is significant movement in the energy, thus limiting the size of the second derivative matrix (which is calculated by finite difference; Brooks et al., 1983).

The key point is that minimization is a very useful tool, but that care should be taken when analyzing the validity of a particular "low-energy" structure. Because of the limits in sampling and the great likelihood of minimizing to a local minimum (which may or may not be representative of the favored low-energy structures), the validity needs to be judged based on the chemical intuition of the modeler and known reliability of the initial model. To better sample space, Monte Carlo or molecular dynamics methods can be applied; although these methods avoid the problem of getting trapped in local minima, limits in sampling only allow partial sampling of conformations "near" the initial structure. Despite the limited sampling, minimization is an extremely useful tool to characterize model structures.

Note that care should be taken when applying restraints along with minimization, such as when using NMR-derived distance restraints, to balance the restraint force constants with those of the force field. If the restraint force constants are too large relative to the force field, unrealistic distortion of the structure may result. As a final practical comment about minimization, it is important to initially use small step sizes during the minimization of high-energy structures. This prevents the minimizer from jumping out of the current well to a potentially higher energy surface, which with repeated large jumps can lead to instability. Moreover, if the step size is too large, the minimizer may effectively get "trapped" in a cycle, jumping back and forth across the bottom of the well and preventing the minimizer from converging.

Monte Carlo

The Metropolis Monte Carlo method is essentially an algorithm for generating a random walk in conformational space such that the conformations obtained are distributed according to the probabilities expected for the equilibrium Boltzmann distribution (Metropolis et al., 1953; Valleau and Whittington, 1977). The algorithm is very simple and requires a single evaluation of the potential energy (and no force calculation) for each step (Figure 7.8.2B). At each step, a random movement is made. In an all-atom representation, this may be the movement of a single particle; with an internal coordinate representation, it is a movement along the normal coordinates (such as a rotation of one of the rigid units). The potential energy difference (ΔE) between the new conformation and the old is calculated. If the energy of the new conformation is lower, the move is accepted. If not, the move is accepted if $e^{-\Delta E/k_B T}$ is less than a random number drawn from the interval [0,1]; otherwise the old conformation is retained. The length of the simulation

relates to the number of attempted moves or configurations. The length of the simulation does not relate to the time scale and only represents the number of configurations sampled.

Use of this procedure leads to a representative set of structures. With sufficient sampling, this will converge to the equilibrium distribution and in principle allow reasonable estimation of any ensemble average, assuming the forces between particles are velocity independent (Metropolis et al., 1953). This assumption allows one to separate out the momenta or kinetic energy from the total energy such that only the potential energy is evaluated in the above expression; in practice, the kinetic component is always ignored. Since uphill moves are only accepted randomly (and not according to any deterministic process), there is no implicit time evolution and the progression of the sampling gives no information about the dynamics. The success of this procedure (with finite sampling) largely relates to the "move set", or set of possible moves. When an all-atom treatment is applied with single particle moves, the stiff internal degrees of freedom can lead to large energy changes and therefore a small step size needs to be used. With small step sizes, many more configurations may need to be evaluated to generate an appropriate ensemble, decreasing the efficiency. Moreover, for systems with disparate frequencies in different degrees of freedom—i.e., both stiff internal degrees of freedom (bonds, angles) and softer modes (correlated movements, dihedral rotation)—such as nucleic acids or single particle or atom moves, this leads to poor sampling. Even for liquid simulation, small atom-based moves lead to low acceptance ratios and poor sampling. In liquid simulation this can be overcome by using moves along normal coordinates, such as bond rotations, coupled with rotations and translations of the entire molecule. In general, it is desirable to avoid move sets that overly reject moves, and a roughly 40% acceptance rate represents a reasonable balance (Jorgensen and Tirado-Rives, 1996).

The difficulty in choosing a proper move set led to the early impression that MD simulations were up to ten times more efficient than MC at generating conformations of the small protein BPTI (Northrup and McCammon, 1980); however, this simulation used an extremely inefficient move set based on movement of atomic centers. Much better behavior is seen with a move set based on internal coordinate rotations about bonds (with rigid bond length and angles; Noguti and Go, 1985). For liquid simulation, MC is likely the most efficient method for generating reasonable ensembles; an excellent MC program for liquid simulation is BOSS developed by the Jorgensen group (Jorgensen, 1995). Use of the MC procedure for the simulation of neat liquids led to the development of very reliable van der Waals parameters for the OPLS force field (Jorgensen et al., 1996). A direct comparison of MC and MD simulation of liquid hexane suggests that MC is roughly three times more efficient than MD when an appropriate move set for the MC is applied (Jorgensen and Tirado-Rives, 1996).

Although MC is more efficient for liquid simulations, it is not clear this generalization follows to the simulation of biomolecules, particularly in solution. For example, with internal coordinate moves with flexible long-chain molecules (such as nucleic acids), a small rotation about a central dihedral angle can lead to a large displacement of the end of the chain. In explicit water, this might lead to extreme van der Waals overlap and likely very high rejection rates (for these types of moves). Although the use of correlated moves (such as crankshaft rotations about two bonds) can counter this effect (Dodd et al., 1993; Deem and Bader, 1996), there has been little published use of MC methods in all-atom simulation of nucleic acids in explicit solvent. For simulations without explicit solvent, MC simulation is widely used for nucleic acid simulation particularly with internal coordinate representations. Examples include its use in structure predictions as previously mentioned (Erie et al., 1993), investigating DNA bending in polyadenine tracts (Zhurkin et al., 1991), and investigating counterion distribution about DNA (Young et al., 1997).

Molecular Dynamics

Molecular dynamics simulation refers to integrating numerically the classical equations of motion (Newton's equations) for all the atoms in the system (Figure 7.8.2C). A simulation is started by assigning random momenta (velocities, v_i) for each of the N particles (of mass m_i) from a Boltzmann distribution about a given temperature, T.

$$\frac{1}{2}\sum_{i}^{N} m_i v_i^2 = \frac{3Nk_B T}{2}$$

Equation 7.8.12

Then the dynamics are propagated by integrating Newton's equations of motion, which for the pairwise potential, U_i, and its first derivative:

$$\nabla_i \sum_{j}^{N} U_{ij}$$

Equation 7.8.13

is represented by the following:

$$m_i \frac{d^2 r_i}{dt^2} = -\nabla_i \sum_{j}^{N} U_{ij}$$

Equation 7.8.14

The integration is typically performed through the use of one of a variety of first-order integration algorithms, such as leap-frog or velocity Verlet (Allen and Tildesley, 1987). This requires the calculation of the forces at each step, so a typical MD step involves calculation of the energy, forces, and velocities, and integration to obtain the coordinates for the next step. An important assumption is the ergodic hypothesis that, in the limit of complete sampling, the time average (that obtained by molecular dynamics) is equivalent to the ensemble average. Given sufficient sampling, it appears that the ergodic hypothesis is valid in practice since both MD and MC lead to equivalent ensembles for liquid hexane even though they use rather different sampling mechanisms (Jorgensen and Tirado-Rives, 1996). Also note that, unlike MC, the MD sampling is based on a dynamic propagation of the molecular mechanical forces; therefore, it is important that the forces are analytical or exact derivatives of the energy (i.e., the forces and energies should match). This is not always true in practice since the force calculation in some cases may be too expensive (particularly for nonpairwise forces), requiring some approximations. This means that the sampling is done on a surface different than the molecular mechanical energy surface.

Issues in MD simulation include the need for stable, reversible, and ideally symplectic (which implies loosely that the algorithm conserves energy and momentum) integrators. The simple first-order Verlet, velocity Verlet, and leap-frog integration algorithms satisfy these conditions in proper usage, in contrast to the more complex higher-order integrators such as the Gear predictor. Stability of the integration directly relates to the integration time step that in turn relates to the expected frequency of motion. A simple rule of thumb for Verlet integrators is that the integration catastrophe, or the time step where the integrator blows up, is the period of the highest frequency motion divided by π. For an all-atom simulation, the highest-frequency motions involve bond stretching of bonds to hydrogen. In the absence of rigid bond lengths (or constrained bonds), time steps are limited to the <1 fsec range. With SHAKE (Ryckaert et al., 1977) applied to constrain the lengths of bonds involving hydrogen, time steps in the 2 fsec range are routinely applied. Larger time steps are possible by limiting the high-frequency motion, such as through the

use of rigid units, which then necessitates inverting the inertia tensor as with internal coordinate treatments or the need for imposition of iteratively solved holonomic constraints. Effectively larger time steps are also possible through the application of multiple time step methods, which treat the slowly varying forces (which ideally represent the more computationally demanding part of the energy and force evaluation) with a longer time step, such as the long-range pairwise forces (Tuckerman et al., 1992; Biesiadecki and Skeel, 1993). With increased masses on hydrogen atoms to limit high-frequency motion, time steps as large as 5 fsec have been applied to systems with explicit water.

The symplectic nature of the integrator relates to preserving the Hamiltonian (or loosely the energy representation) of the system during integration. Not only should this be a property of the integrator, but molecular dynamics simulation in general should conserve energy. This is an excellent test of the methods. Note that in common usage, MD programs do not always necessarily conserve energy. This is often due to SHAKE tolerances that are not stringent enough, integration time steps that are too large, the use of the weak-coupling algorithm for constant pressure, and neglect of pair interactions. In order to speed the calculation, the effective number of pair interactions is often reduced to only those within a given range and a list of in-range pair interactions is maintained for each atom. For speed this "pair list" is not updated every step. Unless a buffer is maintained to not omit (include) pair interactions moving into (out of) range which is conservatively or heuristically updated, energy conservation may not be maintained. How the pair interactions are limited to finite range can also have important consequences, as can temperature or pressure coupling for constant T,P ensembles. In order to deterministically integrate Newton's equations of motion, it is important that systematic force errors be avoided (such as can occur with lack of energy conservation and temperature scaling) since these can lead to artifactual behavior, such as violation of equipartition (Harvey et al., 1998; Chiu et al., 2000). For example, with energy drains and the application of temperature adjustment by uniform scaling of the velocities of all the atoms, energy accumulates in low-frequency modes. This can lead to a growth of the center of mass translation. Given this, and since random velocity assignment likely leads to nonzero center of mass translational motion, it is advisable to remove this motion at the beginning of an MD run. Random force errors, such as those resulting from the slow accumulation of errors due to finite numerical precision, do not seem to lead to artifacts since they are equally likely to add as subtract from the total energy. To this end, differences obtained between sequential and parallel MD runs due to differences in the order of operations do not lead to significant differences and simply manifest the inherently chaotic nature of the integration (Braxenthaler et al., 1997).

Variants of standard MD include Langevin and Brownian dynamics. These MD methods are used to implicitly include the diffusive effects of solvent. In practice, this involves adding to the standard forces stochastic forces from a heat bath (via random fluctuation of forces) and dissipative forces to balance them. For in vacuo simulations, this represents thermal collisions with other molecules and allows coupling of energy among the different internal degrees of freedom. This can be very useful since, in the absence of Langevin forces, during deterministic dynamics, memory of the initial conditions may persist in the form of various correlated motions, and some low-frequency correlated motions may not be able to couple back into other modes of motion (since there are no collisions with other molecules). This can lead to poor sampling. In the limit of very high friction, Langevin dynamics become Brownian dynamics. Brownian dynamics are purely diffusive and effectively add a random coordinate displacement. This method is typically used to model the diffusional motion of molecules (such as the encounter of a substrate into an enzyme binding site). More detailed discussion of MD and its variants can be found in the previously cited books and a very useful published discussion of stochastic dynamics methods (Pastor, 1994).

With molecular dynamics, each particle has a finite kinetic energy; a direct implication is that the rate of barrier crossing will be proportional to the temperature and the length of the simulation. This implies that the probability of crossing large energy barriers during MD is rather small (as is represented schematically in Figure 7.8.2C). A back-of-the-envelope estimation of the rate of barrier crossing can be obtained from transition state theory with some basic approximations.

$$k \cong \kappa\nu \times e^{-G^{\ddagger}_{act}/RT}$$

Equation 7.8.15

The rate k is related to the transmission coefficient κ, which represents the ratio of successful transitions over the barrier, an effective rate at the top the barrier or equivalently a factor loosely representing collisions with the barrier, ν, times a Boltzmann-weighted "free energy of activation," which represents the height of the barrier. For simplicity, it can be assumed that at the top of the barrier, the particle always crosses the barrier rather than reflecting back, leading to $\kappa = 1$ (in contrast to expected values of 0.4 to 1.0 in solution). By classical equipartition ($E \cong k_B T$) and from $E = h\nu$ this barrier rate is:

$$\nu = \frac{k_B T}{h}$$

Equation 7.8.16

where T is the temperature, h is Plank's constant, and k_B is the Boltzmann constant. At room temperature, RT is ~0.6 kcal/mol and ν is ~6.2 psec^{-1}. Remembering that this is only an approximation, this suggests that barriers of ~1 kcal/mol can be surmounted in picoseconds (~1.2 psec^{-1}) and ~5 kcal/mol in nanoseconds (~1.5 nsec^{-1}), but that barriers >10 kcal/mol may take microseconds or longer. This relates to the conformational sampling problem and is a significant limitation of MD. In MC simulations, this problem can be overcome by increasing the size of the moves (at the expense of lower acceptance ratios) or by designing clever move sets. In a similar manner, with MD simulation various methods can be applied to effectively lower barriers to conformational transition, such as adding biasing potentials to lower specific torsion barriers, increasing masses, increasing temperature, smoothing the overall potential surface, or applying mean field approximations, such as with the locally enhanced sampling (LES) methodology. The LES methodology shows much promise in biomolecular simulation, such as by reducing the barriers to conformational transition and thereby allowing transition from an "incorrect" to correct RNA hairpin loop conformation (Simmerling et al., 1998). For review of enhanced sampling methods, see Straub (1996).

The attainable MD simulation time scale relates to the complexity of the potential (and therefore the time required to evaluate the energy and forces) and the integration time step. With longer time steps, longer simulations are possible. Through the deterministic procedure and specification of an integration time step, there is a direct relationship between the number of MD steps and the effective time scale (time step × the number of steps = total time). The current state of the art for MD simulation of nucleic acids in explicit solvent involves simulations of <50 nsec. Although in principle this time scale is "accurately" represented by the dynamics or integration of the classical equations of motion, whether a 1 nsec simulation of DNA actually accurately represents 1 nsec of real motion depends on the empirical potential employed. For equilibrium or ensemble properties (such as the energy, free energy, heat capacity, or density) this exact time scale is not important, assuming independence of the forces on the velocity of a given particle. This is implicitly the case with the standard molecular mechanics potentials. For ensemble properties, what matters is the effective amount of sampling. In principle, one can obtain equivalent ensemble averages even if the masses on all the hydrogens are increased. This

will allow a larger time step and therefore a longer sampling. Although this does not affect the ensemble properties, this will drastically affect time-averaged properties, such as water diffusion or the rate of specific conformational transitions.

The point is that these dynamic properties are very sensitive to the potential (and atomic masses). Since the force field is primarily parameterized to represent structure, it is not clear how accurate the dynamic properties are. Clearly in the absence of viscous damping forces or explicit solvent, the rate of dynamics may be enhanced relative to simulations in explicit solvent. Similarly, under high pressure conditions, such as in an isolated solvent droplet, the dynamics may be reduced. In explicit solvent, MD simulations of proteins and nucleic acids display thermal parameters that are in good accord with experiment and in general properly represent fast (picosecond) time scale motions. Whether these methods can accurately represent longer time scale motions is unclear since realistic MD simulation is currently limited to the nanosecond time scale. Evidence from the simulations is that some properties are well represented; however, there is a clear overestimation of harmonic motion in solvated proteins at low temperature (i.e., suggesting motion that is too slow; Steinbach and Brooks, 1994). On the other hand, one of the most commonly used water models, TIP3P (Jorgensen et al., 1983) diffuses at twice the experimental value. The point of this discussion of time scale and dynamics is not to criticize the methods but to remind potential modelers of the various issues and sensitivity not only to the force field but the representation. Although in general the results are within range, it is important to understand how the methods have been validated by comparison to experiment for a given property before making firm claims about exact details of the time scale for conformational transition. For example, estimating the free energy of binding for a particular water to DNA based on the lifetime of its bound state is likely to be inaccurate and very dependent on the specific water model. Additionally, in order to make claims about the time scale for a given process based on MD, the time scale of the simulation should be significantly longer than the relaxation time of the process of interest. Likely the simulation should be at least an order, if not two orders, of magnitude longer than the relaxation time and, when attempting to make statistically valid claims about the rate of a given transition or specific correlation, many events need to be observed. These and related issues regarding the validation of simulation results are presented in a recent review (van Gunsteren and Mark, 1998).

SUMMARY

In this unit, the basic principles of the common energy representations for nucleic acid models have been presented, and methods for exploring these energy surfaces have been discussed in some detail. To move beyond simple energy evaluation of a single model structure, it is fairly clear that a large number of energy (and possibly force) evaluations will be necessary. This becomes computationally demanding, and it is desirable to limit the computational cost as much as possible without sacrificing accuracy. For systems of reasonable size, there is a need to use empirical potentials as discussed; however, even with the simple pairwise potentials, the number of pair interactions quickly grows as the number of atoms is increased. Thus, for larger systems the authors limit the effective number of intermolecular interactions by limiting the range of the interaction (e.g., by limiting interactions to distances less than some cutoff or utilizing hierarchical approaches that coalesce groups of distant atoms into an approximate single effective particle). Additional complications relate to the representation of nucleic acids since water and associated counterion-ions (salt) are an integral part of nucleic acid structure (see *UNIT 7.9*).

LITERATURE CITED

Allen, M.P. and Tildesley, D.J. 1987. Computer Simulation of Liquids. Oxford University Press, Oxford.

Allinger, N.L. 1977. Conformational analysis 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* 99:8127-8134.

Allinger, N.L., Yuh, Y.H., and Lii, J.H. 1989. Molecular mechanics. The MM3 force field for hydrocarbons.1. *J. Am. Chem. Soc.* 111:8551-8566.

Allinger, N.L., Chen, K., and Lii, J.H. 1996. An improved force field (MM4) for saturated hydrocarbons. *J. Comp. Chem*. 17:642-668.

Biesiadecki, J.J. and Skeel, R.D. 1993. Dangers of multiple time step methods. *J. Comp. Phys.* 109:318-328.

Braxenthaler, M., Unger, R., Auerbach, J., Given, A., and Moult, J. 1997. Chaos in protein dynamics. *Proteins* 29:417-425.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D., Swaminathan, J.S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* 4:187-217.

Brooks, C.L., III, Karplus, M., and Pettitt, B.M. 1988. Proteins. A Theoretical Perspective of Dynamics, Structure, and Thermodynamics. John Wiley & Sons, New York.

Caldwell, J.W., Dang, L.X., and Kollman, P.A. 1990. Implementation of nonadditive intermolecular potentials by use of molecular dynamics: Development of a water-water potential and water-ion cluster interactions. *J. Am. Chem. Soc*. 112:9144-9147.

Chatfield, D.C., Eurenius, K.P., and Brooks, B.R. 1998. HIV-1 protease cleavage mechanism: A theoretical investigation based on classical MD simulation and reaction path calculations using a hybrid QM/MM potential. *Theochem. J. Mol. Struct.* 423:79-92.

Cheatham, T.E. III. and Kollman, P.A. 1997. Molecular dynamics simulations highlight the structural differences in DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes. *J. Am. Chem. Soc.* 119:4805-4825.

Cheatham, T.E. III, Cieplak, P., and Kollman, P.A. 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* 16:845-862.

Chiu, S.W., Clark, M., Subramaniam, S., and Jakobsson, E. 2000. Collective motion artifacts arising in long-duration molecular dynamics simulations. *J. Comp. Chem.* 21:121-131.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179-5197.

Cummins, P.L. and Gready, J.E. 1997. Coupled semiempirical molecular orbital and molecular mechanics model (QM/MM) for organic molecules in aqueous solution. *J. Comp. Chem.* 18:1496-1512.

Deem, M.W. and Bader, J.S.. 1996. A configurational bias Monte-Carlo method for linear and cyclic peptides. *Mol. Phys.* 87:1245-1260.

Dixon, S.L. and Merz, K.M.J. 1997. Fast, accurate semiempirical molecular orbital calculations for macromolecules. *J. Chem. Phys.* 107:879-893.

Dodd, L.R., Boone, T.D., and Theodorou, D.N. 1993. A concerted rotation algorithm for atomistic Monte-Carlo simulation of polymer melts and glasses. *Mol. Phys.* 78:961-996.

Erie, D.A., Breslauer, K.J., and Olson, W.K. 1993. A Monte Carlo method for generating structures of short single-stranded DNA sequences. *Biopolymers* 33:75-105.

Field, M., Bash, P., and Karplus, M. 1990. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulation. *J. Comp. Chem.* 11:700-733.

Fletcher, R. and Reeves, C.M. 1964. Function minimization by conjugate gradients. *Computer J.* 7:149-153.

Foloppe, N. and MacKerell, A.D.J. 2000. All-atom empirical force field for nucleic acids. 1) Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comp. Chem.* 21:86-104.

Gao, J.L. 1996. Hybrid quantum and molecular mechanical simulations–An alternative avenue to solvent effects in organic chemistry. *Acc. Chem. Res.* 29:298-305.

Gorin, A.A., Ulyanov, N.B., and Zhurkin, V.B. 1990. S-N transition of the sugar ring in B-form DNA. *Molekulyarnaya Biologiya* 24:1300-1313.

Halgren, T.A. 1996. Merck molecular force field 1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comp. Chem.* 17:490-519.

Harvey, S.C., Tan, R.K.Z, and Cheatham, T.E. III. 1998. The flying ice cube: Velocity rescaling in molecular dynamics simulations leads to violation of equipartition. *J. Comp. Chem.* 19:726-740.

Hobza, P. and Sponer, J. 1999. Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical ab initio calculations. *Chem. Rev.* 11:3247-3276.

Jorgensen, W. 1995. BOSS, Version 3.6. Yale University, New Haven.

Jorgensen, W.L. and Tirado-Rives, J. 1996. Monte Carlo vs Molecular dynamics for conformational sampling. *J. Phys. Chem.* 100:14508-14513.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926-935.

Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118:11225-11236.

Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donni, D., Cieplak, P., Srinivasan, J., Case, D.A., and Cheatham, T.E. III 2000. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33:889-897.

Lahiri, A. and Nilsson, L. 1997. Properties of dianionic oxyphosphorane intermediates from hybrid QM/MM simulations: Implications for ribozyme reactions. *J. Mol. Struct.* 419:51-55.

Langley, D.R. 1998. Molecular dynamics simulations of environment and sequence dependent DNA conformation: The development of the BMS nucleic acid force field and comparison with experimental results. *J. Biomol. Struct. Dyn.* 16:487-509.

Lavery, R., Zakrzewska, K., and Sklenar, H. 1995. JUMNA (junction minimisation of nucleic acids). *Comp. Phys. Comm.* 91:135-158.

Leach, A.R. 1997. Molecular Modeling: Principles and Applications. Addison-Wesley, Reading, Mass.

Levine, I.N. 1991. Quantum Chemistry. Prentice Hall, Englewood Cliffs, N.J.

Maple, J.R., Hwang, M.J., Stockfisch, T.P., Dinur, U., Waldman, M., Ewig, C.S., and Hagler, A.T. 1994. Derivation of class II force fields. 1. Methodology and quantum force field for the alkyl functional group and alkane molecules. *J. Comp. Chem.* 15:162-182.

McCammon, J.A. and Harvey, S.C. 1987. Dynamics of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087-1092.

Nesterova, E.N., Federov, O.U., Poltev, V.I., and Chuprina, V.P. 1997. The study of possible A and B conformations of alternating DNA using a new program for conformational analysis of duplexes (CONAN). *J. Biomol. Struct. Dyn.* 14:459-474.

Noguti, T. and Go, N. 1985. Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. *Biopolymers* 24:527-546.

Northrup, S.H. and McCammon, J.A. 1980. Simulation methods for protein structure fluctuations. *Biopolymers* 19:1001-1016.

Pastor, R.W. 1994. Techniques and applications of Langevin dynamics simulations. *In* The Molecular Dynamics of Liquid Crystals (G. Luckhurst and C. Veracini, eds.) pp. 85-138. Kluwer Academic Publishers, Amsterdam, The Netherlands.

Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., Debolt, S., Ferguson, D., Seibel, G., and Kollman, P. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structure and energetic properties of molecules. *Comp. Phys. Comm.* 91:1-41.

Rick, S.W., Stuart, S.J., and Berne, B.J. 1994. Dynamical fluctuating charge force fields–application to liquid water. *J. Chem. Phys.* 101:6141-6156.

Ryckaert, J.P., Ciccotti, G., and Berendsen, H.J.C. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.* 23:327-341.

Saunders, M., Houk, K.N., Wu, Y.-D., Still, W.C., Lipton, M., Chong, G., and Guida, W.C. 1990. Conformations of cycloheptadecane. A comparison of methods for conformational searching. *J. Am. Chem. Soc.* 112:1419-1427.

Schlick, T. and Overton, M. 1987. A powerful truncated Newton method for potential energy minimization. *J. Comp. Chem.* 8:1025-1039.

Simmerling, C., Miller, J.L., and Kollman, P.A. 1998. Combined locally enhanced sampling and particle mesh Ewald as a strategy to locate the experimental structure of a non-helical nucleic acid. *J. Am. Chem. Soc.* 120:7149-7155.

Stanton, R.V., Little, L.R., and Merz, K.M. 1995. An examination of a Hartree-Fock molecular mechanical coupled potential. *J. Phys. Chem.* 99:17344-17348.

Steinbach, P.J. and Brooks, B.R. 1994. Protein simulation below the glass-transition temperature. Dependence on cooling protocol. *Chem. Phys. Lett.* 226:447-452.

Stewart, J.J.P. 1990. Semiempirical molecular orbital methods. *In* Reviews in Computational Chemistry (K.B. Lipkowitz and D.B. Boyd, eds.) pp. 45-81. VCH, New York.

Straub, J.E. 1996. Optimization techniques with applications to proteins. *In* New Developments in Theoretical Studies of Proteins (R. Elber, ed.) pp. 137-196. World Scientific, Singapore.

Szabo, A. and Ostlund, N.S. 1989. Modern Quantum Chemistry. McGraw-Hill, New York.

Thiel, W. 1997. Computational methods for large molecules. *J. Mol. Struct.* 398:1-6.

Tuckerman, M., Berne, B.J., and Martyna, G.J. 1992. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* 97:1990-2001.

Valleau, J.P. and Whittington, S.G. 1977. A Guide to Monte Carlo for Statistical Mechanics: 1. Highways. *In* Statistical Mechanics A. A Modern Theoretical Chemistry (B.J. Berne, ed.) pp. 137-168. Plenum Press, New York.

van Gunsteren, W.F. and Berendsen, H.J.C. 1987. Groningen molecular simulation (GROMOS) library manual. BIOMOS, Nijenborgh, Groningen, The Netherlands.

van Gunsteren, W.F. and Berendsen, H.J.C. 1990. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angew. Chem., Int. Ed. Engl.* 29:992-1023.

van Gunsteren, W.F. and Mark, A.E. 1998. Validation of molecular dynamics simulation. *J. Chem. Phys.* 108:6109-6116.

Wang, J., Cieplak, P., and Kollman, P.A. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comp. Chem.* 21:1049-1074.

Wiberg, K.B. 1965. A scheme for strain energy minimization. Application to the cycloalkanes. *J. Am. Chem. Soc.* 87:1070-1078.

Young, M.A., Jayaram, B., and Beveridge, D.L. 1997. Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: Fractional occupancy of electronegative pockets. *J. Am. Chem. Soc.* 119:59-69.

Zerner, M.C. 1991. Semiempirical molecular orbital methods. *In* Reviews in Computational Chemistry (K.B. Lipkowitz and D.B. Boyd, eds.) pp. 313-365. VCH, New York.

Zhurkin, V.B., Ulyanov, N.B., Gorin, A.A., and Jernigan, R.L. 1991. Static and statistical bending of DNA evaluated by Monte Carlo calculations. *Proc. Natl. Acad. Sci.* U.S.A. 88:7046-7050.

Internet Resources

http://www.ccl.net or http://www.netsci.org/Resources/Software/Modeling

Lists of available software.

http://www.msg.ameslab.gov/GAMESS/GAMESS.html

GAMESS website.

http://www.dl.ac.uk/CCP/CCP1/gamess.html

GAMESS-UK website.

http://www.gaussian.com

Gaussian program website.

http://www.schrodinger.com

Jaguar website.

http://www.wavefunction.com,

MNDO4 is available in many different codes from this website and others.

http://www.ccl.net

MOPAC 6.0 and 7.0 are in the public domain and are available from this (the Computational Chemistry mailing list site) and other sites. It is also available from the Quantum Chemistry Program Exchange, Department of Chemistry at the University of Indiana.

http://www.q-chem.com

Q-chem website.

http://www.msi.com

ZINDO website.

Contributed by Thomas E. Cheatham, III
University of Utah
Salt Lake City, Utah

Bernard R. Brooks
National Heart Lung & Blood Institute, NIH
Bethesda, Maryland

Peter A. Kollman
University of California
San Francisco, California