Molecular Modeling of Nucleic Acid Structure

Molecular modeling, loosely defined, relates to the use of models to investigate the three-dimensional structure, dynamics, and properties of a molecule or set of molecules. At the heart of this is specification of a molecular *model*, which provides a molecular structure at an appropriate level of granularity, usually in terms of three-dimensional atomic coordinates. Molecular modeling can be approached on many levels, ranging from energy minimization (finding the set of coordinates that minimizes the energy) with a complete ab initio quantum-mechanical treatment of the energetics, to sampling "reasonable" conformations with a simplified energy representation or potential, to the manipulation of physical models where no implicit energy representation is included. These methods serve not only as tools to aid in the interpretation of experimental data, but to directly complement such data by providing a relationship between the macroscopic behavior observed experimentally and the microscopic properties represented in the model or simulation.

As discussed in previous units, various molecular modeling tools can serve as conformational search engines for sampling conformational space subject to the restraints inferred from nuclear magnetic resonance (NMR; see *UNIT 7.2*) and crystallography (see *UNIT 7.1*) experiments. This is a critical step in the refinement of three-dimensional atomic structure. Inclusion of some representation of the energy, such as through the use of a specially parameterized empirical force field, can aid in this endeavor by limiting sampling to more realistic (in terms of energy) conformations.

As mentioned above, molecular mechanics methods can not only be used as a tool, but can directly complement experimental data. For instance, molecular dynamics simulations can be used to aid in the interpretation of NMR order parameters or to estimate anisotropic rotational diffusion. In addition, computer simulation techniques have the potential to give structural and dynamic insight into the atomic interactions occurring on a time scale (<μsec) typically not observable due to averaging in crystallography and NMR experiments. Ultimately, as methods are proven reliable, they can then be applied in cases where experimentation is limited, difficult, or unfeasible, such as studying highly flexible systems, investigating proposed chemical modifications that have yet to be synthesized, or to represent extremes of pressure, temperature, and concentration. As will become apparent, the methods are steadily improving to the point that reliable predictions are emerging.

A critical point that needs to be made at the outset is that these methods cannot be treated as a "black box" or hands-off procedure; there is no standard protocol that can be applied. Modeling is really more of an art. As each situation has differing requirements and needs, various choices need to be made as to what level of treatment to apply and what model to use. These choices rely on a critical understanding of the limitations in the methods. Therefore, the purpose of this discussion is to open up this black box a bit to allow some understanding of the options and choices a modeler makes, highlighting the tradeoffs that must be made in accuracy, system size, and time. The discussion here and in *UNITS 7.8* to *7.10* is not meant to provide a complete review of nucleic acid modeling, nor to substitute for the more complete treatment discussed in the primary literature. Instead, these units are intended to provide a framework that describes molecular modeling of nucleic acids, points out common issues and limitations, and points the reader to other useful information sources.

Implicit in this discussion is a realization that a molecular model is more than simply a representation of the covalent connectivity or static structure. The model may also include some representation of the energetics of the system and perhaps the dynamics over a particular time scale. Although it increases the utility, supplementing static structure with a representation of the energy and dynamics of molecular motion tremendously increases the cost of the modeling. For example, the simulations required to accurately represent the sequence-specific structure and molecular dynamics of a small, solvated nucleic acid duplex (<20 base pairs) on a nanosecond time scale would likely require weeks to months on available computer workstations, even with simple empirical energy representations. Of course, this added information may not always be necessary. For example, to investigate whether a proposed modification to a DNA base is steri-

Contributed by Thomas E. Cheatham, III, Bernard R. Brooks, and Peter A. Kollman
Current Protocols in Nucleic Acid Chemistry (2000) 7.5.1-7.5.12

cally feasible may only require the crude manipulation of a physical model to see an effect. Therefore, it is critical to understand the applicability, reliability, and limitations of these methods. In other words, the choice of the model depends on the question being asked.

The remainder of the discussion in this unit introduces the simplest levels of molecular modeling applied to nucleic acids. These include generation, evaluation, and characterization of the initial molecular model. At this simplest level, a nucleic acid model is limited to a static representation of the structure in the gas phase. Evaluation of this given model's utility is therefore based on the chemical intuition of the modeler, where manipulations to the model are limited to rotation about single bonds. To move beyond this level, supplement units in this series will delve more deeply into the myriad of issues involved in the computer simulation of nucleic acids. These include describing the common energy representations for nucleic acids that may be applied (*UNIT 7.8*), and discussion of how to properly represent the electrostatic interactions and solvation effects (*UNIT 7.9*). Additionally, various methods to find more representative structures are introduced, with a focus on molecular dynamics simulation methodologies. Finally, a description of practical issues in nucleic acid simulations will be provided (*UNIT 7.10*), such as what force fields are appropriate to apply, how simulations of nucleic acid are set up with explicit solvent and counterions, and how crude relative free energy differences can be estimated from molecular dynamics simulations. In these discussions, the focus will be on the middle ground in terms of size, time scale, and accuracy—that is, the simulation of small nucleic acids (typically less than ~250 base pairs), with explicit representation of the environment (if feasible or necessary), empirical pairwise potential functions, and time scales ranging from the analysis of individual snapshots to nanosecond-length simulations. For those readers more interested in learning about the simulation of larger nucleic acid systems (~1,000 to 15,000 base pairs), a variety of reviews can be consulted (Vologodski and Cozzarelli, 1994; Schlick, 1995; Olson, 1996).

MOLECULAR MODELING

The practice of molecular modeling basically involves the *generation* of an initial molecular model, *evaluation* of the model's utility, and perhaps *manipulation* of the molecular model (followed by further evaluation; see Figure 7.5.1).

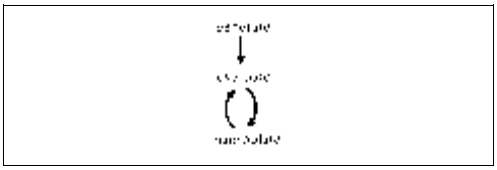

Figure 7.5.1 Schematic representation of molecular modeling analysis.

Prior to generating an initial molecular model, it is necessary to choose its representation or level of detail. For nucleic acids, the structural representation can be approached on many levels, ranging from the atomic level (including electrons) to coarser levels, such as those that model structure using a single point per base pair. The realism of the model directly depends on this choice of representation and further depends on what properties one is trying to represent. As shown in Table 7.5.1, modeling can be considered a tradeoff between the accuracy, the size and granularity of the system, and the time scale to be represented. If the model only concerns a single conformation or small set of conformations of a molecule of <100 atoms, a very accurate energy model and a description that includes all the atoms and electrons can be used (such as ab initio quantum mechanics with a fairly large basis set and even correlation). However, to investigate the supercoiling of a small DNA plasmid over a microsecond time scale, the system can no longer be represented at the atomic level, and a much simpler description of the energetics and a coarser representation of the structure must be imposed. However, this may be sufficient to represent the properties of interest. Between a full quantum mechanical treatment appropriate for small molecules and the coarse-grained single point per base pair model appropriate for large systems, molecular dynamics methods with an empirical potential may give reliable results as long as no "chemistry" is involved (such as bond forming, bond breaking, or electron transfer) and highly polarizable metal ions are treated at a very approximate level. These methods can give reliable insight into the sequence-specific structure and dynamics of a small nucleic acid duplex in solution.

The Static Structure Model

At the simplest level, and where the representation of the model does not include any reality beyond the covalent connectivity, molecular modeling can be performed by creating and manipulating physical models. Physical

Table 7.5.1 Tradeoffs in Molecular Modeling

Accuracy (increasing)	Time scale (decreasing)	System size (decreasing)	Granularity (finer grain)
Effective potential	Microseconds	Supercoiled DNA, plasmid	One point per base pair, elastic rod
Molecular mechanics (implicit solvent)	Nanoseconds to microseconds	<1000 base pairs	All atom, implicit solvent
Molecular mechanics	Nanoseconds	<250 base pairs	All atom, explicit solvent
Quantum mechanics	Individual snapshots	Nucleotide(s), few waters/ions	All atom plus electrons, implicit solvent

models are available that can represent three levels of granularity. At the finest level, there are a variety of atomic and molecular orbital models that represent the atoms and electrons. These molecular orbital models are not really appropriate for larger and more complicated molecules (such as anything larger than perhaps benzene), and therefore their use is really limited to teaching. Much more useful for representing nucleic acid structure are models that represent the atoms and bonds and, therefore, the covalent connectivity of a molecule.

There are a few common types of models in use that can be classified as either space-filling or bond-oriented. The most common space-filling models are of the Corey-Pauling-Koltun (CPK) variety, named after the researchers that developed them. These space-filling models represent the various atoms as cut-out spheres of a size proportional to the van der Waals radius, which are colored and shaped according to atom type and can be connected together (based upon the hybridization state and possible connectivity of the atom). The most common bond-filling models are polyhedral models. These provide a series of pieces that are in various polyhedral shapes with holes for pegs, which represent the bonds. The shape, color, and number of holes represent the various atom types (and hybridization state), and connecting pegs represent the bonds.

Although these models are useful for teaching and for building models of small molecules, they are not appropriate for building macromolecular models, such as of a DNA duplex. To build a larger molecule, special-purpose and more durable physical models can be purchased. These provide larger building units (such as DNA bases) in addition to smaller atom/half-bond units, which can be connected together. The scale of these models is usually in the 1 cm to 1 inch per Å range. Some models that have been used successfully are the Maruzen models, such as the HGS Biochemistry Molecular Model (see Internet Resources). Coarser folded-chain models, such as protein models that represent a connection/bond for each α-carbon, are also in use.

The physical bond-oriented models, although tedious to build and often very fragile, are very useful for gaining insight into atomic structure. In addition, the models can be manipulated (which can lead to problems with larger model structures, as they tend to deform). Although the models have rigid bonds and angles, they typically allow free rotation about single bonds. This can provide insight into the correlated conformational changes that occur upon change in a given coordinate. One example is the change in sugar pucker conformation from C2′-*endo* to C3′-*endo*, which lowers the rise between base pairs and shifts the conformation not only of the atoms in the ribose ring but also of the nucleic acid backbone. In fact, modeling B-DNA with physical models led to the formulation of Calladine's rules, which suggest means to overcome strong steric hinderances between adjacent purines in opposite strands as the base pair propeller twist increases to improve stacking.

Computational Graphics and Energy Models

A problem with physical models is that there is no reliable means to include a description of the energy. With these models, energy can only be represented rather crudely, such as by inhibiting free rotation because of the connectivity or by the addition of physical restraints to prevent rotation about double bonds. This allows a minimal interpretation of the *intra-*

molecular or internal energetics of the system (related to the connectivity of the molecule).

In addition to intramolecular interactions, a realistic depiction of the energy requires representation of the *inter*molecular interactions (e.g., van der Waals or steric repulsion and dispersion attraction interactions, hydrogen bonding, and electrostatic interactions). Although the solid-sphere models can represent steric repulsion, they cannot be used to accurately describe the total energy; however, a realistic treatment of the energetics can readily be calculated by computer. Coupled with molecular graphics (digital display of molecular models), computational energy models open the door for much more realistic and reliable molecular modeling. Prior to the advent of molecular graphics, physical models were routinely used as aids for crystallographic refinement.

Molecular graphics programs are now abundant and allow very nice and realistic display of molecular structure. The generality of the programs removes some of the tedium and cost of building physical models. However, since the computer graphics display is two-dimensional, the ease of seeing the three-dimensional model is lost and needs to be recovered by coloring, shading, or rotating the model to project the third dimension. Alternatively, stereoview displays can be used, which allow three-dimensional viewing with special glasses (either through shuttering, as with the Crystal Eyes display, or with coloring and shading). In addition to more general usage, adding a description of the conformational energy to the molecular model is easier on the computer.

Including a picture of the energy along with the molecular graphics can provide greater insight and help aid in the evaluation of the model. Examples include coloring regions of a molecule based on favorable electrostatic potential or highlighting atoms that show significant steric overlap. The manipulations possible at the simplest level mirror those of physical models and include a variety of coordinate manipulations, such as rotating about bonds or chemically modifying the structure. However, rather than manipulating the model by hand as with physical models, hooks need to be provided in the molecular graphics software to allow selection and rotation of various parts of the molecule.

Given a reliable initial model structure, molecular modeling with simple coordinate manipulations may be sufficient for many applications, such as suggesting that it is not feasible to fit a particular drug into the minor groove of a double-helical nucleic acid without seriously distorting the duplex, or showing that a certain chemical modification to the phosphodiester backbone is incompatible with the model structure. Simple modeling and molecular graphics were used as a guide in the initial design of peptide nucleic acid (PNA), an isosteric and stable backbone modification to DNA proposed for use as an antisense therapeutic agent (Nielsen et al., 1991).

Manipulation of molecular graphics or physical models, when coupled with an appropriate chemical/structural intuition, can give useful information. Examples include understanding steric effects, such as the interaction of drugs with the grooves or base pairs of nucleic acid duplexes or correlated changes in structure due to rotation about particular bonds. However, a major issue with this type of modeling is *evaluation* of the molecular models. Evaluation and interpretation of the meaning of the molecular model depends on the quality of the initial model, the reliability of the energy representation (if any), and the choice of coordinate manipulations to the model that might be made. Without a reliable guide into the conformational energetics and coordinate manipulations necessary to "improve" the model, evaluation of the model depends solely on the chemical intuition of the modeler. This intuition is necessary to rule out unfeasible or unrealistic models or to suggest manipulations to the model that may improve the property of interest.

Because there is no easy way to judge the quality of these models within this simple modeling framework, the conclusions made are often tenuous in the absence of experimental verification. For example, the initial model may not have been at all representative of what is seen experimentally or structural manipulations may lead to a model structure that is energetically unreasonable. Although the situation, in principle, improves with more advanced treatments because the energy is included and unreasonable coordinate manipulations are avoided, there are still many limitations in the methods. This is compounded by the sheer complexity of rugged energy landscapes for biomolecular structures, which makes evaluation of the reliability of a model structure difficult. In this sense, it should not be immediately assumed that "better" results are seen with more advanced treatments only because more reliable methods are used. There is still an essential need to compare the model

with experimental data and to critically evaluate the model.

To aid the modeler with simple molecular modeling, perhaps the ultimate molecular modeling environment might involve viewing a molecular graphics depiction of the model as it updates in real time according to the underlying energy potential, while the model is manipulated according to the whims of the modeler. An example of this type of program is Sculpt (Surles et al., 1994), which allows real-time minimization of the structure as it is manipulated. Further enhancement to this environment could come from visual and aural feedback from the system, such as a bang sound and flash of red light, to discourage manipulations by the modeler that move atoms into sterically forbidden regions. More involved haptic feedback mechanisms are also possible, such as increasing the difficulty of performing a given manipulation in proportion to the energetic penalty. Ultimately, molecular modeling environments of this type will incorporate visual, aural, and tactile feedback mechanisms, coupled with stereoscopic three-dimensional display in a virtual reality "cave" (Cruz-Neira et al., 1992), to guide the modeler as the model is manipulated. Software to perform this type of real-time modeling has become available in recent years, although the complexity of the calculations limits the treatment, and therefore fairly approximate representations of the energetics must be employed.

Nevertheless, this ultimate molecular modeling facility, with realistic energy representations and user feedback to steer the various molecular manipulations, unfortunately does not give a complete understanding of the molecular structure. The energy (enthalpy) alone is insufficient to describe the relative stability of various models, and care needs to be levied in judging the reliability of models based on differences in energy. In addition to describing the energy of the system, it is also necessary to include entropic effects. When entropic effects are included, free energy values may be obtained, providing the connection with reality and experimental measurement. With free energy, the modeler has a handle on the relative population of each state or can equivalently understand the various thermally accessible conformations of the molecule in its native environment.

To add entropic effects, some means of sampling the space of accessible conformations (according to the relative probability of observing a given conformation or equivalently according to the Boltzmann distribution) is needed. To do this, molecular dynamics (MD) or Monte Carlo (MC) simulation (discussed in more detail in *UNIT 7.8*) can be done with the given energy representation. This, however, tremendously increases the cost and complexity of modeling. Whether or not the sampled space of conformations is representative depends on the reliability of the energy description, the amount of conformational sampling, and the reliability of the initial model. However, it should be emphasized that more costly and detailed treatments do not always lead to "better" insight and are not always necessary to address the question at hand.

Generating the Initial Model

The first step in any modeling endeavor is creation of the initial molecular model, where "model" refers to a particular set of three-dimensional coordinates that define the structure of interest. In this discussion, which concerns nucleic acid structure on an atomic level (as opposed to the more coarse-grained bead models appropriate for modeling larger nucleic acid structures), this model is the set of three-dimensional atomic coordinates. Generally, a model of the coordinates is built by hand or received from another source (such as a database of experimentally derived structures). As will become more apparent later in this overview, the quality of the modeling in large part relates to the quality of the initial model or the ability to find or sample the "correct" structure given the initial model. In this regard, studying an unknown RNA structure is likely to be unfeasible at present, since it is unrealistic to imagine correctly folding up the RNA structure in dynamics simulations (due to barriers to conformational transition that cannot be overcome during the time scale of the simulations, and to inaccuracies in the energetic representation). Although there has been tremendous progress in predicting RNA secondary structure, predicting the overall tertiary structure (i.e., three-dimensional atomic coordinates) is still a major unsolved challenge. In spite of this, there have been a few attempts (for review see Brion and Westhof, 1997; Leclerc et al., 1997). Therefore, it is best to base the modeling on experimentally derived structures. Since DNA tends to adopt regular duplex structures, one can often use the canonical structures as an initial guess. The canonical models were derived from fiber diffraction studies of large DNA fibers and give an average idealized geometry and structure

representative of DNA under specific conditions (such as A-DNA under low humidity and B-DNA under physiological conditions; Arnott and Hukins, 1972). Crystallography provides another source of high-resolution structures, such as the left-handed Z-DNA duplex (Wang et al., 1979). The common canonical forms of DNA (A-DNA, B-DNA and Z-DNA) are shown in Figure 7.5.2 as stereo views. A good resource (although somewhat out of date) for general information on the structure of DNA is Saenger's excellent book (Saenger, 1984). High-resolution structures are also emerging from NMR spectroscopy (Ulyanov and James, 1995; also see *UNIT 7.2*). A more recent book surveying nucleic acid structure and interactions as well as NMR and crystallography studies is *Bioorganic Chemistry: Nucleic Acids* (Hecht, 1996).

Many of the experimentally derived nucleic acid structures are freely available through either the Protein Data Bank (PDB; see Internet Resources; Abola et al., 1987) or the Nucleic Acid Database (NDB; see Internet Resources; Berman et al., 1992), both of which contain the coordinates for a variety of nucleic acid structures and protein-nucleic acid complexes derived from crystallography or NMR experiments. The NDB may be a more appropriate place to start, as (1) it has been specifically tailored to assemble and distribute structural information about nucleic acids, (2) it can be searched, and (3) it provides coordinates (in multiple formats) as well as information about the crystal parameters, packing, and experimental conditions. From both of these sources, coordinate files in the commonly used PDB format can be obtained.

If an experimental structure is not available, it may still be possible to generate a reasonable model structure. A tool (or more accurately, a language for molecular manipulation) that can help develop such an initial model is Nucleic Acid Builder (NAB) developed by Tom Macke

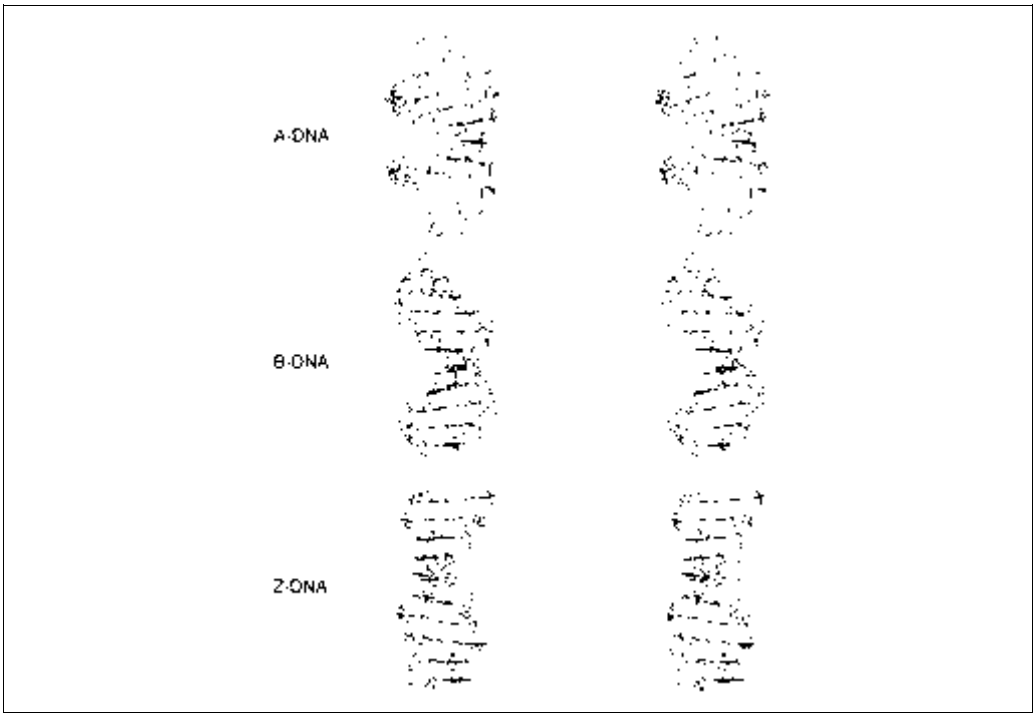

Figure 7.5.2 Canonical structures of DNA shown as stereo views. Shown are canonical models of A-DNA and B-DNA of d[CCAACGTTGG]$_2$ (Arnott and Hukins, 1972) and a 10-mer extended model of the Wang Z-DNA structure of d[CGCGCGCGCG]$_2$ (Wang et al., 1979) as stereo views. Stereo views are common in the literature; these are wall-eyed stereo views as opposed to cross-eyed. Although some people can view these directly, most people resort to one of a variety of hand-held stereo viewers, such as those based on mirrors or better ones that use focusing lenses. The model of Z-DNA was built by overlaying the two 6-mers at the joining region to the root-mean-squared (RMS) best fit overlapping CpG steps, and the A/B-DNA models were built using the NUCGEN module of AMBER 4.1 (Pearlman et al., 1995). The A-DNA and B-DNA models were all-atom RMS best fit to a common reference frame, and the view is into the major groove on top and the minor groove on the bottom.

and Dave Case (Macke and Case, 1998). The NAB molecular manipulation language allows a specification of rigid body translations, specification of restraints, distance geometry methods, and various other tools to aid in the generation of arbitrary structures. This has been used to generate model structures of synthetic Holliday junctions, protein-DNA complexes, RNA pseudoknots, supercoiled DNA, and other structures (Macke and Case, 1998). If the model shares properties with other known structures, such as common secondary structure elements or sequence, it may be possible to model by homology to the known structures or, alternatively, to build up the structure from a library of smaller pieces of known structure. This approach has been used to model RNA tertiary structure (Major et al., 1991) and the structure of DNA single strands (Erie et al., 1993).

Recent surveys of crystal structures in the Cambridge Structure Database (which contains a variety of high-resolution structures of mononucleosides and mononucleotides; Allen et al., 1979) and the NDB (Berman et al., 1992) provide a set of parameters that can serve as the beginnings of a dictionary for standard nucleic acid geometry. These surveys investigate the geometry of the bases (Clowney et al., 1996) and the sugar and phosphate backbone (Gelbin et al., 1996; Schneider et al., 1997). Additionally, recent surveys have investigated the specific hydration of nucleic acids and interaction with metal ions (Schneider et al., 1993; Schneider and Kabelac, 1998). High-level theoretical techniques can also give useful information. Ab initio quantum-mechanical simulations with a reasonable basis set (6-31 G* or better) and some inclusion of correlation can accurately represent geometry and polarization effects, and therefore properly represent nucleic acid interaction with various ions, metals, or nucleic acid bases. Monte Carlo and molecular dynamics simulation can also be used to obtain specific insight into ion association and hydration.

Completing the Initial Model

Often the experimentally determined structures obtained from the PDB or NBD lack explicit hydrogen atoms. Additionally, the nomenclature used is invariably different from that of the given modeling program, and the user has to impose various contortions to coerce the file into the expected naming and numbering conventions. Therefore it is fairly common to have to modify a PDB file to conform to the particular program's pedantic conventions and, additionally, to somehow add hydrogen atoms to the structure. Almost all of the modeling programs are equipped with some facility for adding missing atoms, particularly hydrogens. For more advanced treatments, solvent and counterions can also be added (discussed in *UNIT 7.9*).

It is always a good idea to check the initial structure carefully to determine if the conformation and nomenclature is as expected and whether the hydrogens are added with the correct stereochemistry. It would be very disappointing to discover, after spending weeks running nanosecond-length molecular dynamics simulations of solvated DNA, that one of the H1′ atoms on a particular residue was inadvertently added with the wrong stereochemistry, leading to an α-glycosyl linkage rather than the expected β linkage. It is likewise critical to check the stereochemistry of the structure after manipulations to the molecular model are made. Under some conditions, such as when using distance geometry methods or when performing stringent minimization with large restraints, the structure can be distorted and the stereochemistry altered.

Although not all modeling programs adhere to IUPAC naming conventions (JCBN, 1983; see *APPENDIX 1C*), these conventions are a good reference to check the naming, orientation, and placement of the various atoms. Additionally, there are a variety of tools for characterizing the nucleic acid structure, which are discussed in the next section. However, these methods do not necessarily check stereochemistry, depend on the use of correct hydrogen naming conventions, or enforce IUPAC naming conventions.

Although the PDB format is a common and well-defined standard for three-dimensional atomic coordinates, not all programs understand the standard PDB format, and they instead rely on some subtle variant or expect another coordinate format entirely. To aid in converting between the large set of formats available for many of the various modeling tools, the program babel is very useful (see Internet Resources). Not only can this perform direct conversion among various coordinate file formats, it can assign connectivity, bond orders, and hybridization when this information is not present.

Characterizing Nucleic Acid Structure

In order to characterize the quality of an initial molecular model or to later evaluate the conformational changes that occur as the model

is manipulated (for example, during MD simulation), it is useful to characterize the overall three-dimensional structure. In proteins, one is typically only concerned with the ϕ and ψ backbone angles and perhaps some of the side-chain χ angles; the overall structure is characterized by the particular secondary structure elements and folding class. In contrast, with nucleic acids, there are many angles of interest. These range from the backbone angles α, β, γ, ε, and ζ, to the puckering conformation of the furanose ring, to the χ angle representing the orientation of the sugar to the base (Saenger, 1984; see *APPENDIX 1B*). To characterize the conformation of the sugar moiety (the furanose ring), the Altona and Sundaralingam concept of pseudorotation is generally used (Altona and Sundaralingam, 1972). This defines the sugar pucker amplitude (representing how far the ring is from planar) and the pseudorotation phase angle (representing the correlated values of the individual torsions making up the ring). Various values of the pseudorotation phase angle, more commonly referred to as the sugar pucker, represent different puckerings out of the plane (on the same side as the C5′ atom, *endo*, or to the opposite side, *exo*). Methods for calculating these values are straightforward and are typically included in most modeling packages.

In addition to characterizing the overall backbone structure, sugar pucker, and χ angle of a single polynucleotide strand, it is also desirable to characterize the commonly occurring duplex structures that result from complementary base pairing between strands. Helicoidal analysis is typically applied to characterize global properties of the duplex (such as the helical repeat or overall helical twist), properties between adjacent base pairs (such as the rise), or properties of individual bases (such as the propeller twist). These properties represent the extent of rotation or translation of the bases or base pairs with respect to a common reference frame, typically the helical axis.

The nomenclature and definitions were standardized at an EMBO workshop on DNA curvature and bending (Dickerson et al., 1989). See Figure 7.5.3 for a graphical description of these values. Despite the standard nomenclature and definitions, the precise details of the mathematics were not standardized. Therefore, among the variety of programs commonly used to analyze helicoidal structure, each differs in the details regarding the exact definition of the helical axis, reference frame, and pivot points. Commonly used programs include NEWHELIX by Richard Dickerson, Curves by Heinz Sklenar and Richard Lavery (Lavery and Sklenar, 1988), and programs by Marla Babcock and Wilma Olson (Babcock et al., 1994) among others. The most developed and consistent mathematical treatment of the helicoidal

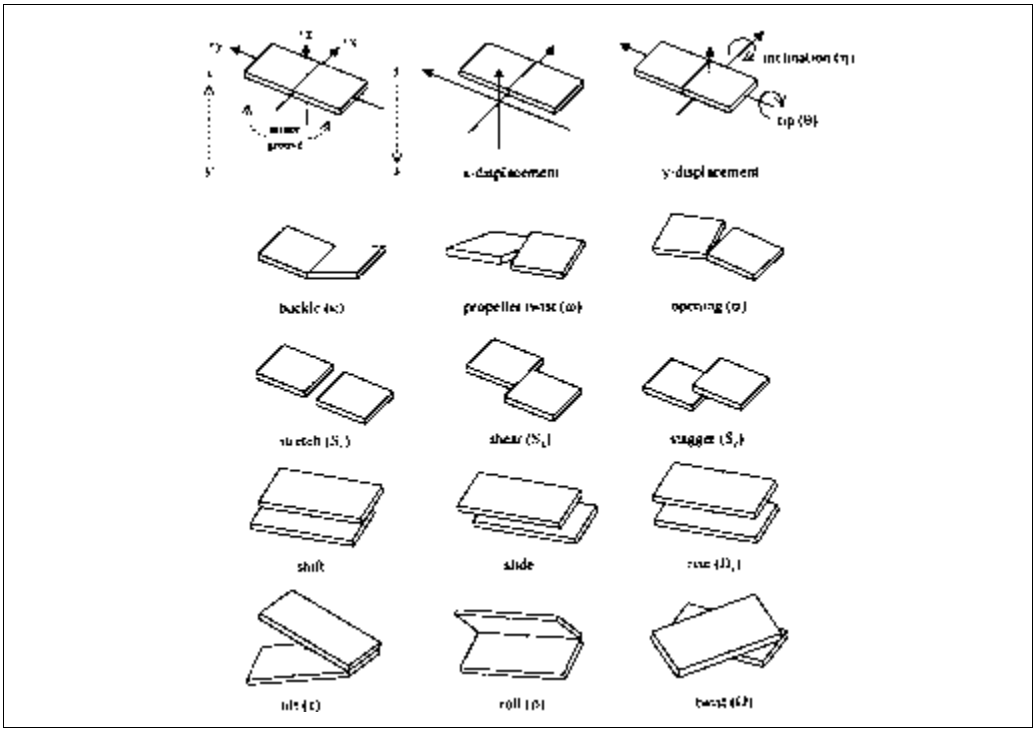

Figure 7.5.3 Pictorial definition of the helicoidal parameters.

parameters is likely either that of Babcock and Olson or that of Elhassan and Calladine, which is fully reversible (Elhassan and Calladine, 1995). The former has symmetrical definitions on a uniform scale for the various rotations and defines pivot points or axes that minimize mathematically induced artifactual correlations between the various rotational and translational parameters. Despite the advantages of these programs, NEWHELIX and Curves are the most commonly used programs to calculate helicoidal parameters. Although these methods give qualitatively comparable results, care should be taken in quantitative comparison of helicoidal values calculated from different programs due to the sensitivity of the method to definition of the reference frame. This is discussed in more detail in recent work by Lu and Olson (Lu and Olson, 1999; Lu et al., 1999).

A further distinction relates to global versus local helicoidal parameters; reference to a local helical axis typically relates to the axis between adjacent base pairs, whereas global helicoidal parameters are in reference to some best-fit global helical axis over the whole duplex. While the global parameters typically lead to more regular values (and less individual variation), the global axis may not be sufficiently determined for small duplexes (such as those with less than a full helical repeat) or distorted duplexes (such as an RNA duplex with a bulge), giving rise to misleading helicoidal parameters. The global axis may therefore not be appropriate. Moreover, given that the overall structure is determined by local interactions between adjacently stacked base pairs, local helicoidal parameters may be more representative. When comparing helicoidal values calculated during modeling to those in the literature, care should be taken to ensure that consistent reference frames (local versus global) and definitions of the values are applied. In addition to standard helicoidal analysis, groove structure is also commonly investigated, such as the relative width and depth of the minor or major groove (see, for example, Stofer and Lavery, 1994).

Helicoidal analysis and calculation of the various backbone angles can also be applied to the individual coordinate snapshots (for like conformations) or a representative coordinate-averaged structure generated during modeling, such as from a molecular dynamics or Monte Carlo simulation. Although it is often the case that average backbone angles calculated as the average of individual values for each coordinate snapshot are close to the values determined from the average structure, this is not typically true for helicoidal parameters, which are very sensitive to the conformation (Cheatham and Kollman, 1997). Modelers should keep in mind that the average structure obtained, such as that seen in crystallography or NMR experiments, hides the detailed dynamics. Moreover, coordinate-averaged conformations are not equivalent to torsion-averaged structures, which do not necessarily give average properties similar to that from the mean of the individual coordinate sets. Therefore, care should be taken in various coordinate comparisons. The common means to compare structures is through the use of best-fit root-mean-squared deviations (RMSd) between the coordinates or torsion angles. This indicator is very useful for determining the degree of similarity between two structures (when the RMSd values are small), but does less well at representing dissimilarity, since small differences in structure can lead to large root-mean-squared differences.

SUMMARY

This unit has introduced molecular modeling of nucleic acids on the simplest level. The modeling process can be described in three stages:

Generation. Create an initial model either by hand building it based on the molecular connectivity or by obtaining the coordinates from a depository of experimentally derived structures. In the absence of a complete experimental structure, base the structure on known (cannonical) structure and/or use tools (e.g., Nucleic Acid Builder) to complete the model.

Evaluation. Is the structure valid? Judge this based on chemical/structural intuition and comparison with experimentally derived structures. The structure can be described in terms of the backbone angles, sugar pucker, glycosidic χ torsion, and helicoidal parameters. Additionally, it is important to check the stereochemistry and hydrogen placement.

Manipulation. Coordinate manipulations can be made by simple rotation around chemical bonds. As possible, include some crude representation of the energy to avoid bad steric overlap and unrealistic rotations.

Other units delve more deeply into methods for evaluating and manipulating the models and representations of nucleic acids that go beyond the single static gas-phase structure model. This includes a discussion of how to properly represent the long-range electrostatic interactions and how to include some representation of the effect of the environment (solvent and ionic strength effects; see *UNIT 7.9*). With a more

realistic representation of the energy (*UNIT 7.8*), the energy can be used as a guide to suggest coordinate manipulations. Evaluation of the model depends on the reliability of the energy and how the system is represented, coupled with the chemical intuition of the modeler and comparison to experimental data.

LITERATURE CITED

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., and Weng, J. 1987. Protein Data Bank. *In* Crystallographic Databases—Information Content, Software Systems, Scientific Applications (F.H. Allen, G. Bergerhoff, and R. Sievers, eds.) pp. 107-132. Data commission of the international union of crystallography, Bonn/Cambridge/Chester.

Allen, F.H., Bellard, S., Brice, M.D., Cartright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R., and Watson, D.G. 1979. The Cambridge Crystallographic Data Centre: Computer-based search, retrieval, analysis and display of information. *Acta Crystallogr.* B35:2331-2339.

Altona, C. and Sundaralingam, M. 1972. Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Am. Chem. Soc.* 94:8205-8212.

Arnott, S. and Hukins, D.W. 1972. Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.* 47:1504-1509.

Babcock, M.S., Pednault, E.P., and Olson, W.K. 1994. Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J. Mol. Biol.* 237:125-156.

Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R., and Schneider, B. 1992. The nucleic acid database—A comprehensive relational database of 3-dimensional structures of nucleic acids. *Biophys. J.* 63:751-759.

Brion, P. and Westhof, E. 1997. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* 26:113-137.

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. Amer. Chem. Soc.* 119:4805-4825.

Clowney, L., Jain, S.C., Srinivasan, A.R., Westbrook, J., Olson, W.K., and Berman, H.M. 1996. Geometric parameters in nucleic acids: Nitrogenous bases. *J. Amer. Chem. Soc.* 118:509-518.

Cruz-Neira, C., Sandin, D.J., DeFranti, T.A., Kenyon, R.V., and Hart, J.C. 1992. The CAVE: Audio visual experience automatic virtual environment. *Commun. ACM* 35:65-72.

Dickerson, R.E., Bansal, M., Calladine, C.R., Diekmann, S., Hunter, W., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H.C.M., Olson, W.K., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D.M., Tung, C.S., Wang, A.H., and Zhurkin, V.B. 1989. Definitions and nomenclature of nucleic acid structure components. *Nuc. Acids Res.* 17:1797-1803.

Elhassan, M.A. and Calladine, C.R. 1995. The assessment of the geometry of dinucleotide steps in double helical DNA; a new local calculation scheme. *J. Mol. Biol.* 251:648-664.

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.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W.K., and Berman, H.M. 1996. Geometric parameters in nucleic acids: Sugar and phosphate constituents. *J. Amer. Chem. Soc.* 118:519-529.

Hecht, S. 1996. Bioorganic Chemistry: Nucleic Acids (S. Hecht, ed.) pp. 512. Oxford University Press, New York.

JCBN. 1983. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Abbreviations and symbols for the description of conformations of polynucleotide chains. Recommendations 1982. *Eur. J. Biochem.* 131:9-15.

Lavery, R. and Sklenar, H. 1988. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* 6:63-91.

Leclerc, F., Srinivasan, J., and Cedergren, R. 1997. Predicting RNA structures: The model of the RNA element binding Rev meets the NMR structure. *Folding Des.* 2:141-147.

Lu, X.-J. and Olson, W.K. 1999. Resolving the discrepancies among nucleic acid conformational analyses. *J. Mol. Biol.* 285:1563-1575.

Lu, X.-J., Babcock, M.S., and Olson, W. K. 1999. Overview of nucleic acid analysis programs. *J. Biomol. Struct. Dyn.* 16:833-843.

Macke, T. and Case, D.A. 1998. Modeling unusual nucleic acid structures. *In* Molecular Modeling of Nucleic Acids (N.B. Leontis and J. Santa Lucia, eds.) pp. 379-393. ACS, Washington, D.C.

Major, F., Turcotte, M., Gautheret, D., LaPalme, G., Fillion, E., and Cedergren, R. 1991. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science* 253:1255-1260.

Nielsen, P.E., Egholm, M., Berg, R.H., and Buchardt, O. 1991. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science* 254:1497-1500.

Olson, W.K. 1996. Simulating DNA at low resolution. *Curr. Opin. Struct. Biol.* 6:242-256.

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.

Saenger, W. 1984. Principles of Nucleic Acid Structure. Springer Advanced Texts in Chemistry (C.E. Cantor, ed.). Springer-Verlag, New York.

Schlick, T. 1995. Modeling superhelical DNA: Recent analytical and dynamical approaches. *Curr. Opin. Struct. Biol.* 5:245-252.

Schneider, B. and Kabelac, M. 1998. Stereochemistry of binding of metal cations and water to a phosphate group. *J. Am. Chem. Soc.* 120:161-165.

Schneider, B., Cohen, D.M., Schleifer, L., Srinivasan, A.R., Olson, W.K., and Berman, H.M. 1993. A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *Biophys. J.* 65:2291-2303.

Schneider, B., Neidle, S., and Berman, H.M. 1997. Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers* 42:113-124.

Stofer, E. and Lavery, R. 1994. Measuring the geometry of DNA grooves. *Biopolymers* 34:337-346.

Surles, M.C., Richardson, J.S., Richardson, D.C., and Brooks, F.P. 1994. Sculpting proteins interactively—Continual energy minimization embedded in a graphical modeling system. *Protein Sci.* 3:198-210.

Ulyanov, N.B. and James, T.L. 1995. Statistical analysis of DNA duplex structural features. *Methods Enzymol.* 261:90-120.

Vologodski, A.V. and Cozzarelli, N.R. 1994. Conformational and thermodynamic properties of supercoiled DNA. *Annu. Rev. Biophys. Biomol. Struct.* 23:609-643.

Wang, A.H., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G., and Rich, A. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* 283:743-745.

INTERNET RESOURCES

Simulation codes

http://www.amber.ucsf.edu/amber

The home page for the AMBER suite of programs for molecular mechanics and dynamics. See also the subpage http://www.amber.ucsf.edu/amber/polyA-polyT/ for a tutorial that describes in detail setting up, equilibrating, and running molecular dynamics simulations using AMBER on a small DNA duplex in solution.

http://igc.ethz.ch/gromos

The GROMOS molecular mechanics/dynamics software home page.

http://honiglab.cpmc.columbia.edu/grasp

The home page for the GRASP continuum electrostatics and molecular graphics display code developed by Anthony Nicholls.

http://www.lobos.nih.gov/Charmm

The CHARMM molecular mechanics/dynamics software home page at the National Institutes of Health. The root of this link discusses the LoBoS "lot's of boxes on shelves" parallel computer developed at the NIH for use in molecular simulation.

http://www.msi.com

The home page for Molecular Simulations, which distributed X-Plor and the commercial version of CHARMM.

http://www.intsim.com

The home page for the company Interactive Simulations, which develops the Sculpt software. This program allows real-time molecular modeling with continuous energy minimization as the model is manipulated.

http://www.ks.uiuc.edu/Research/namd

The home page for the NAMD molecular mechanics/dynamics simulation package developed by Klaus Shulten's group at the University of Illinois.

http://dasher.wustl.edu/tinker

The home page for the TINKER molecular mechanics/dynamics software. Includes an extensive list of WWW links to other MM/MD resources.

Model building and analysis tools, nucleic acid nomenclature

http://www.scripps.edu/case

The home page of Professor David Case at the Scripps Research Institute contains links to the NAB (Nucleic Acid Builder) software and manuals.

http://www.eyesopen.com/babel.html

The home page of the Molecular Structure Information Interchange Hub or the program babel developed in Professor Dan Dolata's group by Pat Walters and Matt Stahl. This program is very useful for interconverting a variety of different molecular modeling program file formats.

http://www.chem.qmw.ac.uk/iupac

A repository of many of the IUPAC naming conventions. This site has a very nice Web page describing in detail the notation and naming conventions that apply to nucleic acids.

http://www.sphere.ad.jp/hgs

The site for the company that makes the Maruzen physical molecular models (HGS). For protein and nucleic acids, of particular interest is the Maruzen Biochemistry Molecular Models.

Coordinate repositories and information resources

http://www.rcsb.org/pdb

The Protein Data Bank server at the Research Collaboratory for Structural Bioinformatics (Rutgers, SDSC, NIST).

http://ndbserver.rutgers.edu

The Nucleic Acid Database server maintained by Helen Berman and others at Rutgers University.

http://www.ccl.net/chemistry

The computational chemistry list archives. This contains information about a number of modeling programs, conference listings, and job postings.

http://cmm.info.nih.gov/intro_simulation/course_for_html.html

This page, sponsored by the Center for Molecular Modeling at the NIH, provides a nice introduction to macromolecular simulation.

Contributed by Thomas E. Cheatham, III and Bernard R. Brooks
National Heart, Lung and Blood Institute, NIH
Bethesda, Maryland

Peter A. Kollman
University of California
San Francisco, California