Computers Chem. Vol. 13, No. 3, pp. 185–189, 1989
Printed in Great Britain. All rights reserved

0097-8485/89 $3.00 + 0.00

DEFINING THE AXIS OF A HELIX*

PETER C. KAHN
Department of Biochemistry and Microbiology, 328 Lipman Hall, Rutgers University, New Brunswick, NJ 08903, U.S.A.

(*Received* 22 *April* 1988)

Abstract—A simple method for finding the axis of a helix is presented. Although described in terms of protein alpha helices, the method is generally applicable to helices whose defining units are more or less regularly spaced. Absolute mathematical regularity of the helical parameters is not required. The procedure can be applied at several levels of rigor. At its simplest it involves simple vector operations only and yields two points that define the axis along with the axis direction cosines. No regression analysis is needed. The minimum length helix required for the algorithm is four residues, which define an axis segment between residues 2 and 3. In its more rigorous form the algorithm scans along the chain one residue at a time, yielding a set of axis segments to which a linear least squares regression fits the axis. Non-linear, iterative procedures are not necessary. For each consecutive set of four residues the pitch, radius, rise per residue, rotation per residue about the axis and number of residues per turn are also obtained. Methods are also outlined for distinguishing between smoothly curved and sharply kinked helices and for extracting from these structures the radius of curvature and the location and angle of the kink.

INTRODUCTION

The recent publication by Aqvist (1986) of a method to find the axis of an alpha helix presents an excellent review of the problem. Aqvist's method involves fitting a cylinder to the helix by means of a non-linear least squares regression. Such regressions are iterative and require initial estimates of the parameters in order to converge. Although Aqvist points out that his method is insensitive to the input estimates and converges rapidly, non-linear regressions are inherently more time consuming and complicated than linear regressions, and considerably more so than estimations that do not involve regressions at all.

Offered here is a method whose level of rigor can be adjusted as needed. At its simplest it finds two points that define the axis by using simple vector operations only. At its most accurate, vector operations yield a set of points to which a linear least squares regression fits the axis. In addition, it yields the rise and angle of rotation per residue and the radius of the helix and provides rigorous measures of local distortions from perfect helical geometry, which, apart from the average overall radius, are not provided by the published method.

The minimum helix required by the algorithm consists of four repeating units. Although it is developed for protein alpha helices, other kinds of helices can be analyzed.

Methods are also outlined for distinguishing between smoothly curved and sharply kinked helices and for extracting the radius of curvature and the locations and angles of kinks.

METHOD

The geometry expressed in terms of convenient vectors is given in Fig. 1.

Construct vector **P1** from the origin to carbon alpha 2 (CA_2). From CA_2 construct vectors **A** and **B** to CA_1 and CA_3 respectively. The bisector of the angle defined by **A** and **B** will be perpendicular to the helix axis and will intersect it. The simplest way to find its direction is to take the vector sum of **A** and **B** and normalize it to 1A. Call the norm **V1**.

Repeat this process with any other CA as the center except the first or last one, which cannot be used because one of the vectors corresponding to **A** or **B** will be undefined. Let **P2** be the vector from the origin to the central CA of this trio and **V2** the normalized angle bisector.

Since **V1** and **V2** are both perpendicular to the axis, their cross product will have the direction of the axis. Normalize **V1** × **V2** and call it **H**.

Needed now is the distance along **V1** and **V2** to the axis. This could be taken as the standard helix radius, but that would involve using idealized helical parameters. We will obtain the radius directly from the data instead.

Let r be the radius and d the distance along the axis from its intersections with the extensions of **V1** and **V2**. Let **H1** and **H2** be vectors, as yet unknown, from the origin to these intersections, i.e. to the points on the axis opposite the central alpha carbon atoms of

[*] N.J. Agricultural Experiment Station Publication Nos. D-01405-5-88 and D-10535-2-88.

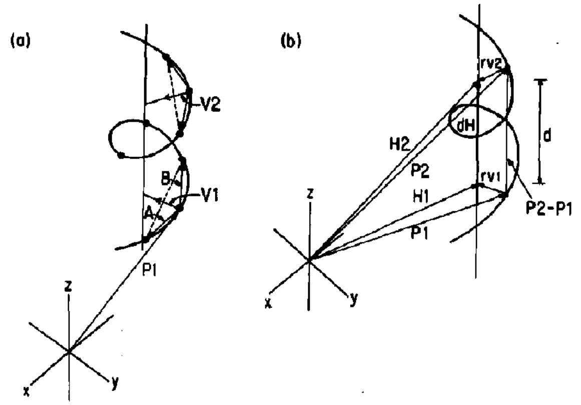

Fig. 1. Vector geometry of the helix axis. (a) Geometry of a trio of alpha carbons. A second trio is indicated also. (b) Angle bisectors and other vectors that define the axis. In the text vectors and their terminal points are designated by the same symbol. The context clarifies which is meant.

the two trios described above. It follows that

$$\mathbf{H1} = \mathbf{P1} + r\mathbf{V1} \qquad (1)$$

$$\mathbf{H2} = \mathbf{P2} + r\mathbf{V2} \qquad (2)$$

and

$$\mathbf{H2} = d\mathbf{H} + \mathbf{H1}. \qquad (3)$$

Substituting **H2** from equation (3) into equation (2):

$$\mathbf{H1} + d\mathbf{H} = \mathbf{P2} + r\mathbf{V2} \qquad (4)$$

and **H1** from equation (1) into equation (4) and rearranging:

$$r\mathbf{V1} + d\mathbf{H} = \mathbf{P2} - \mathbf{P1} + r\mathbf{V2}. \qquad (5)$$

It is clear that each side of this equation represents the vector from point **P1** to point **H2**, the right being the path through point **P2** and the left the path through **H1**. Therefore

$$|r\mathbf{V1} + d\mathbf{H}|^2 = |(\mathbf{P2} - \mathbf{P1}) + r\mathbf{V2}|^2. \qquad (6)$$

When this is expanded in Cartesian coordinates and solved for r, we take advantage of the facts that **V1** and **V2** are normalized to length 1 Å and that $d\mathbf{H}$ and **V1** are perpendicular, making their scalar product zero. We obtain

$$r = \frac{|d\mathbf{H}|^2 - |(\mathbf{P2} - \mathbf{P1})|^2}{2|(\mathbf{P1} - \mathbf{P2})\cdot\mathbf{V2}|}. \qquad (7)$$

To solve equation (7) one needs d, which is obtained as the projection of (**P1** − **P2**) on **H**, since **V1** and **V2** are skew lines, i.e. lines in parallel planes. Thus d is the perpendicular distance between lines perpendicular to the axis and cutting the helix through the carbon alphas at **P1** and **P2**:

$$d = (\mathbf{P2} - \mathbf{P1})\cdot\mathbf{H}. \qquad (8)$$

Back substitution of d from equation (8) into equation (7) yields r, and substitution of r into equations (1) and (2) yields two points on the axis, one opposite the central CA of each trio.

For a rough estimate of the axis one might use the second and penultimate alpha carbons as **P1** and **P2** (but see below). A more rigorous estimate can be had by using CA_2 and CA_3 as trio centers, then CA_3 and CA_4, scanning along the helix by one residue at a time to the penultimate CA. This will produce a set of short axis segments. Taken together these comprise one point opposite CA_2, one opposite CA_{n-1} and two points opposite each intervening CA. The least squares line through them will give the best fit to the axis. Simple, rapid methods to obtain the least squares line in three dimensions are presented elsewhere (Kahn, 1989).

Should the helix be kinked one can detect it by examining the angle between suitably chosen axis fragments. To determine whether a kink is present one could choose as vertex of the kink angle the mean of the two segment end points opposite any CA with the exception of the axis ends. Least squares lines from the vertex to the ends of the axis could then be computed and the angle at the vertex calculated. By moving the vertex along the axis from a point near one end of the helix to a point near the other end, one can find the location of the bend by looking for the smallest angle.

As discussed in the companion paper (Kahn, 1989), one of the linear least squares methods for fitting the axis to the segments yields also the plane that best fits the data. If the helix is smoothly curved or kinked, the plane thus found is the plane containing the curvature. The radius of curvature can be obtained in that plane by standard methods. To distinguish between smooth curvature and a sharp kink, one would obtain the location and angle of the kink as described above as well as the radius of curvature. Tests of the r.m.s. deviation of the data from the smooth curve and from the two lines

Table 1. VAX 8800 CPU time for program versions to calculate axes for all eight helices of myoglobin*

Version No.	CPU time (s)
1	0.045
2	0.072
3	0.18 (0.586†)
4a	0.19‡ (0.452†)
4b	0.22‡ (0.596†)

* The eight helices of myoglobin range from 7 to 26 residues in length with mean and SD 15.1 ± 6.7 residues. The individual lengths are given in Table 3.

† These times include calculations at each tetrad of the local pitch, radius, rise per residue, angle of rotation per residue and the standard deviations in each of these for each helix. The times also include computation of the r.m.s. deviations of the axis segments from the axis lines as well as output time for all these results. Versions 1 and 2 do not contain these calculations and their output. The code has not yet been fully optimized.

‡ Comparison of the CPU times for the least squares algorithms is discussed in greater detail separately (Kahn, 1989).

defining the kinked structure would then distinguish between them.

If successive sets of four alpha carbons are used as described above, the values of d and r that are obtained are the rise per residue and helical radius respectively. The angular rotation per residue about the axis, the number of residues per turn, and the pitch can be obtained easily and complete the description of the helix. The cosine of the angle of rotation is simply $\mathbf{V1}\cdot\mathbf{V2}$. The number of residues per turn is therefore obtained by dividing 360° by $\cos^{-1}$ $(\mathbf{V1}\cdot\mathbf{V2})$, and the pitch by multiplying the number of residues per turn by d from equation (8).

CODING AND TESTING

To test the procedure described here, it has been coded in VAX FORTRAN in four versions. All versions use variables whose names are the same as those in the description above. The programs read an entire crystallographic data file from the Protein Data Bank (PDB) at Brookhaven National Laboratory (Bernstein *et al.*, 1977) and compute axes and related output for all helices in the protein. Sperm whale myoglobin (PDB file 3MBN) has been used as an example. The protein has eight helices of varying lengths and degrees of distortion, and thus provides a good test. The first and last residues of each helix were taken from the HELIX records in the PDB file, i.e. the helix boundaries were as determined in the crystallographic study (Takano, 1977).

Version 1 is the simplest. It uses three points at each end of a helix to define the axis. It returns two points on the axis, one opposite residue 2 and the other opposite residue $(n-1)$. Version 2 uses the tetrad of consecutive residues at each end of a helix to define two axis segments, one opposite residues 2 and 3 and one opposite residues $(n-2)$ and $(n-1)$. Like Version 1 it returns the axis defined by the line between the segments ends opposite CA_2 and $(n-1)$. Versions 3 and 4 are similar in that they both step along the helix, using each possible tetrad of consecutive residues as described above. Version 3 obtains the direction of the final axis by averaging the direction cosines of its segments. Versions 4a and 4b find the direction of the axis by least squares. Versions 3 and 4 return axis ends for each protein helix as do Versions 1 and 2. Versions 3 and 4 also produce for each helix the root mean square (r.m.s.) deviations of the segments from the final axis as well as the means and standard deviations of the rise and rotation per residue, the radius, and the pitch at each point along the axis. Versions 4a and 4b differ in the least squares method used to fit the axes to the segments. A simplified algorithm is used in 4a and an eigensystem solution in 4b (Kahn, 1989). Eigenvalues and eigenvectors are obtained with the International Mathematical and Statistical Libraries (IMSL) routine EVCSF. The VAX 8800 CPU times needed to compute the axes for the eight helices of myoglobin are given for the various versions in Table 1.

In Table 2 the angles between the axes a determined by Version 4b and those found by the other versions are presented. Table 3 compares the r.m.s. deviations of the axis segment ends generated by Versions 3 and 4 (the segments are the same for those versions that generate them) with the overall axis of each helix in each program version. The distances computed to construct Table 3 are the perpendicular distances of the segment end points from the various axes (Osgood & Graustein, 1921).

Table 2. Angles between least squares (Version 4b) axes of myoglobin (in degrees) and those produced by the other versions

Myoglobin helix	A	B	C	D	E	F	G	H
Version 1	21.7	52.5	6.4	4.9	4.5	9.2	8.8	19.7
Version 2	2.0	1.0	0.5	0.3	0.8	1.3	0.5	0.5
Version 3	3.5	1.9	2.8	1.6	1.1	5.6	1.7	2.9
Version 4a	1.3	0.1	0.0	0.0	0.2	0.0	0.1	0.2

Table 3. Root mean square deviations of myoglobin axis segment ends computed by Versions 3–4b* from the overall axes of different program versions

				(Angströms)				
Myoglobin helix	A	B	C	D	E	F	G	H
Helix length (residues)	16	16	7	7	20	10	19	26
Version 1	7.34	24.78	0.71	1.18	1.29	57.05	2.83	11.35
Version 2	0.98	0.23	0.12	0.10	0.65	0.18	0.28	1.39
Version 3	0.57	0.24	0.14	0.08	0.34	0.33	0.26	0.76
Version 4a	0.48	0.14	0.08	0.06	0.31	0.13	0.17	0.57
Version 4b	0.46	0.14	0.08	0.06	0.31	0.13	0.17	0.57

* Versions 3, 4a and 4b all produce the same set of axis segments. They differ in the fitting of the axis to the segments (see text).

DISCUSSION

When the axes are examined visually on a graphics system, it is seen immediately that those produced by Version 1 fit well in some cases but very poorly in others, especially helices A, B, F and H. In the latter group at least one end of the axis is actually outside the helix itself. The ends of protein helices are frequently distorted, and those distortions can lead to large errors when the minimal algorithm is applied to long helices. Version 1, although simple and computationally efficient, should therefore be used only with helices known to be regular, especially at the ends. Upon initial examination of a protein, one seldom knows this, however, which reduces the usefulness of Version 1 for protein work.

Although minor distortions or data errors at the helix ends propagate into large axis errors in Version 1, the errors are much smaller over short sequences of four consecutive alpha carbons. Version 2, which uses one such tetrad at each end to compute two completely independent axis segments, produces much better axes. Their quality can be judged from Tables 2 and 3. Table 2 gives the angles between the axes produced by all the versions and those produced using least squares Version 4b as a standard, while Table 3 presents the r.m.s. deviations of the Version 4 axis segments from the axes of the other versions. The Version 2 directions are certainly good enough for approximate comparisons, for example, of the angle between one helix and another in a protein. For some helices, however, although the Version 2 directions are quite good, the locations in space of the axes defined as the lines between the axis ends, are less so, as shown by the r.m.s. deviations in Table 3.

The angular deviations of the Version 3 axes from those of Version 4 are actually a little larger than those of Version 2, particularly for helices A, E and H, whereas the r.m.s. distance errors are either smaller or the same. As the CPU time (Table 1) for Version 3 is approximately twice that for Version 2, however, one must choose carefully what features are important for one's application.

The results of Versions 4a and 4b are very similar. Given the precision of crystallographic coordinates, the differences between them are probably not significant.

Version 4a requires a small increase in CPU time over Version 3 (Table 1) and yields slightly smaller r.m.s. deviations in every case (Table 3). Where CPU time is not limiting, Version 4 is to be preferred, as it is the most accurate. It should be emphasized, however, that for applications in which CPU time is an important consideration, the results of Version 2 are probably good enough.

At its most stringent the procedure described here involves a linear least squares regression (Versions 4a and 4b), which is simpler and faster to compute than the nonlinear regression of Aqvist (1986). If a somewhat less precise estimation of the axes is all that is needed, it is even simpler, as a few vector operations (Version 2) will suffice. The method is restricted to helices whose repeating units are more or less regularly spaced, however, although absolute mathematical regularity is not required. All protein helices fall into this category. For helices whose defining units are highly irregularly spaced, Aqvist's method is obligatory.

The rise per residue, pitch, radius and angle of rotation about the axis per residue that describe each axis segment is computed by Versions 3 and 4 are calculated at the points H1 and H2, which are opposite residues 2 and 3 of a tetrad of alpha carbons. They are useful indicators of departures from idealized helical geometry. It should be remembered, however, that they are averaged properties of the tetrad as a whole. Although it will be convenient to refer to the location of such distortions as at residues 2 and 3, it should be understood that this is a short-hand for the defining tetrad.

One application of the algorithm concerns the delineation of helical boundaries in interpreting X-ray crystallographic data. By stepping a tetrad of consecutive residues through the entire protein one residue at a time and displaying the helical segments on a graphics terminal, we find in preliminary calculations that the ends of each helix are easily recognized. At those points the curve traced by the segments usually changes direction abruptly. In any case tetrads for which the rise, rotation and radius do not all fit suitable norms are objectively defined as nonhelical. The use of the segments to analyze protein alpha helical geometry in greater detail is in progress.

Lastly, application of the algorithm to DNA helices should be undertaken with caution. With perfectly formed helices having the helical parameters of B-DNA it works perfectly. However, when a crystallographically determined DNA of 12 bases, PDB file 7BNA, was examined using the ribose Cl′ atoms to define the helix, the short axis segments did not approximate a line very well at all. The helix is distorted. B-DNA has 10 residues per turn whereas protein alpha helices have 3.6. The four consecutive residues used in the more rigorous versions of the algorithm thus encompass approximately one full turn in the protein case but considerably less than a turn of DNA. Small distortions from perfect helical geometry appear to have much less effect in the former case than in the latter. This was tested by adding to the coordinates of a perfectly formed DNA helix normally distributed random noise. The mean of the noise distribution was zero and the SD 0.1 Å. This amount of noise is so small that it is difficult to see the difference between the original and the perturbed helices upon visual examination. Nevertheless, the helial radius, rise, and angle of rotation were grossly variable over the length of the helix. The problem was remedied by using as the first defining tetrad residues 1, 3, 5 and 7, as the second 2, 4, 6 and 8, etc. Both the helical parameters and the direction and position of the axis were restored to their proper values. When this procedure was applied to the crystallographically determined coordinates of the Cl′ atoms of the B-DNA mentioned above, the helical parameters were approximately constant over the helical length, the axis segments did approximate a line, and the final overall axis fit the structure quite well upon visual examination.

Acknowledgements—Much of the work for this paper was done during sabbatical leave in the laboratory of Professor F. M. Richards. For making available the facilities of his laboratory and particularly for many stimulating discussions with him, the author is especially grateful. Useful and extensive discussions with Drs J. Hartigan and B. Bush are also gratefully acknowledged. The idea that the triad angle bisector in Fig. 1(a) passes through the helix axis arose in the course of a conversation with C. Levinthal. This work was supported by NIH Grant GM 22778 to F. M. Richards at Yale University, by the Center for Advanced Food Technology of Rutgers University, an initiative of the New Jersey Commission on Science and Technology, and by the New Jersey State Agricultural Experiment Station.

REFERENCES

Aqvist J. (1985) *Computers Chem.* **10**, 97.

Bernstein F. C., Koetzle T. F., Williams G. J. B., Meyer E. F., Brice M. D., Rodgers J. R., Kenard O., Shimanouchi T. & Tasumi M. (1977) *J. Mol. Biol.* **112**, 535.

Kahn P. C. (1989) *Computers Chem.* **13**, 91.

Osgood W. F. and Graustein W. C. (1921) *Plane and Solid Analytic Geometry*, pp. 514–515. Macmillan, New York.

Takano T. (1977) *J. Mol. Biol.* **110**, 537.