Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 16, Issue Number 4, (1999)
©Adenine Press (1999)

Overview of Nucleic Acid Analysis Programs[†]

http://www.albany.edu/chemistry/sarma/jbsd.html

Xiang-Jun Lu,
Marla S. Babcock
and Wilma K. Olson*
Department of Chemistry,
Rutgers, the State University
of New Jersey,
Wright-Rieman Laboratories,
610 Taylor Road,
Piscataway, NJ 08854-8087

Abstract

We outline the mathematical distinctions among seven of the most popular computer programs currently used to analyze the spatial arrangements of bases and base pairs in nucleic acid helical structures. The schemes fall into three basic categories on the basis of their definitions of rotational parameters: matrix-based, projection-based, and combined matrix- and projection-based. The approaches also define and construct base and base-pair coordinate frames in a variety of ways. Despite these mathematical distinctions, the computed parameters from some programs are strongly correlated and directly comparable. By contrast, other programs which use identical methodologies sometimes yield very different results. The choice of reference frame rather than the mathematical formulation has the greater effect on calculated parameters. Any factor which influences the reference frame, such as fitting or not fitting standard bases to the experimentally derived coordinates, will have a noticeable effect on both complementary base pair and dimer step parameters.

Introduction

The orientational and translational parameters now used to describe nucleic acid base-pair configuration—*i.e.*, the Tilt, Roll, Twist angles and Shift, Slide, Rise translations that define neighboring base-pair steps and the Buckle, Propeller Twist, Opening angles and Shear, Stagger, Stretch displacements that position complementary bases (Figure 1)—are not unique. The numerical values derived from the many currently available analysis routines, while mathematically rigorous and virtually identical for ideal B-DNA duplexes, differ significantly in highly distorted structures (1-4; Elgavish and Harvey, unpublished results). Conformational patterns derived from the analysis of representative double helices also depend on computational methodology (5), and the understanding of intrinsic structure and deformability deduced with one approach may not necessarily be transferable to treatments of molecules developed on another basis. This clearly frustrates collective attempts to decipher the subtleties of nucleic acid conformation and to extract the principles that underlie the base sequence-dependent responses of DNA and RNA to proteins, drugs, and other ligands.

Interest in resolving the discrepancies among nucleic acid analysis schemes has led us to reimplement seven of the most popular programs in current use—CEHS (6), CompDNA (7), Curves (8,9), FREEHELIX (10), NGEOM (11,12), NUPARM (13,14), and RNA (15-17)—in a single software package. By controlling the many factors which influence the computed results—including the algorithm for calculating parameters, the choice of reference frame, the construction of the "middle frame" needed to obtain parameters independent of chain direction, the standard base geometries, and the least-squares fitting procedures, we have uncovered the reasons why the various approaches yield different interpretations of nucleic acid structure. As reported elsewhere (5), the computed parameters are highly sensitive to the choice of reference frame. The seven schemes give very similar descriptions of base-pair geometry, even in the most deformed protein-DNA complexes, if the calculations are based on a common reference frame.

[†]This article is dedicated to David Beveridge on the occasion of his 60th birthday.

*Author to whom correspondence should be addressed. Phone: (732) 445-3993; Fax: (732) 445-5958; E-mail: olson@rutchem.rutgers.edu

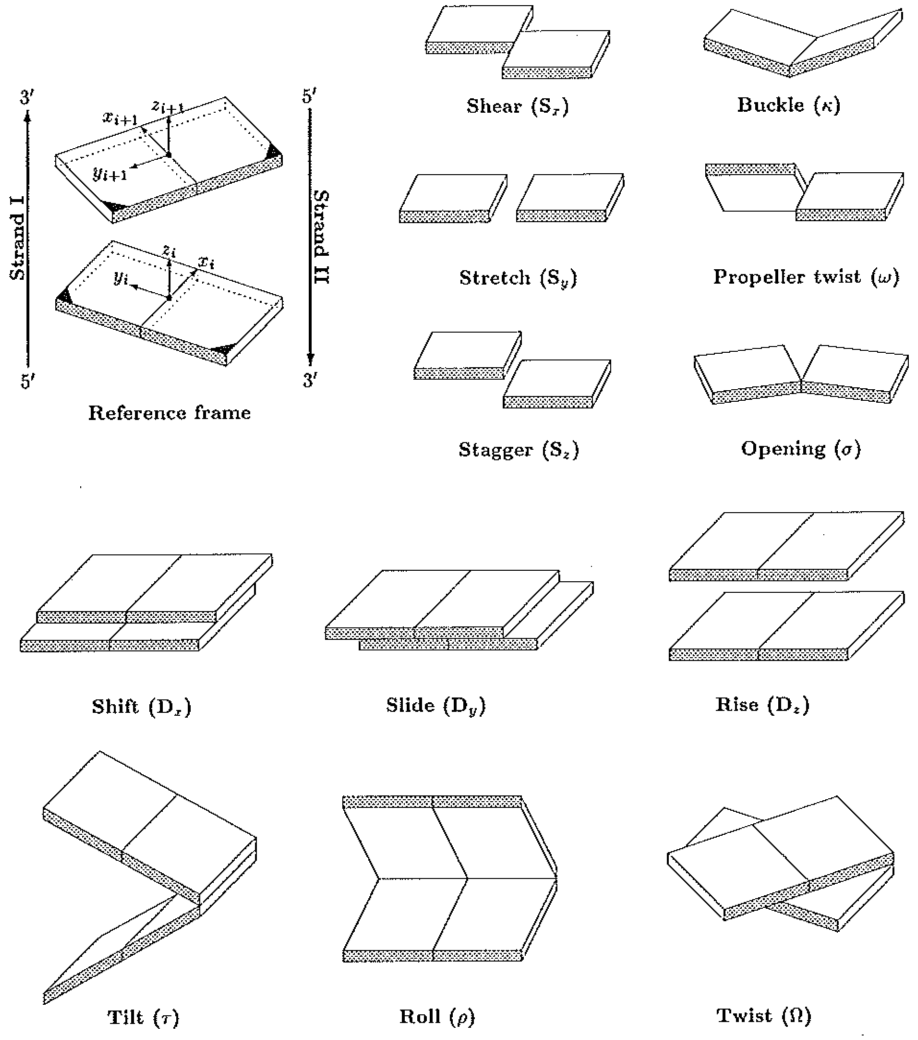

Figure 1: Pictorial definitions of parameters that relate complementary base pairs and sequential base-pair steps. The base-pair reference frame is constructed such that the x-axis points away from the (shaded) minor groove edge. Images illustrate positive values of the designated parameters (42).

In this report, we concentrate on the mathematical aspects of the seven programs as they relate to the six local step parameters and two base-pair parameters—Buckle and Propeller Twist—which are common to these schemes. Local parameters like these are essential for comparative studies of different types of structures and are widely used in both protein (18,19) and DNA conformational analyses (7,20). A global perspective of structure available in some nucleic acid analysis routines (8,9,13,14) provides a complementary description of base-pair geometry with respect to the overall helical axis. The global helical parameters between successive residues are helpful in pinpointing the sites of concentrated global deformations of an individual structure. Local parameters, however, are more typically used in statistical mechanical treatments of the global properties of nucleic acids in solution (21-23). The uncertainties involved in the construction of the overall helical axis (see below) also make the global frame less useful for rigorous comparisons of different structures.

Mathematical Comparisons

Overview

The parameters and coordinate frames used in various computer programs to describe the three-dimensional disposition of nucleic acid base pairs and base-pair steps are summarized in Tables I and II. The computational distinctions lie in (a) the precise definitions of the six parameters, three rotations and three translations, that relate a set of bases or base pairs, (b) the construction of the base and base-pair coordinate frames, (c) the choice of the "middle frame" which assures that the magnitudes of computed parameters are independent of the direction from which the structure is read, and (d) the fitting of reference standards to individual bases and base pairs. To facilitate comparison, the different approaches are described with the same terminology in the tables. Here $\hat{\mathbf{x}}_i$, $\hat{\mathbf{y}}_i$, $\hat{\mathbf{z}}_i$ denote unit vectors along the axes of the designated reference frame, the subscripts i and $i+1$ referring to adjacent base pairs and the subscript m to the "middle frame".

Table I

Comparative definitions of base-pair step rotations in seven popular nucleic acid analysis schemes.

Matrix-based Definitions

Program	Reference Matrix	Direct Parameters[†]	Tilt	Roll	Twist
CEHS	$\mathbf{R}_{\hat{k}}(\frac{\Omega}{2}-\phi)\mathbf{R}_{\hat{j}}(\Gamma)\mathbf{R}_{\hat{k}}(\frac{\Omega}{2}+\phi)$	Ω,Γ,ϕ	$\Gamma\sin\phi$	$\Gamma\cos\phi$	Ω
NGEOM	$\mathbf{R}_{\hat{k}}(\frac{\Omega}{2}-\delta)\mathbf{R}_{\hat{i}}(\Gamma)\mathbf{R}_{\hat{k}}(\frac{\Omega}{2}+\delta)$	Ω,Γ,δ	$\Gamma\cos\delta$	$-\Gamma\sin\delta$	Ω
RNA	$\mathbf{R}_{\hat{u}}(\varphi)$	$\hat{\mathbf{u}},\varphi$	$u_1\varphi$	$u_2\varphi$	$u_3\varphi$

Projection-based Definitions

Program	Tilt	Roll	Twist
CompDNA	$\sin(\frac{\tau}{2})=-\hat{\mathbf{y}}_m\cdot\hat{\mathbf{z}}_{i+1}$ [¶]	$\sin(\frac{\rho}{2})=\hat{\mathbf{x}}_m\cdot\hat{\mathbf{z}}_{i+1}$ [¶]	$\cos\Omega=P(\hat{\mathbf{y}}_i,\hat{\mathbf{z}}_m,\hat{\mathbf{y}}_{i+1})$
FREEHELIX	$\cos\tau=P(\hat{\mathbf{z}}_i,\hat{\mathbf{x}}_m,\hat{\mathbf{z}}_{i+1})$	$\cos\rho=P(\hat{\mathbf{z}}_i,\hat{\mathbf{y}}_m,\hat{\mathbf{z}}_{i+1})$	$\cos\Omega=P(\hat{\mathbf{y}}_i,\hat{\mathbf{z}}_m,\hat{\mathbf{y}}_{i+1})$
NUPARM	$\sin(\frac{\tau}{2})=-\hat{\mathbf{y}}_i\cdot\hat{\mathbf{z}}_m$	$\sin(\frac{\rho}{2})=\hat{\mathbf{x}}_i\cdot\hat{\mathbf{z}}_m$	$\cos\Omega=P(\hat{\mathbf{y}}_i,\hat{\mathbf{z}}_m,\hat{\mathbf{y}}_{i+1})$

Combined Matrix and Projection-based Definitions

Program	Tilt	Roll	Twist
Curves	$\cos(\frac{\tau}{2})\cos(\frac{\rho}{2})=\hat{\mathbf{z}}_m\cdot\hat{\mathbf{z}}_{i+1}$ [¶]	$\sin(\frac{\rho}{2})=\hat{\mathbf{x}}_m\cdot\hat{\mathbf{z}}_{i+1}$ [¶]	$\cos\Omega_+=\left(\mathbf{R}_{\hat{x}_m}(\frac{\tau}{2})\hat{\mathbf{y}}_m\right)\cdot\hat{\mathbf{y}}_{i+1}$ $\cos\Omega_-=\left(\mathbf{R}_{\hat{x}_m}(-\frac{\tau}{2})\hat{\mathbf{y}}_m\right)\cdot\hat{\mathbf{y}}_i$ $\Omega=\Omega_++\Omega_-$

[†] The NGEOM and CEHS phase angles are defined with respect to a common bending axis but measured in the opposite direction so that ϕ - $\delta = 90°$.

[¶] Simplified expressions derived from the original manuscript and verified by numerical methods.

Base-pair Step Parameters

While there is little ambiguity with regard to the translational parameters relating coordinate frames (see below), there are several alternative representations of the rotational parameters (Table I). Rotation angles (see Figure 1) are defined in terms of either the general rotation matrix $\mathbf{R}_{\hat{u}}(\varphi)$ (Eq. [1]) describing a rotation of magnitude φ about an arbitrary unit vector $\hat{\mathbf{u}} = u_1\,\hat{\mathbf{i}} + u_2\,\hat{\mathbf{j}} + u_3\,\hat{\mathbf{k}}$, where the $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, $\hat{\mathbf{k}}$ are unit vectors along the axes of the local Cartesian frame (24):

$$\mathbf{R}_{\hat{u}}(\varphi)=\begin{bmatrix}\cos\varphi+(1-\cos\varphi)u_1^2 & (1-\cos\varphi)u_1u_2-u_3\sin\varphi & (1-\cos\varphi)u_1u_3+u_2\sin\varphi\\ (1-\cos\varphi)u_1u_2+u_3\sin\varphi & \cos\varphi+(1-\cos\varphi)u_2^2 & (1-\cos\varphi)u_2u_3-u_1\sin\varphi\\ (1-\cos\varphi)u_1u_3-u_2\sin\varphi & (1-\cos\varphi)u_2u_3+u_1\sin\varphi & \cos\varphi+(1-\cos\varphi)u_3^2\end{bmatrix},\ [1]$$

or the projection $P(\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\mathbf{c}})$ (Eq. [2]) equal to the cosine of the "torsion angle" formed by the $\hat{\mathbf{a}}$ - $\hat{\mathbf{b}}$ - $\hat{\mathbf{c}}$ unit vector sequence, *i.e.*, the angle between vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{c}}$ projected onto the plane perpendicular to $\hat{\mathbf{b}}$:

$$P(\hat{\mathbf{a}},\hat{\mathbf{b}},\hat{\mathbf{c}})=\frac{\hat{\mathbf{a}}\cdot\hat{\mathbf{c}}-(\hat{\mathbf{a}}\cdot\hat{\mathbf{b}})(\hat{\mathbf{b}}\cdot\hat{\mathbf{c}})}{\sqrt{\left(1-(\hat{\mathbf{a}}\cdot\hat{\mathbf{b}})^2\right)\left(1-(\hat{\mathbf{b}}\cdot\hat{\mathbf{c}})^2\right)}}. \qquad [2]$$

The three matrix-based methods in Table I—CEHS, NGEOM, and RNA—define the three base-pair step rotations in terms of elements of the transformation matrix that brings coordinate frames of neighboring residues into coincidence. The first two packages make use of a sequence of symmetric Euler rotations about arbitrarily chosen coordinate axes, equating Twist to one of the matrix operations and expressing Roll and Tilt in terms of the net bending and the phase angle that "symmetrizes" the coordinate transformation. The third program, by contrast, relates the orientational variables to the components of the single rotational operation that brings the base-pair frames into coincidence. All three approaches provide sufficient information for a rigorous description, *i.e.*, reconstruction, of nucleic acid structure at the base-pair level.

The projection-based schemes in the Table—CompDNA, FREEHELIX, and NUPARM—equate the base-pair rotations to the angles formed by different sets of coordinate axes. All three of the above packages use the same definition of Twist, namely the pseudo-torsion angle describing the orientation of the long axes of neighboring base pairs with respect to the normal of the intervening "middle frame". While FREEHELIX defines Tilt and Roll by analogous torsions, CompDNA and NUPARM determine bending parameters that are analogs of valence angles. Regeneration of molecular coordinates from a set of valence angles and torsions is straightforward (25). The assorted valence angle projections in Table I also do not matter so long as the reference frames are orthogonal.

Finally, the Curves package takes advantage of both matrix and projection methods in defining dimer step angles (see Table I). The matrix operations used in determining Twist relate the "middle frame" to the base-pair coordinate axes with transformations identical to those employed in CEHS and NGEOM. Roll is a projected valence angle obtained much like the CompDNA and NUPARM values, whereas Tilt is deduced from Roll and the net bending angle.

Translational parameters are defined in all seven programs in terms of the projections of the virtual bond vector linking the origins of consecutive base pairs $(\mathbf{o}_{i+1} - \mathbf{o}_i)$ onto the x-, y- and z-axes of the "middle frame" (see Figure 2 and the discussion below). All programs define Shift and Slide consistently as the x- and y-components of this translational vector. Rise is similarly described in all programs, except for Curves, as the z-component of this vector. The Curves Rise is the sum of two equal distances, $|\mathbf{p}_i - \mathbf{o}_i|$ and $|\mathbf{p}_{i+1} - \mathbf{o}_{i+1}|$, illustrated in Figure 2.

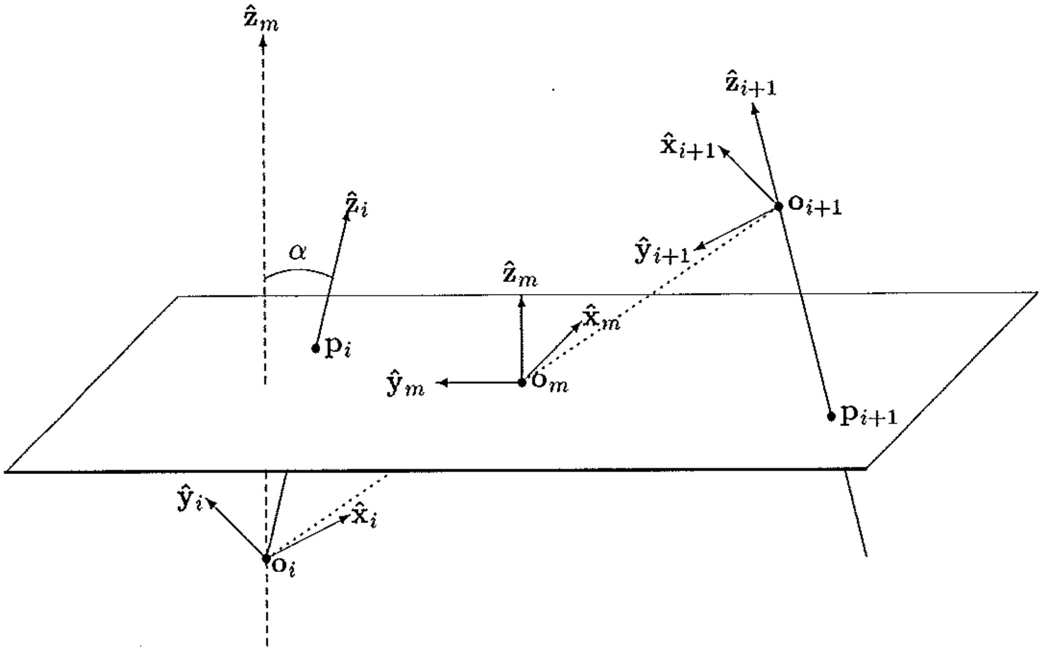

Figure 2: Schematic of base-pair frames i and $i+1$ and the "middle frame" m used to calculate local step parameters. The angle α is half the bending angle between successive base-pair normals.

Table II

Comparative constructions of base, base-pair, and dimer ("middle") coordinate frames in seven popular nucleic acid analysis packages.

Program	Base Frame	Base-pair Frame				Dimer/Middle Frame
		Origin	x-axis	y-axis	z-axis	
CEHS[†]	Normal from least-squares fit of base atoms (origin at N1(R)–C4(R) or N3(Y)–C6(Y) midpoint)	C8(R)–C6(Y) midpoint	$\hat{\mathbf{y}} \times \hat{\mathbf{z}}$	C8(R)–C6(Y)	Base-pair normal from least-squares fit (orthogonality correction)	"Middle frame" between adjacent base pairs (orthogonal)
CompDNA	Fitted standard with embedded reference frame (origin along a corrected C8(R)–C6(Y) line)	Mean of base origins	$\hat{\mathbf{y}} \times \hat{\mathbf{z}}$	Mean of y-axes (corrected for orthogonality)	Mean of base normals	$\hat{\mathbf{z}}_m$ from mean base-pair normals; $\hat{\mathbf{y}}_m$ from mean y-axes (corrected for orthogonality); $\hat{\mathbf{x}}_m = \hat{\mathbf{y}}_m \times \hat{\mathbf{z}}_m$
Curves	Fitted standard with embedded reference frame (origin at B-DNA fiber axis)	Mean of base origins	Mean of x-axes (corrected for orthogonality)	$\hat{\mathbf{z}} \times \hat{\mathbf{x}}$	Mean of base normals	$\hat{\mathbf{z}}_m$ from mean base-pair normals; $\hat{\mathbf{x}}_m$ from mean x-axes (corrected for orthogonality); $\hat{\mathbf{y}}_m = \hat{\mathbf{z}}_m \times \hat{\mathbf{x}}_m$
FREEHELIX	Normal from least-squares fit of base atoms; no other axes located	C8(R)–C6(Y) midpoint	$\hat{\mathbf{y}} \times \hat{\mathbf{z}}$	C8(R)–C6(Y)	Base-pair normal from least-squares fit	$\hat{\mathbf{z}}_m$ from mean base-pair normals; $\hat{\mathbf{y}}_m$ from mean y-axes (corrected for orthogonality); $\hat{\mathbf{x}}_m = \hat{\mathbf{y}}_m \times \hat{\mathbf{z}}_m$
NGEOM[¶]	Principal axes reference frame (origin at geometric center of base atoms)	Geometric center of base-pair atoms	Principal axes reference frame (naturally orthogonal)			"Middle frame" between adjacent base pairs (orthogonal)
NUPARM	Normal from least-squares fit of base atoms; no other axes located	C8(R)–C6(Y) midpoint or geometric center	$\hat{\mathbf{y}} \times \hat{\mathbf{z}}$	C8(R)–C6(Y) or C1′(R)– C1′(Y)	Mean of base normals	$\hat{\mathbf{x}}_m$, $\hat{\mathbf{y}}_m$ from mean base-pair axes; $\hat{\mathbf{z}}_m = \hat{\mathbf{x}}_m \times \hat{\mathbf{y}}_m$
RNA[§]	Fitted standard with embedded reference frame (origin along a corrected C8(R)–C6(Y) line)	Mean of pivot point-adjusted base origins	"Middle frame" between complementary bases (naturally orthogonal)			"Middle frame" between adjacent base pairs (orthogonal)

[†] CEHS "middle frame" corresponds to average of frames generated by rotation of base pairs through half the bending angle, *i.e.* $\Gamma/2$ in Table I.

[¶] NGEOM "middle frame" is identical to that of CEHS for rotational parameters and to that of RNA for translational parameters.

[§] RNA "middle frame" corresponds to a half-way rotation between coordinate frames, *i.e.*, square-root of rotation matrix $\mathbf{R}_{\hat{\mathbf{u}}}(\varphi)$ in Table I.

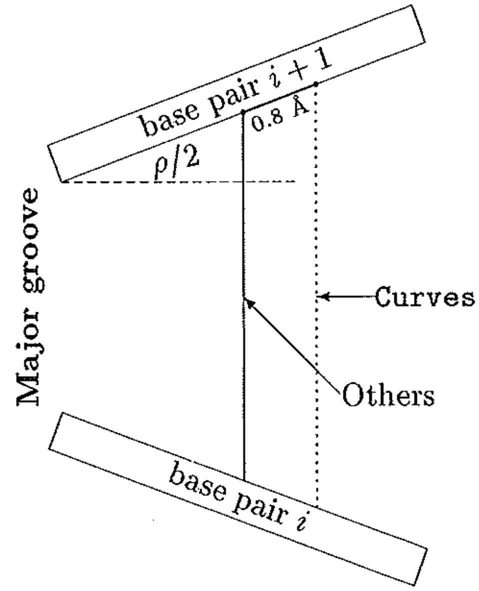

Figure 3: Comparative locations of local base-pair reference frames in Curves compared to other approaches. A positive Roll angle, as illustrated here, opens the minor groove and moves the Curves base-pair origins further apart.

Coordinate Frames

There is general agreement among all analysis packages in assigning a right-handed reference frame to bases and base pairs, and most programs also correct for orthogonality. The programs, however, differ in terms of the methods used to fit and define these axes (see discussion below and Table II). For example, the CompDNA, Curves, and RNA packages use least-squares methods to fix idealized chemical structures with embedded coordinate frames to individual bases, the first two programs placing the origin on the inside of the bases, *i.e.*, along what would be the dyad axis if the bases formed ideal Watson-Crick pairs, and the last program displacing the origin toward the centers of the bases. As noted in Figure 3, the Curves origin is also displaced along the dyad axis relative to other programs. By contrast, CEHS, FREEHELIX, and NUPARM use the best-fit normals to the base pairs and the C8(R)–C6(Y) long axis to construct the base-pair frame. The CompDNA, Curves, and RNA base-pair frames are averages of the two base frames (see Table II for details).

The NGEOM analysis differs from all other programs in taking the principal axes of inertia as the reference frame of bases and base pairs and locating the origins of these frames at the geometric centers of the respective units. Subtle distinctions in the chemical structure of individual Watson-Crick base pairs yield sequence-dependent coordinate frames and intrinsic differences in the local base-pair and dimer step parameters of regular structures.

The "middle frame" is typically constructed as an average of adjacent base-pair frames. The three matrix-based programs—CEHS, NGEOM, and RNA—define this frame by half rotations, either the axis positions generated by a half Twist and a half rotation about the bend axis (3) or the frame resulting from half the single rotational operation that brings the base-pair frames into coincidence. The projection methods express the "middle frame" as geometric averages of neighboring base-pair frames and must be corrected to ensure orthogonality. The averaging procedure is ambiguous in that the orientation of the "middle frame" depends on the order in which vectors are averaged and orthogonality is corrected. For example, NUPARM finds the mean dyad and long axes, *i.e.*, $\hat{\mathbf{x}}_m$ and $\hat{\mathbf{y}}_m$, prior to finding the "middle frame" normal, whereas other projection schemes find $\hat{\mathbf{z}}_m$ followed by $\hat{\mathbf{x}}_m$ or $\hat{\mathbf{y}}_m$. Because the various combinations give slightly different results, the geometric meaning of the dimer parameters is not quite as clear as in the matrix-based schemes.

Least-squares Fitting Procedures

a. Standard-bases with embedded coordinate frames

The least-squares procedures used in CompDNA, Curves, and RNA to fit a standard base with embedded reference frame to an observed base structure are mathematically equivalent. Both CompDNA and Curves take advantage of the conventional rotation matrix approach of McLachlan (26), while RNA implements a closed-form solution of absolute orientation using unit quaternions developed by Horn (27). Because the unit quaternion can be transformed to the familiar rotation matrix, the two algorithms should yield identical fittings. The three programs, however, employ different sets of standard base geometries and embedded coordinate frames so that numerical results disagree. For example, RNA uses data derived by Taylor and Kennard (28) from early high resolution crystal structures of model nucleosides, nucleotides, and bases, whereas the Curves and CompDNA standard is a B-DNA fiber model (29). The former data more closely fit the idealized base geometries recently compiled by Clowney *et al.* (30) from the Cambridge Structural Database (31) (see Table III).

Table III
RMS deviations (Å) between the coordinates of ring atoms in the standard bases used in DNA analysis programs vs. the updated dataset from Clowney *et al.* (30).

	A	C	G	T	U	average
RNA	0.003	0.003	0.004	0.006	0.004	0.004
CompDNA & Curves	0.009	0.009	0.008	0.018	0.012	0.011

The rotation matrix that brings standard and observed base frames into coincidence is found by a simple comparison of the relative positions of all atoms in the two forms (27). The *xyz*-coordinates of individual atoms $i = 1 \cdots \mathrm{N}$ in the standard and experimental bases, where N is the number of atoms, are represented respectively by the 1×3 vectors, $\mathbf{s}_i$ and $\mathbf{e}_i$. The atomic positions are then re-expressed in terms of their displacements, $\mathbf{s}'_i = \mathbf{s}_i - \bar{\mathbf{s}}$ and $\mathbf{e}'_i = \mathbf{e}_i - \bar{\mathbf{e}}$, from the geometric centers of each structure, $\bar{\mathbf{s}}$ and $\bar{\mathbf{e}}$, and the 3×3 covariance matrix **C** is constructed from the $\mathrm{N} \times 3$ matrices of collective coordinates, **S′** and **E′** (32),

$$\mathbf{C} = \frac{1}{\mathrm{N}-1}\left[\mathbf{S}'^{T}\mathbf{E}' - \frac{1}{\mathrm{N}}\mathbf{S}'^{T}\mathbf{i}\mathbf{i}^{T}\mathbf{E}'\right]. \qquad [3]$$

Here the superscript *T* signifies the transpose of a matrix, and **i** is an $\mathrm{N} \times 1$ column vector consisting of only ones.

Diagonalization of the 4×4 real symmetric matrix **M** constructed from the elements of **C**, *i.e.*, c_{ij}, where $i,j = 1,2,3$,

$$\mathbf{M} = \begin{bmatrix} c_{11}+c_{22}+c_{33} & c_{23}-c_{32} & c_{31}-c_{13} & c_{12}-c_{21} \\ c_{23}-c_{32} & c_{11}-c_{22}-c_{33} & c_{12}+c_{21} & c_{31}+c_{13} \\ c_{31}-c_{13} & c_{12}+c_{21} & -c_{11}+c_{22}-c_{33} & c_{23}+c_{32} \\ c_{12}-c_{21} & c_{31}+c_{13} & c_{23}+c_{32} & -c_{11}-c_{22}+c_{33} \end{bmatrix}, \qquad [4]$$

yields the unit quaternion (27). Specifically, the unit eigenvector, $\mathbf{q}_i$ ($i = 0 \rightarrow 3$), corresponding to the largest eigenvalue of **M** gives the rotation matrix **R** that brings the two frames into coincidence:

$$\mathbf{R} = \begin{bmatrix} q_0^2+q_1^2-q_2^2-q_3^2 & 2(q_1q_2-q_0q_3) & 2(q_1q_3+q_0q_2) \\ 2(q_2q_1+q_0q_3) & q_0^2-q_1^2+q_2^2-q_3^2 & 2(q_2q_3-q_0q_1) \\ 2(q_3q_1-q_0q_2) & 2(q_3q_2+q_0q_1) & q_0^2-q_1^2-q_2^2+q_3^2 \end{bmatrix} \qquad [5]$$

It should be noted that there is a sign ambiguity with regard to the unit quaternion: both $\mathbf{q}_i$ and $-\mathbf{q}_i$ satisfy the eigensystem associated with **M**. The sign, however, has no influence on **R**.

Once superimposed onto the experimental base, the fitted origin of the standard base is given by $\mathbf{o} = \underline{\mathbf{e}} - \underline{\mathbf{s}}\mathbf{R}^{T}$, and the coordinates of standard base atoms are transformed to $\mathbf{f}_i = \mathbf{s}_i\mathbf{R}^{T} + \mathbf{o}$. The root-mean-square (RMS) deviation between the standard and experimental structures is thus $\sqrt{\frac{\sum_{i=1}^{N}|\mathbf{e}_i - \mathbf{f}_i|^2}{\mathrm{N}}}$.

b. Base Normals

Rather than fit a standard base to experimental coordinates, CEHS, FREEHELIX, and NUPARM perform a least-squares fitting of a plane to a set of atoms (33,34) in order to define the base and base-pair normals. A covariance matrix based on the $\mathrm{N} \times 3$ matrix of Cartesian coordinates **E** is diagonalized to find the vector normal to the best plane. Specifically, **C** is obtained using Eq. [3] with both **S′** and **E′** substituted by **E**, and the normal vector lies along the eigenvector that corresponds to the smallest eigenvalue. It is also worth noting that the best-fitted global linear helical axis can be found with exactly the same algorithm, although two subroutines with similar functionalities are implemented in FREEHELIX.

c. Other Fittings

The diagonalization of the moments of inertia tensor in NGEOM automatically yields the least-squares points, line, and plane through a structure (35). Curves is unique in finding an optimized "curved" global helical axis by minimizing the variations in helical parameters between successive bases and the curvature between successive helical axis segments.

See our website, http://rutchem.rutgers.edu/~olson/jmb/prog_comp.html, for further details and examples of the various least-squares fitting procedures.

Numerical Comparisons

Correlations Among Programs

As is clear from Table I, the CEHS and NGEOM definitions of base-pair rotations are virtually identical, as are the CompDNA and NUPARM angular definitions. The calculated parameters obtained from each pair of methods, however, are not well correlated (see Figure 4). The seemingly different approaches from CompDNA, Curves, and RNA on one hand, and CEHS, FREEHELIX, and NUPARM on the other, however, do yield comparable results (Figure 4). As detailed in Table II, these

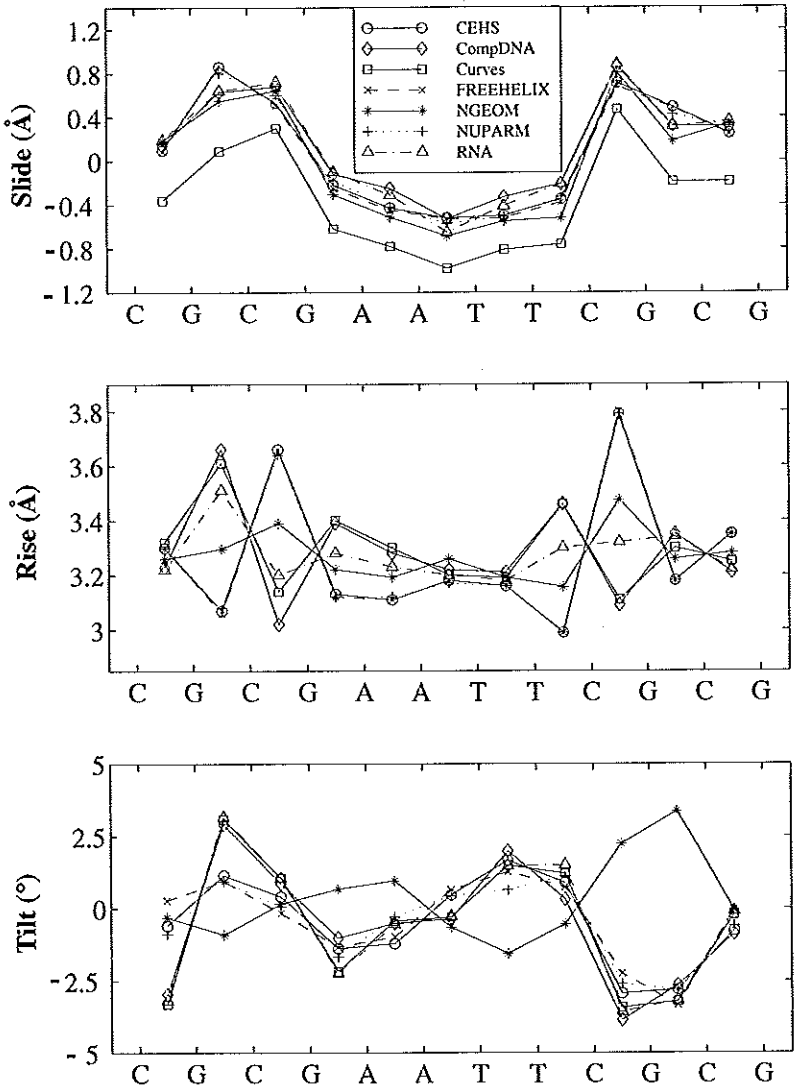

Figure 4: Comparison of the three local step parameters—Slide, Rise, and Tilt—in the 1.4Å high resolution B-DNA dodecamer duplex, d(CGCGAATTCGCG)$_2$ (BDL084) (36). It is clear that the Curves Slide is consistently ~0.5Å smaller than other values. Rise and Tilt obtained with CEHS, FREEHELIX, and NUPARM match closely, as do the CompDNA, Curves, and RNA values. The NGEOM results stand out from other parameters. The computed values of the three remaining base-pair step parameters—Slide, Roll, and Twist—show close agreement among the seven programs (data not shown).

subsets of computations make use of similar base-pair frames. The choice of reference frame is thus more critical than the precise algorithm used to determine base-pair parameters. Not surprisingly, "adjustments" must be made in the NGEOM package (12) to bring computed parameters (notably Twist) based on its principal axes of inertia frames in agreement with CEHS values based on the C8(R)–C6(Y) long axes and base-pair normals.

The large discrepancies in Curves parameters compared to all other methods stem primarily from an ~0.8Å displacement of the base origin towards the minor groove side of each base pair (Figure 3). This offset gives rise to systematic discrepancies of ~0.5Å in Slide (Figure 4) and ~0.8Å in global x-displacement (data not shown) in Curves compared to other programs, and also contributes to significant differences in Rise at kinked dimer steps (Figure 4). As detailed elsewhere (5), distorsions, such as Roll and Buckle, which displace base-pair origins, account for the most of the discrepancies in Rise. The unusual definition of the Curves Rise (Figure 2) also contributes to the computed differences in Figure 4, particularly at the third base-pair step, CG, where Roll ≈ 12°.

Fitting vs. Not Fitting Standard Bases

The Curves package gives the user the option of computing base-pair and dimer step parameters directly from the experimentally derived base positions or from ideal standards which are fitted to the observed coordinates. The experimental variations in base structures may influence the reference frames and thus affect the computed parameters. For example, the recently reported Curves analysis of base-pair geometry in the high resolution B-DNA d(CGCGAATTCGCG)$_2$ dodecamer duplex (36), Nucleic Acid Database (NDB) (37) entry: BDL084, was based on the experimental coordinates. Our re-analysis of the structure with fitted base standards shows noticeable differences in some parameters. Figure 5 illustrates the changes in Propeller Twist found by fitting vs. not-fitting idealized bases to this structure. The computed non-planarity of complementary base pairs as measured by Propeller Twist can vary by as much as 6° with the two approaches. The differences in computed parameters are even greater in the original Drew-Dickerson dodecamer of the same sequence (38), NDB entry: BDL001, which was refined without the benefit of new bond length and valence angle standards (30, 39). The average deviation in Propeller Twist computed with and without fitting standard bases in the lower resolution structure is ~5° vs. ~3° in the high resolution duplex (36), and some base pairs in the original dodecamer duplex show as much as ~9° difference in Propeller Twist with the two approaches.

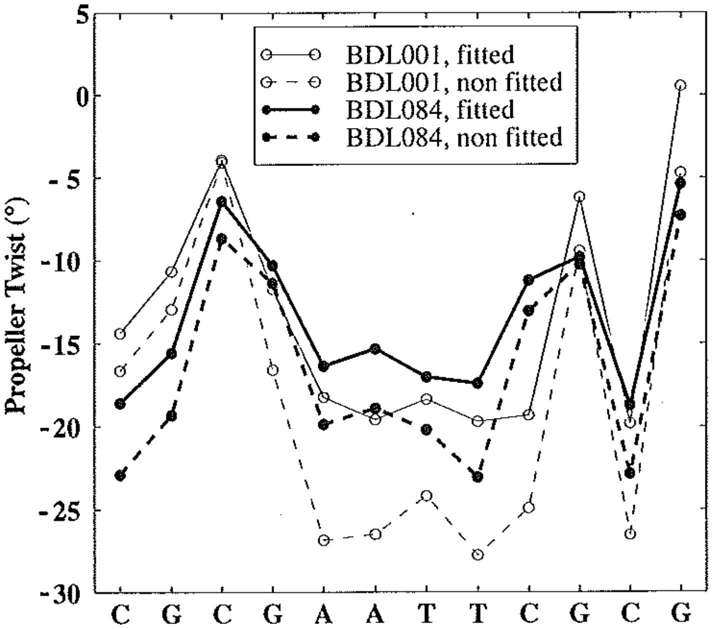

Figure 5: Effect of fitting vs. not fitting a standard base to the experimentally derived coordinates on the Curves Propeller Twist of the high resolution (BDL084) (36) and original (BDL001) (38) d(CGCGAATTCGCG)$_2$ dodecamer duplex structure.

Table IV

Dependence of R^2, the square of the correlation coefficient, on the method used to calculate complementary base-pair and dimer step parameters of the high resolution (1.4Å) d(CGCGAATTCGCG)$_2$ dodecamer duplex by Williams and co-workers (36) vs. the original Drew-Dickerson structure of the same sequence (38).

	Shear (Å)	Stretch (Å)	Stagger (Å)	Buckle (deg)	Propeller (deg)	Opening (deg)
Curves non-fitting	5	2	15	68	56	78
Curves fitting	8	12	46	58	67	74
CompDNA	9	2	39	48	72	71
	Shift (Å)	Slide (Å)	Rise (Å)	Tilt (deg)	Roll (deg)	Twist (deg)
Curves non-fitting	62	85	54	71	76	42
Curves fitting	47	89	29	65	71	48
CompDNA	47	83	37	51	71	49

The apparent influence of resolution on Propeller Twist and other parameters in the d(CGCGAATTCGCG)$_2$ dodecamer clearly depends upon computational methodology. As reported by Williams and co-workers (36), the deviations in Propeller Twist are appreciably greater in the original duplex than in their new 1.4 Å structure analyzed on the basis of the X-ray coordinates (dashed lines in Figure 5). The differences between corresponding points, however, are much less if the comparison of structures is based on parameters derived from sets of fitted bases (solid lines in Figure 5). Earlier arguments (36,40) concerning coordinate errors and parameter reliability in poorer resolution structures based on these differences are weakened when the standard base analysis is performed. Also bear in mind that Curves generates global base-pair parameters, which are also influenced by the overall deformations of a given structure (see below).

The local, based-fitted CompDNA Propeller Twist values show even smaller distinctions between the two dodecamer structures. Here experimental variations in base structures are minimized, and any differences between global helical axes are removed. The square of the correlation coefficient R^2 between CompDNA Propeller Twist values at corresponding residues in the two structures is 0.72 versus the value of 0.56 reported previously by Williams and co-workers (36) for the non-fitted Curves analysis and 0.67 for the base-fitted Curves parameters (Figure 5 and Table IV). Indeed, the extremely poor correlations of some base-pair parameters, *e.g.* Stagger, disappear when bases are fitted to the two structures (see Table IV). Notably, the inter-structural correlations of the three major step parameters—Twist, Roll, and Slide—are large and independent of computational methodology. These parameters are apparently well determined in the poorer resolution structure and are presumably key to the extraction of a sequence-dependent conformational code in the DNA base pairs (41).

Uncertainties in Global Parameters

Unlike local parameters which are independent of sequence context, the choice of the overall helix axis has an influence on the calculated global parameters (42). For example, the three global Tilt angles obtained with Curves for the first four d(CGCG)$_2$ base pairs of the Williams *et al.* structure are (–2.1°, 3.6°, and 1.5°) when analyzed in terms of the tetramer fragment but (–4.4°, 0.1°, and –1.1°) when computed on the basis of the overall dodecamer axis. The fitting of a straight line to the strongly curved DNA structures seen in many protein-DNA complexes is clearly meaningless. In such cases, global parameters make more sense if determined with respect to the optimal curved axial pathway (8,9).

Acknowledgments

It is a distinct pleasure to dedicate this paper to David L. Beveridge whose own contributions have illustrated the importance of base-pair parameters in understanding the deformations of DNA. We are also grateful to Stephen Harvey and Tricia Elgavish for sharing their unpublished data and to Zippora Shakked for valuable insights on base-pair analysis. This work was taken in part from the dissertation of Marla S. Babcock written in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Rutgers University, 1993. Our research has been generously supported by the U. S. Public Health Service under research grant GM20861. Computations were carried out at the Rutgers University Center for Computational Chemistry.

References and Footnotes

1. L. G. Fernandez, J. A. Subirana, N. Verdaguer, D. Pyshnyi, L. Campos and L. Malinina, *J. Biomol. Struct. Dynam. 15*, 151-163 (1997)
2. G. Guzikevich-Guerstein and Z. Shakked, *Nature Struct. Biol. 3*, 32-37 (1996).
3. X.-J. Lu, M. A. El Hassan and C. A. Hunter, *J. Mol. Biol. 273*, 668-680 (1997).
4. M. H. Werner, A. M. Gronenborn and G. M. Clore, *Science 271*, 778-784 (1996).
5. X.-J. Lu and W. K. Olson, *J. Mol. Biol.*, (in the press) (1999).
6. M. A. El Hassan and C. R. Calladine, *J. Mol. Biol. 251*, 648-664 (1995).
7. A. A. Gorin, V. B. Zhurkin and W. K. Olson, *J. Mol. Biol. 247*, 34-48 (1995).
8. R. Lavery and H. Sklenar, *J. Biomol. Struct. Dynam. 6*, 63-91 (1988).
9. R. Lavery and H. Sklenar, *J. Biomol. Struct. Dynam. 6*, 655-667 (1989).
10. R. E. Dickerson, *Nucl. Acids Res. 26*, 1906-1926 (1998).
11. D. M. Soumpasis and C. S. Tung, *J. Biomol. Struct. Dynam. 6*, 397-420 (1988).
12. C. S. Tung, D. M. Soumpasis and G. Hummer, *J. Biomol. Struct. Dynam. 11*, 1327-1344 (1994).
13. D. Bhattacharyya and M. Bansal, *J. Biomol. Struct. Dynam. 6*, 635-653 (1989).
14. M. Bansal, D. Bhattacharyya and B. Ravi, *Comput. Appl. Biosci. 11*, 281-287 (1995).
15. M. S. Babcock, E. P. D. Pednault and W. K. Olson, *J. Mol. Biol. 237*, 125-156 (1994).
16. M. S. Babcock and W. K. Olson, *J. Mol. Biol. 237*, 98-124 (1994).
17. E. P. D. Pednault, M. S. Babcock and W. K. Olson, *J. Biomol. Struct. Dynam. 11*, 597-628 (1993).
18. J. Singh and J. M. Thornton, *J. Mol. Biol. 211*, 595-615 (1990).
19. J. Singh and J. M. Thornton, *Atlas of Protein Side-Chain Interactions*, Vol. *I*, IRL Press at Oxford University Press, New York (1992).
20. M. A. El Hassan and C. R. Calladine, *Phil. Trans. R. Soc. Lond. A 355*, 43-100 (1997).
21. W. K. Olson, N. L. Marky, R. L. Jernigan and V. B. Zhurkin, *J. Mol. Biol. 232*, 530-554 (1993).
22. N. L. Marky and W. K. Olson, *Biopolymers 34*, 109-120 (1994).
23. J. A. Schellman and S. C. Harvey, *Biophys. Chem. 55*, 95-114 (1995).
24. H. Goldstein, *Classical Mechanics*, Addison-Wesley Publishing Co., Reading, MA (1981).
25. P. J. Flory, *Statistical Mechanics of Chain Molecules*, Wiley-Interscience Publishers, New York (1969).
26. A. D. McLachlan, *J. Mol. Biol. 128*, 74-79 (1979).
27. B. K. P. Horn, *J. Opt. Soc. Am. A 4*, 629-642 (1987).
28. R. Taylor and O. Kennard, *J. Am. Chem. Soc. 104*, 3209-3212 (1982).
29. A. G. W. Leslie, S. Arnott, R. Chandrasekaran, D. L. Birdsall and R. L. Ratlif, *Nature 283*, 743-746 (1980).
30. L. Clowney, S. C. Jain, A. R. Srinivasan, J. Westbrook, W. K. Olson and H. M. Berman, *J. Am. Chem. Soc. 118*, 509-518 (1996).
31. F. H. Allen, O. Kennard and D. G. Watson, in *Structure Correlation* (J. D. Dunitz and H.-B. Bürgi, eds.), Vol. *1*, pp. 71-110. VCH Publisher, New York, (1994).
32. T. P. E. Auf der Heyde, *J. Chem. Educ. 67*, 461-469 (1990).
33. D. M. Blow, *Acta Cryst. 13*, 168 (1960).
34. V. Schomaker, J. Waser, R. E. Marsh and G. Bergman, *Acta Cryst. 12*, 600-604 (1959).
35. D. M. Soumpasis, C. S. Tung and A. E. Garcia, *J. Biomol. Struct. Dynam. 8*, 867-888 (1991).
36. X. Shui, L. McFail-Isom, G. G. Hu and L. D. Williams, *Biochemistry 37*, 8341-8355 (1998).
37. H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S. H. Hsieh, A. R. Srinivasan and B. Schneider, *Biophys. J. 63*, 751-759 (1992).
38. H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura and R. E. Dickerson, *Proc. Natl. Acad. Sci. U.S.A. 78*, 2179-2183 (1981).
39. G. Parkinson, J. Vojtechovsky, L. Clowney, A. T. Brünger and H. M. Berman, *Acta Cryst. D52*, 57-64 (1996).
40. S. Neidle, *Nature Struct. Biol. 5*, 754-756 (1998).
41. W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock and V. B. Zhurkin, *Proc. Natl. Acad. Sci. U.S.A. 95*, 11163-11168 (1998).
42. R. E. Dickerson, *J. Biomol. Struct. Dynam. 6*, 627-634 (1989).

Date Received: December 1, 1998

Communicated by the Editor Ramaswamy H. Sarma