HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2018 April 24.

Published in final edited form as:

J Chem Inf Model. 2017 April 24; 57(4): 864–874. doi:10.1021/acs.jcim.6b00721.

Characterization of Biomolecular Helices and Their Complementarity Using Geometric Analysis

Kevin Hauser[1,#], **Yiqing He**[2], **Miguel Garcia-Diaz**[3], **Carlos Simmerling**[1,4,*], and **Evangelos Coutsias**[4,5,*]

[1]Department of Chemistry, Stony Brook University, Stony Brook, New York, 11794

[2]Great Neck South High School, Great Neck, New York, 11023

[3]Department of Pharmacological Sciences, Stony Brook University, Stony Brook, New York, 11794

[4]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York, 11794

[5]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York, 11794

Abstract

A general method is presented to characterize the helical properties of potentially irregular helices, such as those found in protein secondary and tertiary structures, and nucleic acids. The method was validated using artificial helices with varying numbers of points, points per helical turn, pitch and radius. The sensitivity of the method was validated by applying increasing amounts of random perturbation to the coordinates of these helices; 399,360 helices in total were evaluated. In addition, the helical parameters of protein secondary structure elements and nucleic acid helices were analyzed. Generally, at least seven points were required to recapitulate the parameters of a helix using our method. The method can also be used to calculate the helical parameters of nucleic acid-binding proteins, like TALE, enabling direct analysis of their helix complementarity to sequence-dependent DNA distortions.

Introduction

The helix is a ubiquitous geometric form in structural biology. It is essential to characterize the geometric properties of helices traced by atoms in a biomolecule because function is driven by structure and dynamics. There are two major problems with characterizing the geometric parameters of biomolecular helices. Irregularities between the atoms that trace the helix make it difficult to fit the points directly to the parametric equations of a helix. Also,

*Corresponding address: carlos.simmerling@stonybrook.edu, evangelos.coutsias@stonybrook.edu.
#Current address: Schrodinger, Inc., 222 Third Street, Cambridge, Massachusetts, 02142

Supplementary Information: Coordinates used to characterize Nievergelt's helix and the BurrH-DNA complex, analysis of Nievergelt's helix, helical axis deviations, $\%AAE_{twist}$ and $\%AAE_{rise}$ for the test helices, fitting residuals for α-, π- and 3_{10}-helical secondary structure elements, summary of nucleic acid helical parameters calculated using 3DNA, Curves+ and our method; software: all source code, libraries, examples, test cases and compilation instructions

biomolecular helices frequently trace less than one helical turn. Few methods exist that can derive accurate helical parameters from one-turn helices[1], and irregularities make the problem even more difficult.

In structural biology, empirical methods are available that can characterize the geometric parameters of irregular helices, but are only applicable to the molecular fragments for which they were trained. For example, DSSP[2] assumes a specific chemical connectivity associated with peptide secondary structures[3-5] (e.g. integral H-bonding between amides). HELFIT[6] and HELANAL[7] assume Cα-Cα chain connectivity in peptides to define an internal coordinate system (i.e. interlinked torsions along a chain of nine sequential Cα). For nucleic acids[8-9], 3DNA[10] and Curves+[11] find a helical axis by RMSD-fitting a nucleobase pair to a reference base pair structure whose helical axis is pre-defined[12].

A general method that does not require empirical constraints - such as a prior knowledge of the helical axis, helix radius or chemical topology - and is less affected by helical irregularities would enable the analysis of any biomolecular helical geometry. Such a method is needed to characterize the structures of superhelical nucleic acid binding proteins (NBPs) and modular superhelices[13]. The only such method known to us was developed by Nievergelt, a total least squares (TLS) approach that can characterize irregular helices consisting of points comprising a helix that traces just 90°[14]; for context, a single amino acid in a helical secondary structure element (SSE) traces ~100°.

The method we introduce here is an extension of TLS into a 2D linear least squares form. The simplification from 3D to 2D leads to several advantages; mainly, requiring fewer points to achieve the same level of accuracy. Like TLS, our method provides the ability to accurately calculate, without constraints, the helical parameters of helices traced by irregularly spaced points. Our method can also be used to characterize the helices traced by superhelical NBPs along with the nucleic acids they bind. The resulting helical parameters for the protein and the nucleic acid are directly comparable, providing a novel tool to analyze complementarity of the helical geometries involved in protein-nucleic acid binding. Such a method is likely to prove useful as the need to design highly specific genome editing enzymes[15] that recognize sequence-dependent DNA helix distortions continues to rise.

Theory

A cylindrical helix projects a circle on the *x*-*y* plane only if the *z*-axis is parallel to the helical axis. We deconstruct the problem into four parts. In part **1**, a general rotation matrix is derived. The algebra is simplified by assuming that the unit plane being rotated always passes through the origin of the coordinate systems (original and rotated). In part **2**, the rotation matrix is cast in spherical coordinates. In part **3**, the fitting problem is posed as a linear least squares problem and its solution described. In part **4**, we describe how the helical parameters are obtained from the optimized helix frame. The derived parameters represent the best helical curve through which the user-supplied points pass; the more points per helical turn, the smoother the helix will appear.

1. Rotate a helix in 3D, then project the helix on the *x-y* plane

The optimal helical axis must first be located in 3D space. The helical axis is defined as the normal vector to the plane onto which the helix projects a circle. To find this plane, a rotation matrix, **R**, is needed that relates the old coordinates of the points to rotated coordinates of the points. The original coordinates of the points are denoted (*X,Y,Z*) and the rotated coordinates are denoted (*x,y,z*).

A unit normal vector, $\hat{n}$, representing the plane ($aX + bY + cZ = 0$) containing the coordinates (*X,Y,Z*) is defined using direction cosines, $\hat{n} = (a,b,c)$. The basis vectors ($\hat{e}_X,\hat{e}_Y,\hat{e}_Z$) and ($\hat{e}_x,\hat{e}_y,\hat{e}_z$) represent the frame of the original coordinates (*X,Y,Z*) and the rotated coordinates (*x,y,z*), respectively. The unit vector $\hat{e}_z$ is chosen to be the unit normal vector of the plane (*a, b, c*):

$$\hat{e}_z = a\hat{e}_X + b\hat{e}_Y + c\hat{e}_Z \quad (1)$$

We define a right-handed coordinate system:

$$\hat{e}_X \cdot \hat{n} = a \quad (2)$$

The component of $\hat{e}_X$ that is parallel to $\hat{e}_z$ is

$$\hat{e}_{X\parallel} = a\hat{e}_z \quad (3)$$

and the perpendicular component $\hat{e}_{X\perp}$ is the rest of $\hat{e}_X$:

$$\hat{e}_{X\perp} = \hat{e}_X - a\hat{e}_z \quad (4)$$

Having defined $\hat{e}_z$ in equation (1), it can be substituted into equation (4):

$$\hat{e}_{X\perp} = \left(1 - a^2\right)\hat{e}_X - ab\hat{e}_Y + ac\hat{e}_Z \quad (5)$$

Because $\hat{e}_x$ is a unit vector equal to $\hat{e}_{X\perp}/|\hat{e}_{X\perp}|$, equation (5) can be written as

$$\hat{e}_x = \left(\sqrt{1-a^2}\right)\hat{e}_X - \left(\frac{ab}{\sqrt{1-a^2}}\right)\hat{e}_Y - \left(\frac{ac}{\sqrt{1-a^2}}\right)\hat{e}_Z \quad (6)$$

The cross product of the vectors $\hat{e}_z$, and $\hat{e}_x$ yields the vector $\hat{e}_y$, the result of which can be conveniently written as the symbolic determinant

$$\hat{e}_y = \zeta \begin{vmatrix} \hat{e}_X & \hat{e}_Y & \hat{e}_Z \\ a & b & c \\ 1-a^2 & -ab & -ac \end{vmatrix} = \zeta\,(c\hat{e}_y - b\hat{e}_z) \tag{7}$$

where ζ is $\frac{1}{\sqrt{1-a^2}}$. The matrix **R** that rotates the frame housing the original coordinates (*X,Y,Z*) into a new frame housing coordinates (*x,y,z*) is

$$\mathbf{R} = \begin{pmatrix} 1/\zeta & -ab\zeta & -ac\zeta \\ 0 & c\zeta & -b\zeta \\ a & b & c \end{pmatrix} \tag{8}$$

We can now rotate (any) Cartesian coordinates (*X,Y,Z*) using the rotation matrix in equation (8) by changing the direction cosine angles of the unit plane $\hat{n} = (a,b,c)$.

2. Spherical coordinates are used to rotate the helix frame

The unit plane described above passes through the origin. Thus, the two angle components of a spherical coordinate system are sufficient to rotate a set of coordinates over all 3D orientations using equation (8). The radial spherical component, ρ, can be obsoleted because the plane always passes through the origin. Therefore, the rotation of a helix in 3D Cartesian space using direction cosines with only two angles is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/\zeta & -ab\zeta & -ac\zeta \\ 0 & c\zeta & -b\zeta \\ a & b & c \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (9) \tag{9}$$

where $a = \sin\phi\cos\theta$, $b = \sin\phi\sin\theta$ and $c = \cos\phi$, ϕ and θ are the two polar spherical coordinates. As usual for spherical coordinates, ϕ is bounded by [0°,180°) and θ is bounded by [0°,360°).

3. A linear least squares problem is solved for the circle projected by the helix

For each rotation, the *x* and *y* coordinates of the rotated points are fit to a circle in the form of the linear least squares problem (**AX** = **B**)

$$\begin{pmatrix} 1 & 2x & 2y \end{pmatrix} \begin{pmatrix} k \\ x_0 \\ y_0 \end{pmatrix} = \left(x^2+y^2\right) \quad (10)$$

The **A** and **B** matrices in equation (10) are dimension 3 × *N* and 1 × *N*, respectively, where *N* is the number of points being fit. We use singular value decomposition (SVD) to calculate the best estimate of the parameters in the **X** matrix in equation (10). The *k* parameter in the **X** matrix contains the radius of the circle, *r*,

$$r = \sqrt{x_0^2+y_0^2+k} \quad (11)$$

which is simply the radius of the helix. The point where the helical axis intercepts the plane (*a*,*b*,*c*) is the circle center, (x_0,y_0). The residual of the fitting, χ, is

$$\chi^2 = \sum_{i=1}^{N} (x-x_0)^2+(y-y_0)^2+\rho^2-2\rho\sqrt{(x-x_0)^2+(y-y_0)^2} \quad (12)$$

A complete scan of spherical coordinate rotation angles ϕ and θ is performed, and the residual calculated for each discrete rotation. The rotation with the smallest residual (best fit) corresponds to the angles ϕ and θ of the optimal helix frame.

Figure 1 illustrates the rotation-projection method. The coordinates of the points tracing a suspected helix are projected onto the *x*-*y* plane. These points are then fit to a circle and the residual noted inside a program. The method continues to scan the range of (ϕ,θ) defined by the user, with spherical coordinate rotation scan-resolution also defined by the user. Once rotating and fitting across the full user-defined scan range is complete, the method determines which rotation yielded the projection that best fit a circle. Due to symmetry in the projections, a complete scan can be accomplished using ϕ in [0°,180°) and θ in [0°, 180°), rather than the usual bounds of ϕ in [0°,180°) and θ in [0°,360°).

4. Helix pitch, twist and rise are calculated from the optimal helix frame

A helical curve is defined by three properties: radius, pitch and helix axis (Figure 2). A discrete set of points that trace a helix curve can be used to derive these three properties. A helical step describes the geometry between successive points tracing a helix. Twist is the radial angle subtended by a helical step. Rise is the axial displacement spanned by a step. The height of a discrete helix is the displacement between the last and first point along the helical axis. Helical sweep is the sum of the twists. If the helical sweep, Φ, is 360°, then the height of the helix is the pitch because pitch is the axial displacement of the helix per turn (360°). In a unit helix, the radius and pitch have equal magnitude. A helix whose pitch is greater than its radius is shown in Figure 2a, and a helix whose pitch is smaller than its radius is shown in Figure 2b.

Methods

Nievergelt's helix

The Cartesian coordinates of Nievergelt's helix were obtained directly from the publication[14] (Table S1). A spherical coordinates scan over ϕ (from 0° to 180°) and θ (0° to 180°) was performed with 0.5° spherical coordinate scan resolution.

Generating artificial helices

The shape of a helix can be defined by the ratio of its radius and its pitch, *R/κ ratio*. Fractional values of *R/κ* represent thread-like (long and narrow) helices while values of *R/κ* larger than 1 represent ring-like helices (short and wide). The degree to which the helix resembles a polygon[16] (the tips of the translucent planes in Figure 2) or a smooth space curve (the blue curve in Figure 2) is determined by the number of points per helical turn, *PPT*. The number of turns in a helix with *N* total points is thus *N/PPT*.

For each helix shape *R/κ*, a matrix (represented as a shaded square in Figure 3) was generated by varying the number of points per helix, *N*, from 4 to 16, and varying points per turn, *PPT*, from 3 to 8. This generated 78 helices of the same shape, with varying length and granularity.

Helices were generated using the parametric equation:

$$x=R\sin\omega, y=\kappa\omega, z=R\cos\omega \quad (13)$$

where ω is the twist, which is 360°/*PPT*. The helical axis of these artificial helices coincide with the *y*-axis of the coordinate system. The spherical coordinates (ϕ,θ) of the helix frame (*a,b,c*) are (90°,0°). Helices were randomly oriented by rotating them using equation (9), with random values for ϕ bounded by [0, °180°) and θ bounded by [0°,360°).

Noisy helices generated by perturbing (x,y,z)—Noisy helices were generated by perturbing the Cartesian coordinates of each point in the artificial helices described above. Each component of each point was perturbed separately to mimic random (thermal) deviations. Noise levels were based on a coefficient of variation (CV^+),

$$CV^+=\frac{\sigma}{\mu} \quad (14)$$

where σ denotes standard deviation and μ denotes the mean. Here, μ is simply the original Cartesian component of a point from the ideal helices described in the previous section. To generate perturbed coordinates subject to a level of noise prescribed by equation (14), values for the perturbed coordinate components were drawn from a normal distribution whose mean was μ and whose standard deviation was σ.

The superscript "+" in "CV^+" is used to distinguish the variation applied to perturb the Cartesian coordinates (CV^+) from the resulting variation in helical parameters (CV_{twist} and CV_{rise}) calculated for the noisy helices. The following CV^+ values were used: 0.01, 0.05, 0.15 and 0.33. This random perturbation was carried out 256 times for each cell in the *N vs. PPT* matrix for each shape R/κ and for each value of CV^+. Overall, 399,360 noisy helices were generated. Figure 3 summarizes the set of noisy helices used in this study.

The accuracy of our results were quanitifed using percent average absolute error (%AAE). The absolute value of the difference between a calculated value and a reference value was divided by the reference value then multiplied by 100%: The equation for %AAE is

$$\%\mathrm{AAE}=100\% \times \left| \frac{b_{cal} - b_{ref}}{b_{ref}} \right| \quad (15)$$

where b_{cal} represents the value of the calculated parameter and b_{ref} represents the value of the reference value.

Generating biomolecular helices

Generating helical peptide secondary structure elements—Three polypeptide α-helix elements were generated using the LEaP module in Amber[17]. Three slightly different α-helical geometries were used to test the sensitivity of the method. Poly-alanine 32-mer peptides were generated by imposing backbone torsion angles of (-60°,-45°)[18] for ϕ and ψ (henceforth referred to as, $\alpha_{-60,-45}$) backbone torsion angles of (-57°,-47°)[19] for ϕ and ψ (henceforth referred to as, $\alpha_{-57,-47}$), and backbone torsion angles of (-60°,-40°) for ϕ and ψ (henceforth referred to as, $\alpha_{-60,-40}$).

One 3_{10}-helix peptide was generated using LEaP as above, except the backbone torsion angles for ϕ and ψ were (-49°,-27°)[19]. In addition, one π-helix was generated using LEaP as above, except the backbone torsion angles for ϕ and ψ were (-57°,-70°)[19]. Only the Cα atoms from the peptides were used in our fitting procedure.

The expected rise and twist between consecutive Cα atoms of each amino acid in the helical SSEs was calculated from the literature values[20] for pitch and residues per turn. Rise is pitch/residues per turn and twist is 360°/residues per turn.

Generating nucleic acid helices—Three nucleic acid duplexes were generated using 3DNA[10]. A-DNA was generated by imposing base pair (bp)-step twist and rise values of 32.7° and 2.548 Å, respectively. B-DNA was generated by imposing bp-step twist and rise values of 36.0° and 3.375 Å, respectively. A-RNA was generated by imposing bp-step twist and rise values of 32.7° and 2.812 Å, respectively. Curves+[11] was used to characterize the bp-step twist and rise using standard procedures[21]. The C1' atoms of these nucleic acids were used in the fitting. The method has the capability to utilize both helices in dsDNA and dsRNA to find the optimal common helical axis. Both strands were used in our analyses of double-stranded nucleic acids.

Analyzing a superhelical protein-DNA complex—The atomic coordinates of a modular DNA-binding protein with a superhelical tertiary structure, BurrH[22], was obtained from the Protein Data Bank[23] (PDB ID: 4CJA[22]). Superhelical Cα atoms were selected using our previously described approach[24]. Briefly, amino acids tracking the DNA binding cleft were identified (Table S2), one from each module of the protein, and the points of the helix were represented by the Cα atoms of these amino acids. The C1' atoms of both strands of the bound DNA were used to find the optimal DNA helical axis.

Using the test helices to estimate accuracy of an analysis—The number of points required to accurately characterize a helix using our method can be estimated by consulting the results of the analysis of the test helices. In the validation performed here, the R/κ ratio and PPT of helical secondary structure elements (proteins) and the helices of nucleic acids (single and double-stranded DNA and RNA) were known from the literature. The unknown parameter was the minimum number of points needed to accurately characterize these helices. If in future use of this method the R/κ and PPT is not known but the number of points is known, then the test helix results can be consulted to project the expected range of accuracy.

Results

The method requires fewer points than TLS to achieve the same accuracy

First, the method was compared directly to Nievergelt's related TLS[14] approach. The ten Cartesian coordinates published by Nievergelt trace an irregular helix whose pitch and radius were 600 and 195, respectively. We sought to determine the minimum number of points required by our method to reproduce the parameters obtained by Nievergelt's TLS approach. Figure S1 shows the results of the helix-fitting using our approach. With all ten points, our method achieves 0.5% and 0.4% relative error in pitch and radius respectively. Less than 1% relative error in pitch and radius was achieved with seven points and < 5% relative error in pitch and radius with only six points. For comparison, previous work showed that HELFIT requires all ten points to achieve 9% relative error in radius, and 0% relative error in pitch[6].

Validation tests of ideal artificial helices

The search for a helical axis over a scan of spherical coordinate rotations (step-size in ϕ and θ) is the only component of the fitting that is under user control. Scan step-size determines how finely rotations of the helix frame are made, ostensibly increasing the accuracy by increasing the likelihood that the helix can be ideally projected. The user-defined fitting parameters ϕ-range and θ-range were not evaluated during this test because changing them would amount to applying fitting constraints. This would improve the accuracy of the method in an obvious manner; therefore there is no need to test situations in which the user already knows some bounds of the solution.

How does the scan resolution affect accuracy?—As the scan is made coarser, it becomes less likely that the helix can be properly aligned during the projection operation, resulting in potential inaccuracies in calculated helical parameters. As a control, an ideal helix with $PPT = 6$ points per turn, $R/\kappa = 1$ radius/pitch ratio and 12 points, $N = 12$, was

characterized using different spherical coordinate step size (scan resolution) for the spherical coordinate scan. This helix was placed in 64 different random orientations prior to the spherical coordinates scan, and helical parameters were calculated for each orientation. Ideally, the scan is fine enough that the same parameters are calculated for all 64 orientations. The measured percent average absolute error (%AAE) indicates the inaccuracy expected solely from helical axis alignment errors due to the scan spacing (Figure 4a). As expected, the coarse scan resolutions (6° and 3°) did not recapitulate the input helix twist (60°) and rise (360°/6) as accurately as the finer scan resolutions because the structures could not be perfectly oriented. Scan resolutions of 1° grid steps or smaller resulted in good accuracy (AAE < 0.1%).

Next we tested whether the sensitivity for scan resolution was different for a typical non-ideal helix, as expected in biomolecular structures. A representative noisy helix was generated by perturbing the coordinates of the above regular helix by applying noise to the *x*-, *y*-, *z*-coordinates of each point in the helix. The reference pitch and radius for calculation of %AAE were obtained from the calculation using a fine scan with 0.25° scan resolution, which provided the best estimate of the otherwise unknown parameters. The helix was again placed in 64 random orientations. The results were comparable to those for the ideal helix, and increasing scan resolution displayed a sharp increase in accuracy (below 0.1% AAE) at 3° scan resolution, remaining below 0.1% AAE from scan resolutions of 1° scan resolution and below (Figure 4b). Overall, these results indicate that the accuracy of the method does not depend on scan resolutions below 1°, and performs comparably with noisy and ideal helices.

How sensitive are calculated helix parameters to random perturbations of Cartesian coordinates?—The goal of the following analysis was to determine the sensitivity of our method to uncertainty in the positions of the points tracing a helix. A broad validation was performed, in which all 78 helix geometries depicted schematically in Figure 3 were tested. The approach validates the method across a diverse range of helical parameters and coordinate perturbations. These perturbations were applied to the Cartesian coordinates of the helix points rather than the helical parameters themselves (twist, rise, radius), which were affected indirectly by the perturbations. We chose to perturb the helical parameters indirectly because our goal was to determine the dependence of the derived helical parameters on noise in the coordinates to represent the mix of distortions seen in structures of biomolecular helices, or uncertainties in selecting atoms to represent the helical superstructure. Noisy helices were generated by applying random perturbations to the *x*-, the *y*- and the *z*-coordinates of each point in the ideal helix subject to one of four CV^{+}'s (see **Methods**). Each helix-matrix (R/κ, CV^{+}) contained 78 distinct helix geometries, and each geometry contained 256 independent (perturbed) samples. A spherical coordinate scan resolution of 1° was used in the analysis. In the analysis below, “CV” is used to denote the coefficient of variation that was measured after using the method, whereas “CV^{+}” denotes the coefficient of variation that we applied to the data before using the method (see **Methods** for details).

We expected the test helices with the fewest points per helical turn (*PPT*) and the most points to be the least susceptible to noise because the helices would have more turns, and

more data to fit. Multiple turns help distinguish the points as a helix (a circle on the projection plane); conversely, if the points trace less than one turn and are noisy (e.g. $N = 5$, $PPT = 8$) they might appear to trace a 2D arc with ambiguous helix axis. In addition, we expected that increasing levels of applied coordinate perturbations (CV^+) would introduce increasing uncertainties (CV measured) to the derived helical parameters.

Figure 5 reveals the dependence of derived helical twist (CV_{twist}) on the noisy helices with diverse geometries. Helices with $R/\kappa > 1$ are wide-short helices, $R/\kappa = 1$ are unit helices, and $R/\kappa < 1$ are narrow-tall helices (Figure 2). The points of the helices were perturbed by four increasing levels of noise applied to their Cartesian coordinates ($CV^+ = 0.01$, $CV^+ = 0.05$, $CV^+ = 0.15$, $CV^+ = 0.33$, see equation (14)). As expected, the results show that the helices with the fewest total points (small N) and the fewest turns (larger PPT at each N) were the most susceptible to noise (large CV_{twist}), and as expected the effect grew with increasing applied CV^+. The analysis also revealed the dependence of derived helical parameter sensitivity with R/κ. For short helices with wide radii ($R/\kappa > 1$), the derived helical twist was less susceptible to noise than narrow-tall helix geometries. Conversely, the helical twist derived by the method was more sensitive to noise for extended helices with narrow radii. The method was sensitive to coordinate perturbations when $R/\kappa < 1$ because even for small applied CV^+ of 5%, helices with fewer than nine points were incorrectly rotated from the expected helical axis (Figure S2). It is important to note that the test helices in Figure 5 are very short – 4 to 16 points in total – representing the most challenging geometries expected in biomolecules. The %AAE for these the test helices also indicates that the shapes frequently observed in biomolecules ($R/\kappa < 1$) were also those that were most accurately calculated by our method (Figure S3).

Sensitivity of helical rise for regular helices with positional noise—Next we characterized how sensitive the derivation of the rise parameter was to CV^+ noise in the (x,y,z) coordinates of the points tracing the test helices. Since rise and twist are orthogonal parameters (the former depends only on $\hat{e}_z$ while twist depends on $\hat{e}_x$ and $\hat{e}_y$), we expected that rise might depend on helix shape (R/κ) in an opposite way as twist. Helices tracing less than one turn and helices with $N = 4$ points were still expected to be the most susceptible to noise because the former becomes degenerate with a noisy 2D arc, and the latter offers too few data distinguish noise (CV^+) from signal (the underlying helix). The sensitivity of helix rise for diverse geometries with increasing amounts of applied coordinate perturbations (CV^+) is shown in Figure 6. With the smallest applied CV ($CV^+ = 0.01$), the resulting helical rise matched our expectation of an R/κ-dependence opposite to that of twist. For the CV^+ values larger than 0.01, the results indicate that helix rise is more sensitive to noise than twist. CV_{rise} was at least 0.2 for all geometries, when CV^+ values of 0.05 and greater were applied. Like the results of the analysis above for twist, the percent AAE of derived rise for these the test helices indicates that the helices with $R/\kappa < 1$ were also those that were most accurately calculated by our method (Figure S4).

Summary of the analysis of test helices—Overall, twist was more precisely recapitulated than rise when characterizing noisy helices. Twist is a parameter of the helix that is fully contained within cylindrical slices (i.e. a projection plane), thus twist is

optimized directly by the method (*x*-, *y*-components) while rise is optimized indirectly. It is possible that twist is calculated more precisely than rise because twist is regularized during the fitting procedure. It is also possible that twist is more precisely characterized because twist contains two dimensions of information ($\hat{e}_x$ and $\hat{e}_y$) while rise contains only one ($\hat{e}_z$). The diverse test helices represent the limiting case because they possess the minimal geometric properties (number points and turns) required to unambiguously define a helix. The tests performed in this section represent the "worst case scenarios" potential users of the method might experience. The low R/κ ratio helices (1/2 and 1/4) led to the lowest %AAE for rise (Figure S2), twist (Figure S3) and most accurately derived helical axis orientations (Figure S4). These helix shapes are representative of protein secondary structure helices, nucleic acid helices and protein tertiary structure helices.

It is important to note the potential application of the method to augment the fitting of electron density maps in X-ray studies, or even more importantly, proton NMR or electron microscopy due to their inherent inferiorities. In structural NMR for instance, a helix is mainly characterized by NH/NH nuclear Overhauser effect volumes that are translated to distance restraints. It is believed that the NOE distances are accurate to 10%.[25-26] The method could be useful in structural experiments such as NMR if it can accurately characterize the geometry of helical secondary structure elements of proteins.

Testing helical secondary structure elements

α-helix secondary structures—The radius, pitch and *PPT* of an α-helix are 2.3 Å, 5.5 Å and 3.6,[20] respectively, and its R/κ is 0.42. Here, the Cα atoms from the amino acids were used as the points to carry out the fitting. Referencing Figures 5 and 6 for results on ideal helices with comparable *PPT* and R/κ (consulting test helices with *PPT*=4 and R/κ=1/2), we expected ~7 points would be required to accurately define the helical parameters of an α-helix. Figure 7 shows the results of our analysis of the helical parameters for α-helical secondary structure elements. Three slightly different ϕ/ψ backbone torsion angles within the α-helix region of the Ramachandran map were tested. The residual of the fitting for these three helical shapes rises linearly with the number of Cα atoms included in the fit because each atom adds to the total residual (Figure S5). As expected, seven Cα atoms were required to accurately calculate the helical rise (1.5 Å, Figure 7a), twist (100°, Figure 7b) and radius (2.3 Å, Figure 7c) for these ideal structures. Two turns are likely required to characterize helical secondary structure elements because the points tracing these helices are highly polygonal (< 4 *PPT*).

π-helix and 3_{10} secondary structures—The radius, pitch and *PPT* of a π-helix are 2.7 Å, 4.1 Å and 4.2,[20] respectively, and its R/κ is ~0.66. From the results of Figures 5 and 6 (consulting test helices with *PPT*=4 and R/κ=1/2), we expected ~8 points would be required to accurately calculate the helical parameters of a π-helix. The radius, pitch and *PPT* of a 3_{10}-helix is 1.9 Å, 5.8 Å[20] and 3.0 respectively and its R/κ was 0.33; (consulting test helices with *PPT*=6 and R/κ=1/4), we expected ~7 points would be required to calculate accurately the helical parameters of a 3_{10}-helix. Figure 8 shows the results of the analysis of the helical parameters for π- and 3_{10}-helical secondary structure elements. Seven and nine Cα atoms were required to accurately calculate the helical rise (Figure 8a), twist (Figure 8b) and radius

(Figure 8c) for π- and 3_{10}-helical elements, respectively. As with the a-helices, the residual of the fitting for these three helical shapes rises linearly with the number of Cα atoms included in the fit because each atom adds to the total residual (Figure S6). Overall, our method requires no more than nine atoms to achieve high-accuracy helix parameters for ideal protein secondary structure elements.

It stands to reason that our method could be used to determine the vector associated with a helix dipole because of its ability to determine the helical axis. *"In an α-helix the peptide dipole moments are aligned nearly parallel to the helix axis, the axial component being 97% of the dipole moment."* (Direct quote from reference[27]) The method could therefore be used to characterize the direction of the dipole moment in helical secondary structure elements.

Validation of nucleic acid helices: single- and double-stranded DNA and RNA

The method can calculate the helical parameters of single-stranded (ss) and double-stranded (ds) nucleic acids. However, 3DNA[10] and Curves+[11] can only define a helix if base pairs are present; the nucleic acid must be double stranded, precluding these methods from comparison with ours in the assessment of single-stranded nucleic acid helices. It is also important to note that 3DNA and Curves+ require multiple atoms per nucleotide to define a helix frame, whereas our method requires at minimum only one. Here, we used C1' atoms. Table S3 compares dsA-DNA, dsB-DNA and dsA-RNA rise and twist parameters obtained using our approach, Curves+[11] and 3DNA[10], which was used to generate the DNA structures. The overall agreement is excellent between the three methods, < 1% AAE for both parameters of all three nucleic acids, indicating that the method can be used to characterize helix parameters of nucleic acids.

How does the accuracy of the method depend on the number of bp (in dsDNA and dsRNA) or nucleotides (in ssDNA and ssRNA) used in the fitting? The accuracy of helix rise, radius and twist were expected to depend on the total number of C1' atoms used in the fitting. The radius, pitch and *PPT* of B-DNA are 10 Å, 34 Å and 10,[28] respectively, and its R/κ is ~0.29. The radius, pitch and *PPT* of A-DNA are ~11 Å, 28 Å and 11,[28] respectively, and its R/κ is 0.39; the radius, pitch and *PPT* of A-RNA are ~11 Å, 30 Å and 11,[28] respectively, and its R/κ is ~0.37. One turn of a B-DNA double helix has ten base pairs (bp), with ten atoms per strand. Based on our analysis of the test helices (consulting test helices with *PPT*=11 and R/κ=1/4), we expected to obtain accurate parameters when ~8 atoms in total were used, i.e. four bp of dsB-DNA. Figure 9 shows the results of the analysis of dsA-DNA, dsB-DNA and dsA-RNA. Four bp (eight atoms) were needed to accurately calculate the helical rise (Figure 9a) and twist (Figure 9b) of all three double-stranded nucleic acids.

The helical properties of single-stranded nucleic acids were characterized next. As with dsB-DNA, eight atoms were required to accurately calculate the helical parameters for ssB-DNA (Figure 10a,b). However, only seven atoms were required to accurately calculate helical rise (Figure 10a) and helical twist (Figure 10b) for ssA-DNA and ssA-RNA. Eight atoms trace significantly less the one helical turn - 288° for ssB-DNA and 229° for ssA-DNA and ssA-RNA. Considering the peptide results in the previous section, it is possible that seven atoms represents the lower-limit required by the method to obtain reliable and accurate helix parameters of biomolecular helices. Seven points may represent a critical value, which can

be used to either trace more turns of a coarse helix, or better define the curvature but using fewer turns.

Characterizing superhelix protein tertiary-structure

Unlike other helix analysis approaches, our method can characterize the helical parameters of superhelical protein tertiary structures without empirical constraints. Based on a simpler version of the method presented here, we previously calculated[24] the helical parameters of a modular human transcription factor MTERF1[29], which is structurally homologous to TALE proteins[30]. Thus, we were motivated to determine the generality of the method and characterize an alternate modular superhelical protein, the TALE protein BurrH[22], along with the helical parameters of the nucleic acid to which the protein was bound. We expected that the helical parameters of the protein and the DNA would be similar because of the high degree of apparent structural complementarity (Figure 11). Points tracing the BurrH superhelix were defined as the Cα atoms of one amino acid from the same location in each module (Table S2). Without covalent connections between these atoms or a reference frame to define the superhelical axis, ours is the only approach capable of characterizing an irregular superhelical tertiary structure. The results of the analysis indicate that, as expected, the average rise and twist of the protein (3.28 Å ± 1.1 Å and 32.6° ± 2.4° respectively) were nearly identical to those calculated for the DNA (3.34 Å ± 0.4 Å and 31.8° ± 4.7° respectively). The radius of BurrH (20.5 Å) was larger than the radius of the DNA (6.6 Å) because the protein traces a wider helix that wraps around DNA. Deviations in the parameters reflect local variation of the steps between superhelical Cα atoms for steps between modules. The deviations in rise and twist for BurrH - 1.1 Å and 2.4° respectively - reflect the average helix irregularity between modules (steps). Superhelical repeat-step parameters, their variation and their complementarity to the bp-step parameters of DNA will be expanded upon elsewhere.

Conclusion

We developed and tested a general method of helix-fitting that was geared towards characterizing geometries observed in structural biology applications. Validation tests were based on 399,360 test helices whose geometric parameters were representative of diverse biomolecules whose atoms might be perturbed from an ideal helix. The test helices were perturbed with known levels of noise to determine the sensitivity of the method. Helices frequently observed in structural biology applications were tested (peptide helical secondary structure elements, and nucleic acid single- and double-helices). The method was also used determine the helical complementarity of a TALE protein's superhelical tertiary structure and the DNA to which it was bound. Overall, the method introduced here is general, accurate and robust to noisy helical geometries. Based purely in geometry, our method can be used to characterize complementarity in protein-nucleic acid complexes, with potential applications in the design of genome editing reagents and biomaterials, astronomy and particle physics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

KH thanks Mosaverrul Hassan and James Maier for helpful discussions.

Funding Sources: National Institutes of Health (NIH) Ruth L. Kirschstein National Research Service Award [F31-GM101946 to K.H.]; Chemical Biology Training Program Fellowship [T32-GM092714 to K.H.]; National Science Foundation (NSF) Louis Stokes Alliance for Minority Participation Bridges to the Doctorate Fellowship [HRD-0929353 to K.H.]; NSF Alliance for Graduate Education and the Professoriate-Transformation Fellowship [HRD-1311318 to K.H.]; NIH and National Institute of General Medical Sciences (NIGMS) [R01-GM090205 to E.A.C.]; NIH NIGMS [R01-GM100021 to M.G.D.]; NIH NIGMS [R01-GM107104 to C.S.]; NSF Petascale Computational Resource (PRAC) Award from the National Science Foundation [OCI-1036208]. C.S. and E.A.C. acknowledge support from Henry and Marsha Laufer.

References cited

1. Christopher NA, Swanson R, Baldwin TO. Algorithms for Finding the Axis of a Helix: Fast Rotational and Parametric Least-Squares Methods. Comput Chem. 1996; 20(3):339–345. [PubMed: 8673326]
2. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers. 1983; 22(12):2577–2637. [PubMed: 6667333]
3. Astbury WT, Street A. X-ray Studies of the Structure of Hair, Wool, and Related Fibres. I. General. Philos Trans R Soc, A. 1932; 230:75–101.
4. Astbury W, Woods Ht. X-ray Studies of the Structure of Hair, Wool, and Related Fibres. II. The Molecular Structure and Elastic Properties of Hair Keratin. Philos Trans R Soc, A. 1934; 232:333–394.
5. Pauling L, Corey RB, Branson HR. The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. Proc Natl Acad Sci U S A. 1951; 37(4):205–11. [PubMed: 14816373]
6. Enkhbayar P, Damdinsuren S, Osaki M, Matsushima N. HELFIT: Helix Fitting by a Total Least Squares Method. Comput Biol Chem. 2008; 32(4):307–10. [PubMed: 18467178]
7. Bansal M, Kumar S, Velavan R. HELANAL: A Program to Characterize Helix Geometry in Proteins. J Biomol Struct Dyn. 2000; 17(5):811–9. [PubMed: 10798526]
8. Franklin RE, Gosling RG. Molecular Configuration in Sodium Thymonucleate. Nature. 1953; 171(4356):740–1. [PubMed: 13054694]
9. Watson JD, Crick FH. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953; 171(4356):737–8. [PubMed: 13054692]
10. Zheng G, Lu XJ, Olson WK. Web 3DNA – A Web Server for the Analysis, Reconstruction, and Visualization of Three-Dimensional Nucleic-Acid Structures. Nucleic Acids Res. 2009; 37(Web Server issue):W240–6. [PubMed: 19474339]
11. Blanchet C, Pasi M, Zakrzewska K, Lavery R. CURVES Plus Web Server for Analyzing and Visualizing the Helical, Backbone and Groove Parameters of Nucleic Acid Structures. Nucleic Acids Res. 2011; 39:W68–W73. [PubMed: 21558323]
12. Babcock MS, Pednault EP, Olson WK. Nucleic Acid Structure Analysis. Mathematics for Local Cartesian and Helical Structure Parameters that Are Truly Comparable Between Structures. J Mol Biol. 1994; 237(1):125–56. [PubMed: 8133513]
13. Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, Bradley P. Rational Design of α-Helical Tandem Repeat Proteins with Closed Architectures. Nature. 2015; 528(7583):585–588. [PubMed: 26675735]
14. Nievergelt Y. Fitting Helices to Data by Total Least Squares. Computer Aided Geom Des. 1997; 14(8):707–718.
15. Porteus M. Genome Editing: A New Approach to Human Therapeutics. Annu Rev Pharmacol Toxicol. 2016; 56:163–90. [PubMed: 26566154]
16. Whitworth, WA. The Regular Polygon in Space. In: Whitworth, W AllenPendlebury, CTR., Glaiser, JWL., editors. The Oxford, Cambridge and Dublin Messenger of Mathematics. Vol. 4. Macmillan & Co.; Cambridge: Trinity Street, Corner of Green Street: 1875. p. 88-89.

17. Case D, Babin V, Berryman J, Betz R, Cai Q, Cerutti D, Cheatham Iii T, Darden T, Duke R, Gohlke H. Amber 14. 2014
18. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015; 11(8):3696–713. [PubMed: 26574453]
19. Armen R, Alonso DO, Daggett V. The Role of α-, 3_{10}-, and π-Helix in Helix→Coil Transitions. Protein Sci. 2003; 12(6):1145–1157. [PubMed: 12761385]
20. Guo ZY, Kraka E, Cremer D. Description of Local and Global Shape Properties of Protein Helices. J Mol Model. 2013; 19(7):2901–2911. [PubMed: 23529181]
21. Pasi M, Maddocks JH, Beveridge D, Bishop TC, Case DA, Cheatham TC, Dans PD, Jayaram B, Lankas F, Laughton C, Mitchell J, Osman R, Orozco M, Perez A, Petkeviciute D, Spackova N, Sponer J, Zakrzewska K, Lavery R. μABC: A Systematic Microsecond Molecular Dynamics Study of Tetranucleotide Sequence Effects in B-DNA. Nucleic Acids Res. 2014; 42(19):12272–12283. [PubMed: 25260586]
22. Stella S, Molina R, Lopez-Mendez B, Juillerat A, Bertonati C, Daboussi F, Campos-Olivas R, Duchateau P, Montoya G. BuD, A Helix-Loop-Helix DNA-Binding Domain for Genome Modification. Acta Crystallogr Sect A. 2014; 70:2042–2052.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–242. [PubMed: 10592235]
24. Hauser K, Essuman B, He Y, Coutsias E, Garcia-Diaz M, Simmerling C. A Human Transcription Factor in Search Mode. Nucleic Acids Res. 2016; 44(1):63–74. [PubMed: 26673724]
25. Wüthrich, K. NMR of Proteins and Nucleic Acids. Wiley; 1986.
26. Guntert P, Mumenthaler C, Wuthrich K. Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. J Mol Biol. 1997; 273(1):283–98. [PubMed: 9367762]
27. Hol WG, van Duijnen PT, Berendsen HJ. The α-Helix Dipole and the Properties of Proteins. Nature. 1978; 273(5662):443–6. [PubMed: 661956]
28. Saenger, W. Principles of Nucleic Acid Structure. Springer Science; 1984.
29. Yakubovskaya E, Mejia E, Byrnes J, Hambardjieva E, Garcia-Diaz M. Helix Unwinding and Base Flipping Enable Human MTERF1 to Terminate Mitochondrial Transcription. Cell. 2010; 141(6): 982–93. [PubMed: 20550934]
30. Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL. The Crystal Structure of TAL Effector PthXo1 Bound to its DNA Target. Science. 2012; 335(6069):716–9. [PubMed: 22223736]

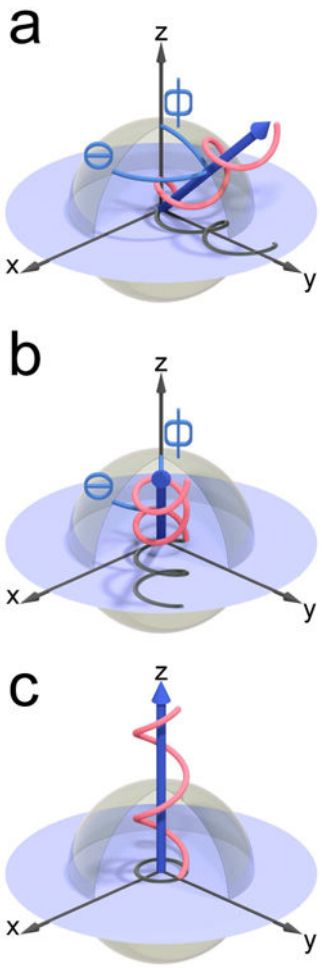

Figure 1.

The frame of a helix is rotated using spherical coordinates to find the projection that best fits a circle. The points tracing a suspected helix (pink) in a frame (x,y,z) = (a,b,c); the points are projected (black) onto the x-y plane (blue disk) and are then fit to a circle using SVD. (**a**) A helix is projected on the x-y plane, with its frame defined by spherical coordinates (45°,90°). (**b**) The helix is rotated, its coordinates projected on x-y plane, and fit again; spherical coordinates (45°,45°). (**c**) After all rotations are complete, the rotation whose projection best fit a circle is the helix frame; spherical coordinates (0°,0°).

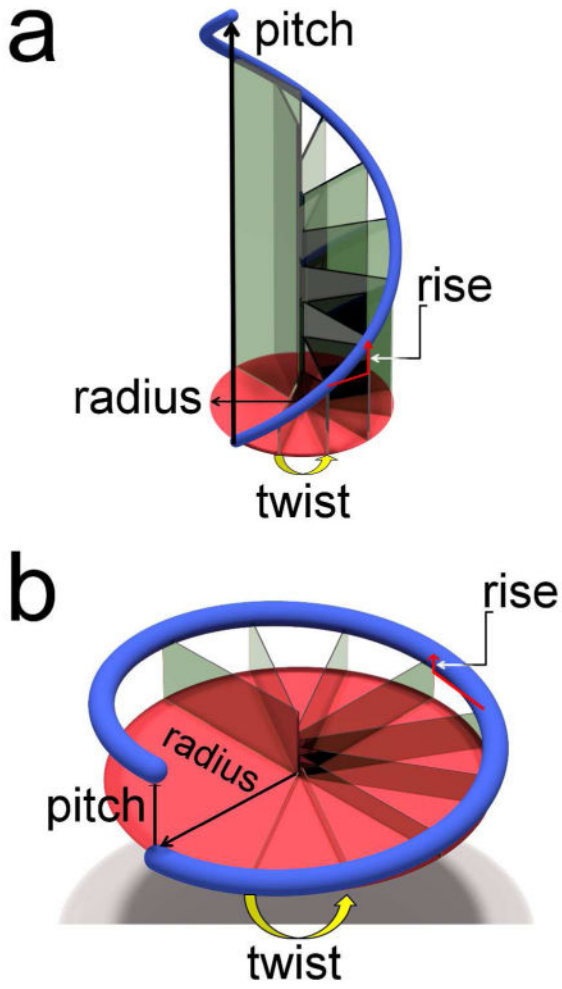

Figure 2.

A helix and its parameters. A helix (blue) whose pitch is (**a**) greater than its radius or (**b**) smaller than its radius. The red disk represents the *x*-*y* plane. Vertical planes (grey) represent steps whose height is the *z*-component of the point and position is the *x*- and *y*-component of the point (i.e. radius^2 = x^2 + y^2). Twist is the angle between successive points. The increase in height between successive steps is rise. Pitch is the height of the helix for one (360°) revolution.

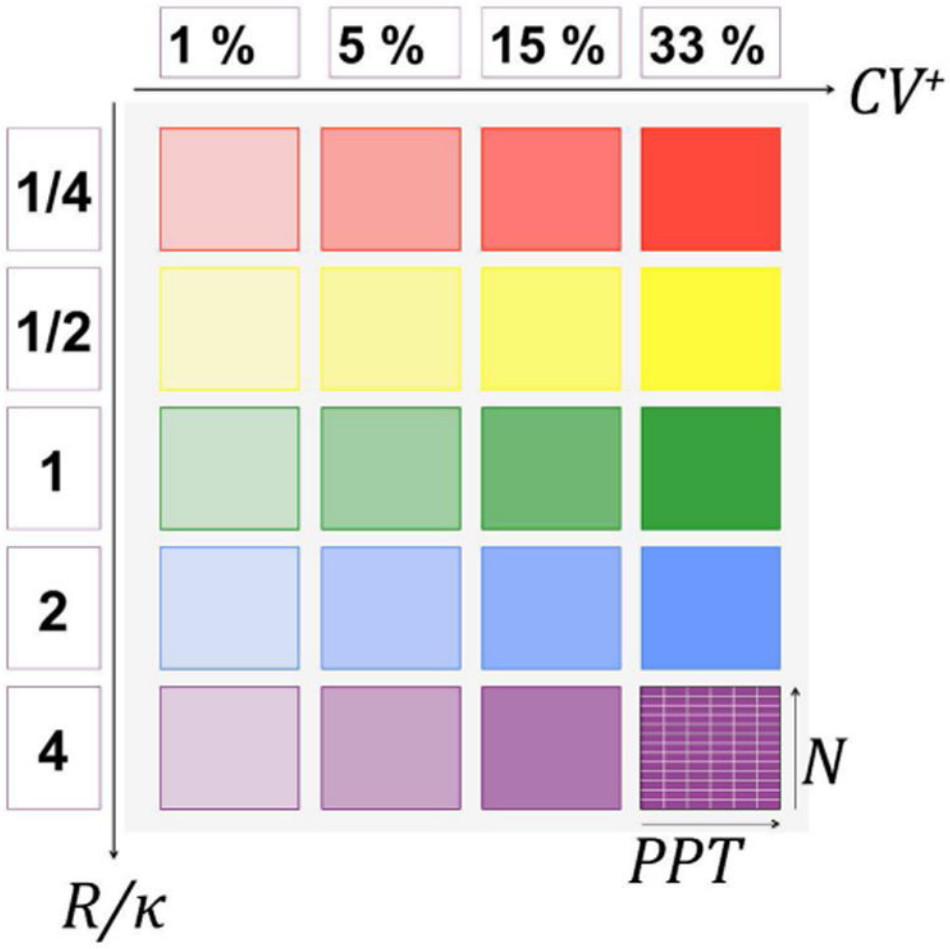

Figure 3.

Overview of the test set of noisy helices, with varying shapes (R/κ), degrees of applied noise (CV^+), number of points per helix (N) and points per helical turn (PPT). Five R/κ shape ratios (rows of one color) and four CV^+-noise levels (columns with different colors) provided 20 shape-noise helix-matrices (each colored square depicts one helix-matrix). Each helix-matrix contains varying numbers of points per helical turn and total number of points in the helix. Each cell in this matrix (6 columns of PPT, 13 rows of N, 78 total cells) represents 256 different random perturbations of the helix. 399,360 total helices were evaluated below: 256 helices per cell, 78 cells per shape-noise matrix, 20 shape-noise matrices.

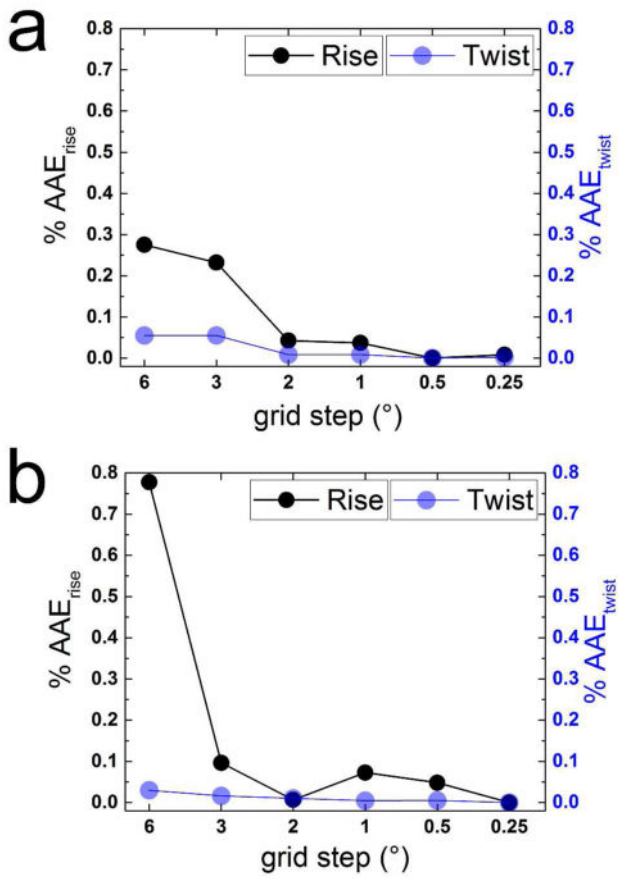

Figure 4.

The effect of scan resolution on accuracy measured by percent AAE (%AAE) of helical parameters for a set of 64 randomly oriented copies of a helix. Smaller %AAE values indicate less sensitivity to random orientation. (**a**) An ideal helix characterized using six scan resolutions (grid steps) of 6°, 3°, 2°, 1°, 0.5° and 0.25° (X axis). The reference rise and twist were 1 and 60, respectively. The Y axes show the measured %AAE for rise and twist. Black and blue symbols represent the measured %AAE of rise and twist respectively. (**b**) The coordinates of the ideal helix shown in panel (**a**) with random Cartesian coordinate perturbations applied to mimic the structure of an irregular, noisy helix.

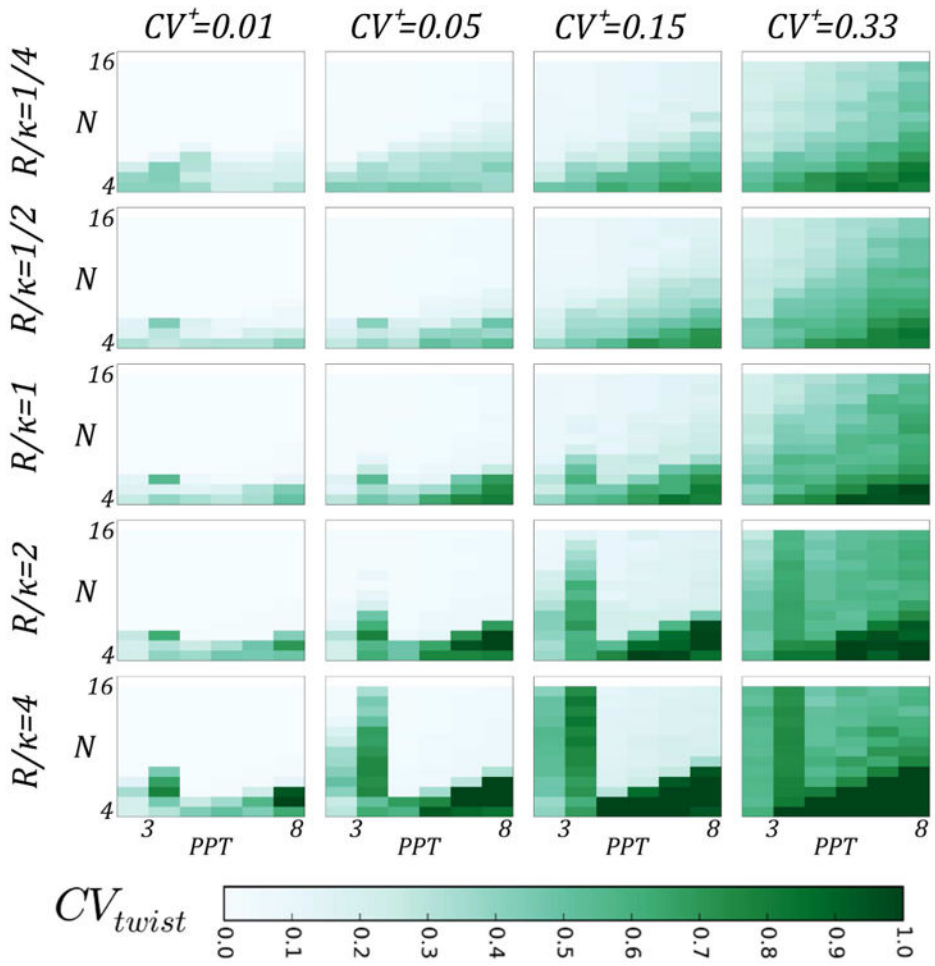

Figure 5.

Twenty heat-maps of derived CV_{twist} for noisy helices with diverse geometries. Families of helices with five radius-pitch ratios, R/κ, are shown, each of which contains a matrix of helices with varying points/turn, *PPT*, and varying numbers of points, *N*. For a given *N*, increasing *PPT* decreases the number of turns present. For each shape geometry (colored cells, e.g. $R/\kappa = 1/4$, $CV^{+} = 0.01$ $N = 16$, $PPT = 3$, the top left-most cell), 256 irregular helices were generated subject to Cartesian coordinate perturbations with CV^{+}, 0.01, 0.05, 0.15 and 0.33. If a cell in a helix-matrix is light-colored (white), the measured CV_{twist} is low and the method precisely characterizes helix twist. However, if a cell is dark-colored (green), the measured CV_{twist} is large and the method does not precisely characterize helix twist; the method is susceptible to noise.

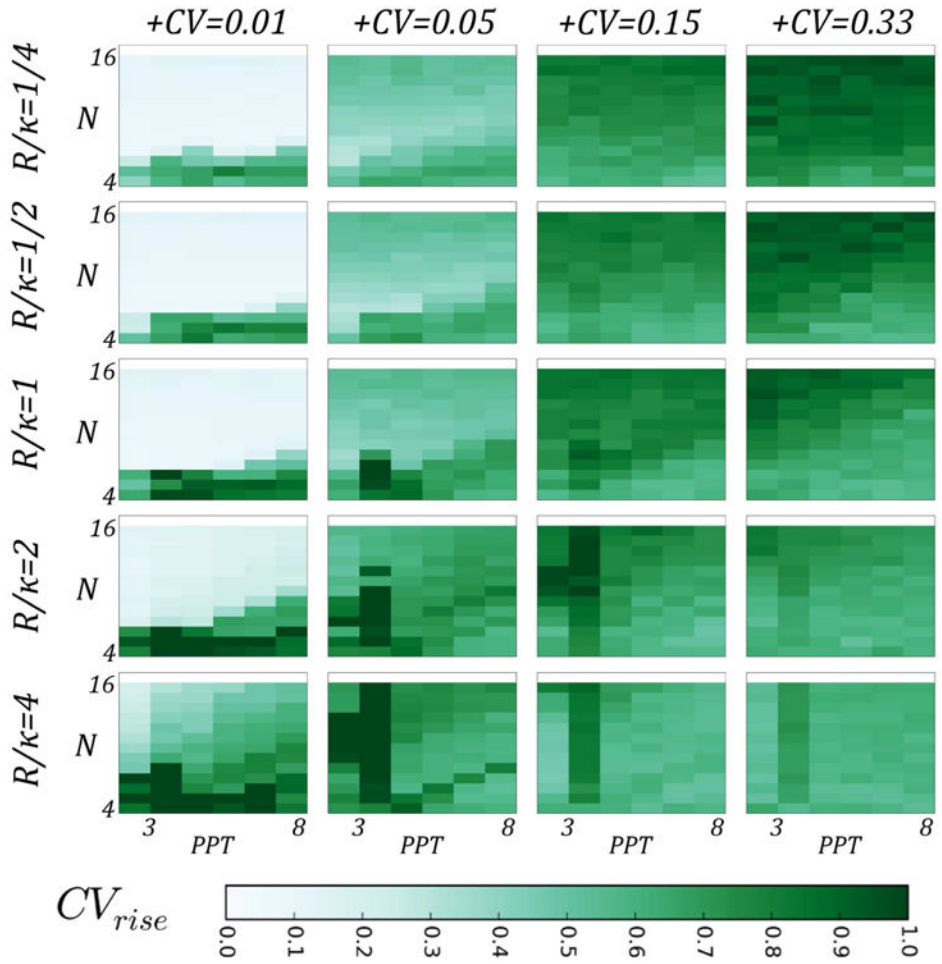

Figure 6.

Twenty heat-maps of derived CV_{rise} for noisy helices with diverse geometries. The helices analyzed and the layout of the data are the same as in Figure 5.

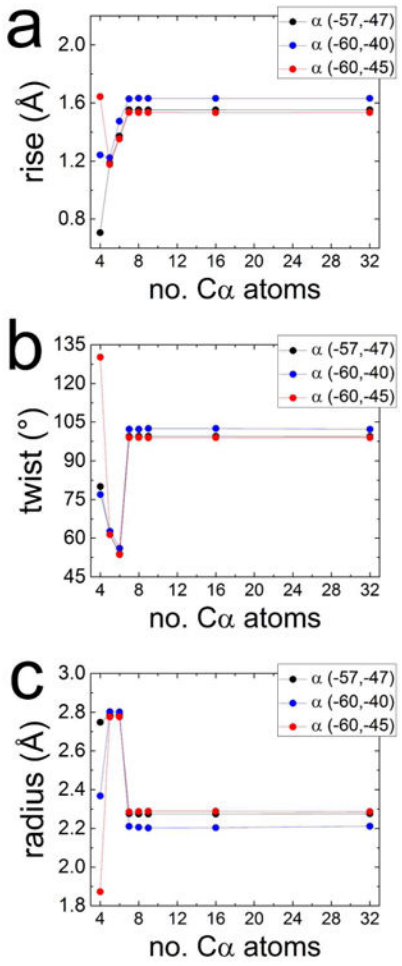

Figure 7.

Helical parameters of α-helical secondary structure elements defined using three different pairs of φ/ψ backbone torsions (black, blue and red symbols represent $\alpha_{-57,-47}$, $\alpha_{-60,-40}$ and $\alpha_{-60,-45}$ respectively), and an increasing number of Cα atoms used in the fitting (X axis). Shown on the Y axes are the derived values for (**a**) rise; (**b**) twist; (**c**) radius.

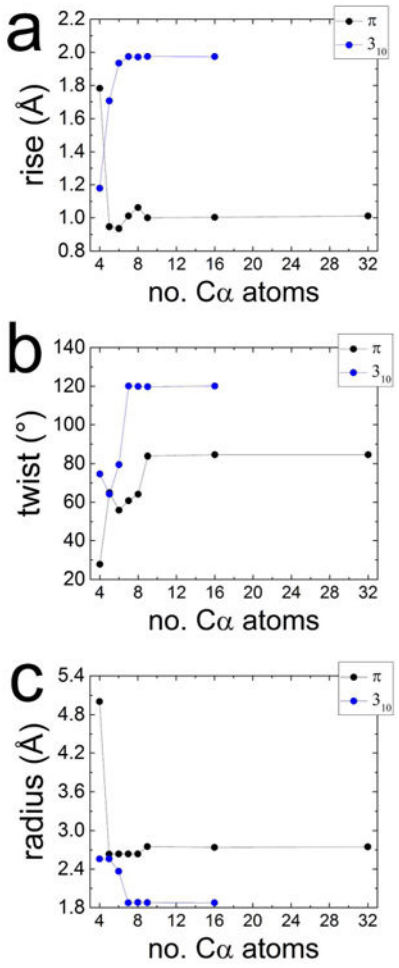

Figure 8.

Helical parameters of **π**- and 3_{10}- helical secondary structure elements, black and blue symbols respectively, with an increasing number of Cα atoms used in the fitting (X axis). Shown on the Y axes are the derived values for (**a**) rise; (**b**) twist; (**c**) radius.

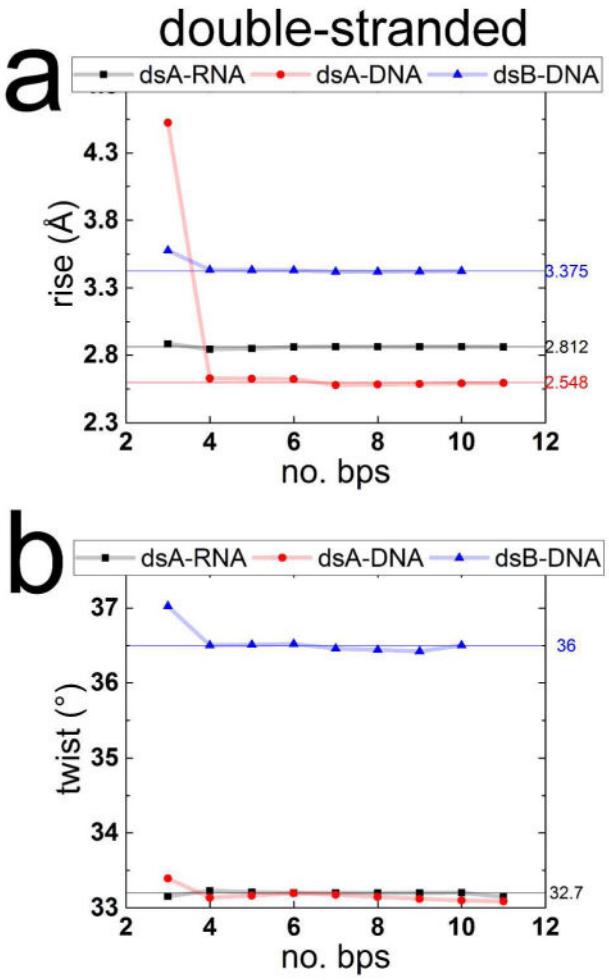

Figure 9.

Helical parameters of double-stranded nucleic acids using atoms from both strands. The X axis shows the number of atoms used in the fitting. Results for dsA-RNA, dsA-DNA and dsB-DNA are shown as black, red and blue symbols respectively. (**a**) Helical rise, with horizontal lines showing the reference values[10] for dsA-RNA (black line), dsA-DNA (red line) and dsB-DNA (blue) line. (**b**) Helical twist, with horizontal lines showing the reference values[10] for dsA-RNA (black line), dsA-DNA (red line) and dsB-DNA (blue) line. Two atoms per bp were used in the fitting.

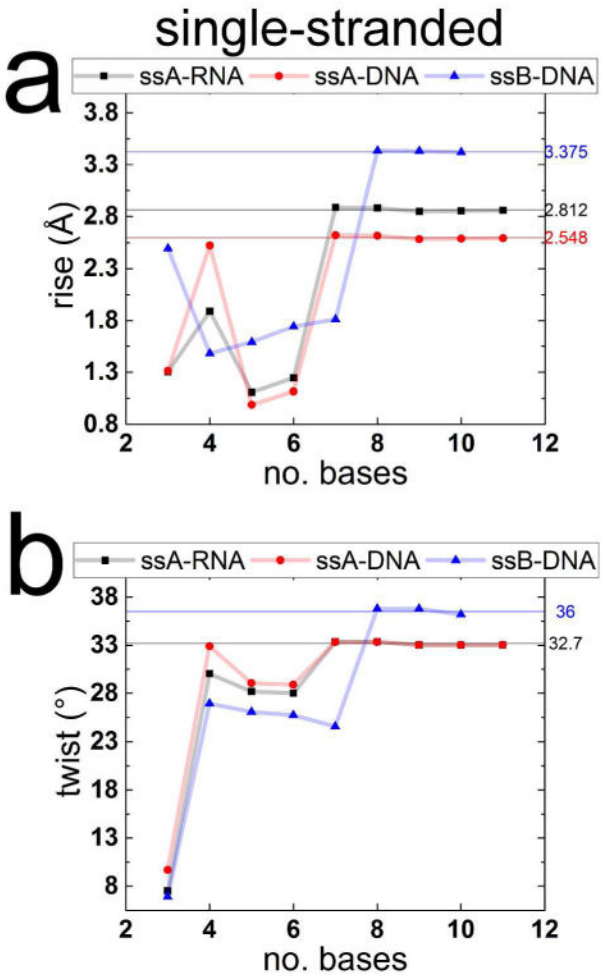

Figure 10.

Helical parameters of single-stranded nucleic acids using atoms from both strands. The X axis shows the number of atoms used in the fitting. Results for ssA-RNA, ssA-DNA and ssB-DNA are shown as black, red and blue symbols respectively. (**a**) Helical rise, with horizontal lines showing the reference values[10] for ssA-RNA (black line), ssA-DNA (red line) and ssB-DNA (blue) line. (**b**) Helical twist, with horizontal lines showing the reference values[10] for ssA-RNA (black line), ssA-DNA (red line) and ssB-DNA (blue) line.

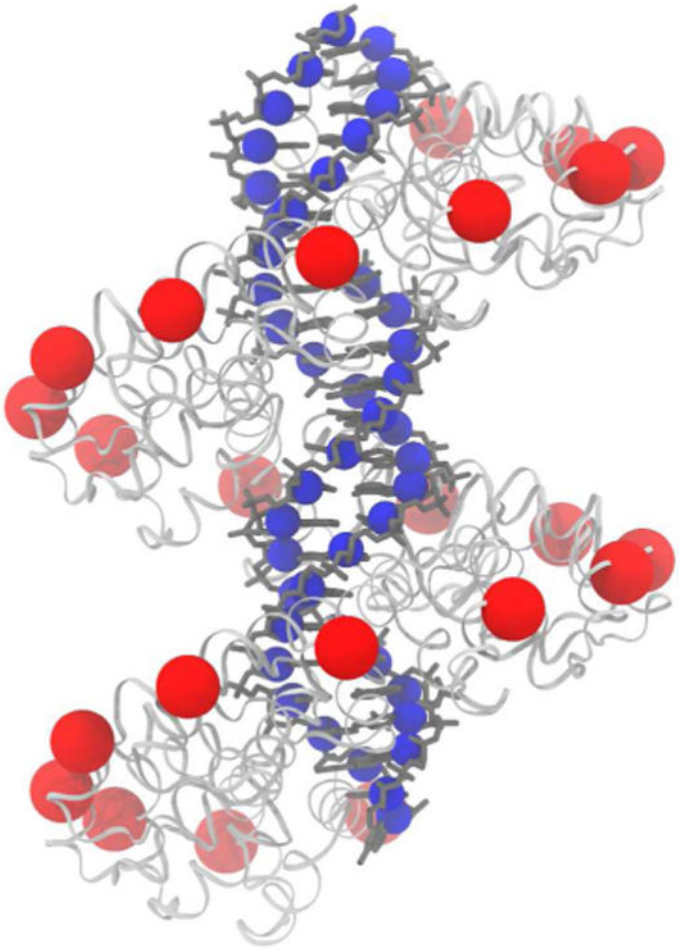

Figure 11.

Helix complementarity in the BurrH-DNA complex (PDB ID: 4CJA[22]). Protein (light grey ribbons), superhelical Cα atoms (red spheres), DNA (dark grey sticks) and C1' atoms (blue spheres).