Curve Encirclement and Protein Structure

Patrick Kessler* and Oliver M. O'Reilly
Department of Mechanical Engineering, University of California, Berkeley, 94720-1740
(Dated: July 20, 2007)

A method of measuring the extent to which space curves encircle one another is introduced. The method provides a family of sets which characterize encircling curves, allowing curve pairs that engage, (and also single curves that self-engage) to be distinguished. The method is applied to the backbone chains of protein molecules.

PACS numbers: 02.40.-k, 02.40.Ma, 87.14.Ee, 87.15.By
Keywords: physical knots, curve geometry, protein conformation

Introduction.– This work is motivated by the question of whether curves in $\mathbb{R}^3$ wrap around or encircle one another. The concept of *encirclement* has many practical applications (e.g., vines and other biological fibers), however there is no clear way to describe it, and no rigorous method for distinguishing between the many possible kinds of encircling curves, for instance between a curve segment that coils around another *tightly* and one that coils around another *loosely*. Neither the linking number of two closed curves nor the writhing number [1] of a single closed curve help with this question. Familiar local quantities (e.g., Serret-Frenet curvature and torsion), and global quantities (e.g., writhe, normal injectivity radius, and global radius of curvature [2]), don't appear to help either. Our question relates to *structural complexity* [3], which is an emerging area in the study of turbulent flow geometries.

We find success with a new approach which features the intersection of one space curve with triangles that have vertices on a second space curve. These intersections correspond to a set of points in $\mathbb{R}^2$ which is easy to visualize and interpret. We call this set an *encirclement set*, and use it to identify structural relationships between the curves. We find that the encirclement set is also meaningful for a single curve; in this case it highlights structural relationships between different pieces of the same curve. We introduce a distance function on the collection of all encirclement sets which allows for a systematic comparison of different space curves.

We apply our approach to the backbone chains of protein molecules, over 40,870 of which are cataloged in the publicly accessible Protein Data Bank (PDB). The categorization of these chains (which qualify as *physical knots* [4, 5]) is an active area of research [6–8]. Presently, one of the leading strategies for organizing proteins is based on a comparison algorithm called DALI [9], which works with the matrix of pairwise distances between atoms in a protein chain. We find that protein encirclement sets give a way to distinguish between different proteins, and so we hope to use these as an alternate basis for protein comparison and classification.

Encirclement.– Let $\mathbf{x}$ and $\mathbf{y}$ be arc-length parametrized curves in $\mathbb{R}^3$, as illustrated in Figure 1. Pick a point $\mathbf{x}(s)$ on $\mathbf{x}$, and a scale value $d > 0$ such that $\mathbf{x}(s-d)$ and $\mathbf{x}(s+d)$ are defined. The points $\mathbf{x}(s-d)$, $\mathbf{x}(s)$, and $\mathbf{x}(s+d)$ are the vertices of an open triangle in $\mathbb{R}^3$ that we call the (d-scale) *encirclement triangle* based at s. If the curve $\mathbf{y}$ intersects this triangle at the point $\mathbf{y}(t)$, then the pair (s,t) is said to be an element of the (d-scale) *encirclement set* E_d of $\mathbf{x}$ about $\mathbf{y}$. Elements of an E-set (short for encirclement set) almost always comprise a finite collection of open curves (called *strands*) in the s-t plane. We identify the *sign* of an E-set element $(s,t) \in E_d$ with the sign of the scalar triple product $[\mathbf{x}(s-d) - \mathbf{x}(s),\ \mathbf{x}(s+d) - \mathbf{x}(s),\ \mathbf{u}(t)]$, where $\mathbf{u}(t)$ is the unit tangent vector at $\mathbf{y}(t)$.

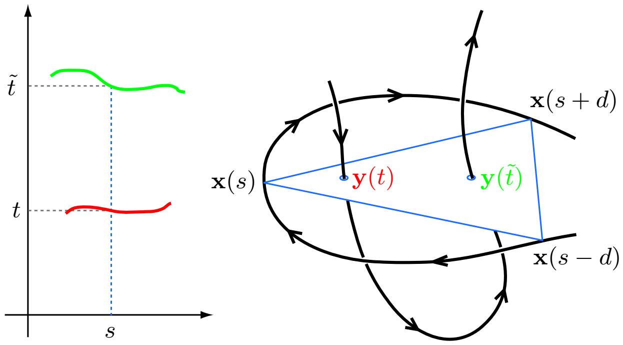

FIG. 1: (In color online.) A d-scale E-set (left) shows the encirclement of space curves (right). The point s on the E-set x-axis corresponds to the point $\mathbf{x}(s)$ on the space curve $\mathbf{x}$. The triangle with vertices $\mathbf{x}(s)$, $\mathbf{x}(s-d)$, and $\mathbf{x}(s+d)$ intersects the space curve $\mathbf{y}$ at $\mathbf{y}(t)$ and at $\mathbf{y}(\tilde{t})$, and so (s,t) and $(s,\tilde{t})$ are E-set elements. The encirclement at $(s,\tilde{t})$ is positive, and the encirclement at (s,t) is negative.

The structure of two space curves is reflected in the shape of their corresponding E-set strands. Consider a curve $\mathbf{x}$ that coils many times around the line segment $\mathbf{y}$. If for all points along $\mathbf{x}$ the radius of curvature of $\mathbf{x}$ is roughly r and the center of curvature of $\mathbf{x}$ is roughly at the same point on $\mathbf{y}$, then the $2r$-scale E-set of $\mathbf{x}$ about

$\mathbf{y}$ will contain a strand that is flat. In contrast, if the centers of curvature of $\mathbf{x}$ move along $\mathbf{y}$ (as they do for the red stripes wrapping around the axis of a candy cane), then the corresponding E-set strand will be tilted. If $\mathbf{x}$ coils tightly around $\mathbf{y}$ (that is, if many coils occur over a short length of $\mathbf{x}$), then the minimum scale d at which this coiling is reflected in the E-set of $\mathbf{x}$ about $\mathbf{y}$ will be less than it would be if the coils in $\mathbf{x}$ were loose. Encirclement is also meaningful when the curves $\mathbf{x}$ and $\mathbf{y}$ coincide (see Figure 2); we refer to this as *self* encirclement.

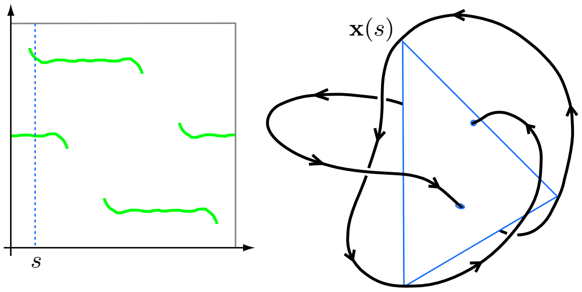

FIG. 2: (In color online.) Self-encirclement of a trefoil knot. The arc-length position of the triangle vertex $\mathbf{x}(s)$ corresponds to the dotted vertical line crossing the E-set x-axis at s. The E-set strands intersecting this dotted line correspond to the two points at which $\mathbf{x}$ intersects the triangle.

Although an E-set conveys information about curve structure, it does not contain enough information to reconstruct a curve. In many cases, different curves will have the same E-sets, for instance the self E-sets for a line segment and an arc of a circle are both empty for all values of the encirclement scale value d.

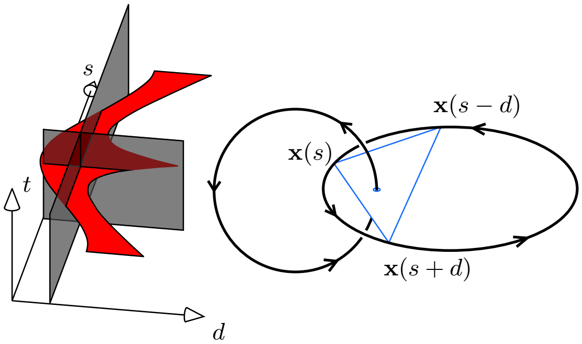

FIG. 3: (In color online.) E-sets are sections of a surface embedded in $\mathbb{R}^3$. Here we show this surface for two linked circles. The intersection of the surface with the two planes at left corresponds to the intersection of the triangle and curve at right.

E-sets are d-indexed cross sections of an open (often disconnected) surface embedded in a three dimensional space with coordinates s, t, and d (see Figure 3). Information about encirclement and curve shape can be obtained from an E-set, that is, from the intersection of this surface with a plane perpendicular to the d-axis. Interestingly, information about curve topology can be obtained from the intersection T of this surface with a plane *perpendicular* to the s-axis. Holding s constant and varying d corresponds to moving the edge of the encirclement triangle from $\mathbf{x}(s-d)$ to $\mathbf{x}(s+d)$ through a ruled surface with boundary $\mathbf{x}$. The sum of the signed intersections of this surface with $\mathbf{y}$ is the linking number of $\mathbf{x}$ and $\mathbf{y}$. These intersections correspond to a subset of the endpoints of strands in T, (like an E-set, T consists of a finite collection of open strands).

Comparing E-Sets.– Similar E-sets and E-subsets (as in Figure 4) correspond to similar space curve structures and substructures. Here we quantify the difference between two E-sets; in the next section we discuss using this quantification to organize protein molecules.

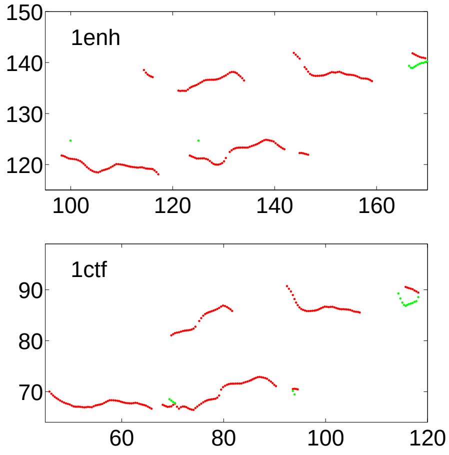

FIG. 4: (In color online.) Similar E-subsets from two different protein molecules; these images are magnifications of the boxed regions in Figure 5. The similarity in these E-subsets suggests a correspondence between the associated protein substructures. The axes indicate position (in Ångstroms) along the protein backbones.

Let A and B be two E-sets and let $f_B : A \longrightarrow \mathbb{R}$ map from $\mathbf{a} \in A$ to

$$f_B(\mathbf{a}) = \inf\{\|\mathbf{a}-\mathbf{b}\| \mid \mathbf{b} \in B\}. \qquad (1)$$

This function tells how far a single element of A is from all the elements of B. A distance function $D(A,B)$ is obtained by integrating (1) along the strands comprising A and B,

$$D(A,B) = \frac{1}{L_A + L_B}\left(\int_A f_B(\mathbf{a}(\tau))\, d\tau + \int_B f_A(\mathbf{b}(\tau))\, d\tau\right),$$

where L_A and L_B are the total lengths of the strands in A and B respectively, and where τ is an arc-length parameter.

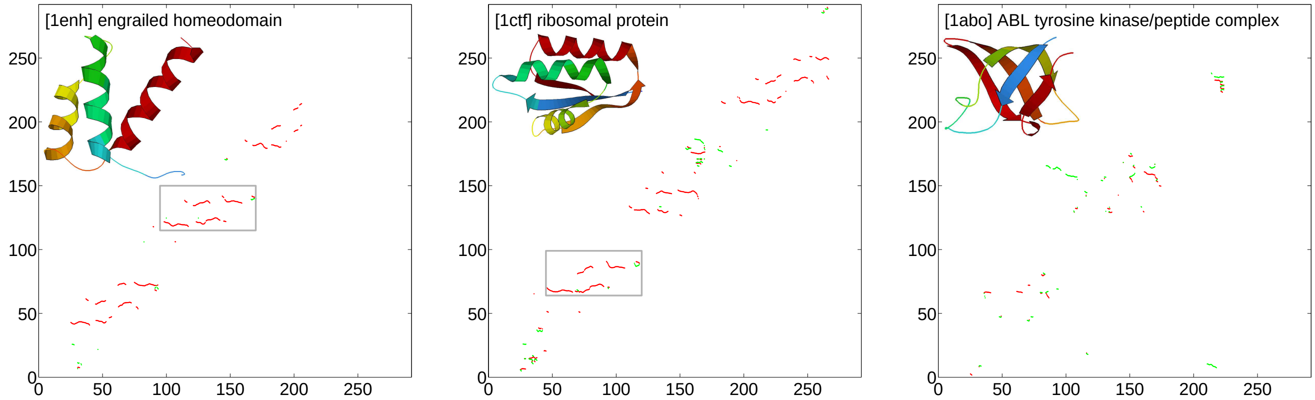

FIG. 5: (In color online.) Self encirclement sets for three different proteins, with $d = 25$Å. The axes indicate position (in Ångstroms) along the protein backbone. There is an especially good correspondence between the E-subsets boxed in gray, in that the strands comprising these subsets *look* the same, (more formally, the distance function D applied to these subsets returns a small number). These subsets are shown with greater magnification in Figure 4. (These protein images were generated with KiNG Version 1.39, see http://kinemage.biochem.duke.edu/)

Imagine partitioning the strands in A and B into small segments of equal length. To each segment in A (B), assign the distance from that segment to the nearest element in B (A). The function $D(A, B)$ simply returns the average of these distances. If A and B are unequal, then $D(A, B) > 0$. For instance, if A consists of the vertical line l_A from $(0, -1)$ to $(0, 1)$, and if B consists of the horizontal line l_B from $(-1, 0)$ to $(1, 0)$, then $D(A, B) = 1/2$. As A and B get closer together, e.g., as l_A rotates about the origin and becomes increasingly aligned with l_B, $D(A, B)$ gets smaller. If A and B are equal then $D(A, B) = 0$, and conversely.

We call D a *distance function* on E-sets because if each of the E-sets A and B is contained in a disk in the s-t plane, then $d_{min} < D(A, B) < d_{max}$, where d_{min} is the diameter of the largest circle that can pass between the disks, and where d_{max} is the diameter of the smallest circle that contains the disks. Unfortunately, although D is positive definite (and trivially symmetric), it is not a true metric because it sometimes violates the triangle inequality. For instance, when A, B, and C are E-sets consisting of the open intervals on $\mathbb{R}$ given by $(-1, 0)$, $(-1, 1)$, and $(0, 1)$ respectively, $D(A, B)+D(B, C) = 1/3$, and $D(A, C) = 1/2$.

A change in the (arc-length) parameterization of a space curve can translate and reflect its E-set in the s-t plane. When comparing E-sets (or E-subsets), D can be minimized over these translations and reflections. Also, D can be evaluated separately for the positive and negative parts of two E-sets.

Protein.– Protein molecules consist of chains of atoms coiled into compact hierarchically structured curves in $\mathbb{R}^3$. Each protein molecule in the PDB is given a unique alpha-numeric identifier, such as 1enh. The proteins with PDB identifiers 1enh, 1ctf, and 1abo and their self encirclement sets (with $d = 25$Å) are illustrated in Figure 5. The proteins with PDB identifiers 1enh and 1ctf both contain three α-helices, while the protein with PDB identifier 1abo contains none. The similarities and differences in these curve structures are reflected in their corresponding ($d = 25$Å) E-sets; the E-sets for these molecules at other scales also reflect these similarities and differences. Our hope is that this novel method of distinguishing between different protein structures (and substructures) will lead to a new and interesting organization of protein molecules.

Similar protein molecules usually have corresponding secondary structures that are interconnected differently. Finding which segments of one protein best correspond to which segments of another involves sifting through a huge number of possible pairings. One of the most successful algorithms for doing this is DALI [9], which uses a Monte Carlo approach to slowly assemble a collection of pairings for which an overall similarity score is high. In DALI, this similarity score is based on comparing the distances between curve nodes. We are currently developing an algorithm like DALI, but with a similarity score based on comparing encirclement subsets rather than distance matrices. We are considering weighted sums of E-set comparison scores D for E-sets over a range of scale values. Also, we anticipate having to recompute encirclement on larger scales as the collection of pairings gets bigger. Once we develop our E-set based similarity score, we will be able to organize protein molecules into trees, clusters, and other structural families [6, 7, 9].

Although protein self encirclement sets show distinctive patterns for different protein structures, protein chains generally do not experience the kind of encirclement that first motivated our investigation, in which one curve wraps many times around another. However, this kind of encirclement does occur between a protein chain and a smoother version of the chain (see Figure 6). Let $\mathbf{X}$ be a $3 \times N$ array with columns containing the xyz coordinates of the atoms comprising a protein's backbone chain, and let $A_k(\mathbf{X})$ denote the k^{th} spatial average of $\mathbf{X}$, defined by

$$\begin{aligned} &A_0(\mathbf{X}) = \mathbf{X}, \\ &[A_{k+1}(\mathbf{X})]_1 = [\mathbf{X}]_1, \ \ [A_{k+1}(\mathbf{X})]_N = [\mathbf{X}]_N, \\ &[A_{k+1}(\mathbf{X})]_i = \frac{[A_k(\mathbf{X})]_{i-1} + [A_k(\mathbf{X})]_{i+1}}{2}, \end{aligned}$$

for $i = 2, \ldots, N-1$. As k grows, $A_k(\mathbf{X})$ approaches the straight line connecting the endpoints of $\mathbf{X}$.

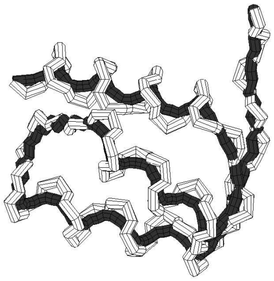

FIG. 6: The backbone $\mathbf{X}$ of the protein with PDB identifier **1enh** is shown in white, and the smoother spatial average $A_5(\mathbf{X})$ of this curve is shown in black. The encirclement of $A_5(\mathbf{X})$ by α-helices in $\mathbf{X}$ shows up clearly in the $d = 4$Å E-set for these curves.

For small values of k, $A_k(\mathbf{X})$ is encircled by the α-helices in $\mathbf{X}$. For larger values of k, $A_k(\mathbf{X})$ engages with the higher order coiling structures in $\mathbf{X}$ (e.g., barrels), and the corresponding E-sets reflect this with signature markings. We are working on incorporating the E-sets for a protein and its average into our E-set based similarity score.

The authors are grateful to Adam Arkin for his advice and suggestions concerning this work, and also to Tye Lidman for very helpful discussions about metrics.

* Corresponding Author: watchwrk@me.berkeley.edu

[1] F. Fuller, Proc. Nat. Acad. Sci. USA **68(4)**, 815 (1971).

[2] O. Gonzalez and J. Maddocks, Proc. Nat. Acad. Sci. USA **96**, 4769 (1999).

[3] R. Ricca, in *Encyclopedia of Nonlinear Science*, edited by A. Scott (Routledge, New York and London, 2005), pp. 885–887.

[4] J. Calvo, K. Millett, and E. Rawdon, eds., *Physical Knots: Knotting, Linking, and Folding Geometric Objects in* $\mathbb{R}^3$, vol. 304 of *Contemporary Mathematics* (American Mathematical Society, 2002), ISBN 0-8218-3200-X.

[5] J. Calvo, K. Millett, E. Rawdon, and A. Stasiak, eds., *Physical and Numerical Models in Knot Theory*, vol. 36 of *K&E Series on Knots and Everything* (World Scientific, 2005), ISBN 981-256-187-0.

[6] J. Hou, G. Sims, C. Zhang, and S. Kim, Proc. Nat. Acad. Sci. USA **100**, 2386 (2003).

[7] J. Hou, S. Jun, C. Zhang, and S. Kim, Proc. Nat. Acad. Sci. USA **102**, 3651 (2005).

[8] A. Lesk, ed., *Introduction to Protein Architecture* (Oxford, 2001), ISBN 0-19-850474-8.

[9] L. Holm and C. Sander, J. Mol. Biol **233**, 123 (1993).