Implementing knot-theoretical characterization methods to analyze the backbone structure of proteins: Application to CTF L7/L12 and carboxypeptidase A inhibitor proteins

Gustavo A. Arteca, Orlando Tapia† and Paul G. Mezey

Departments of Chemistry and Mathematics, University of Saskatchewan, Saskatoon, Canada
†Department of Physical Chemistry, University of Uppsala, Uppsala, Sweden

In this work we apply a recently developed method for characterizing the shape of the tertiary structure of proteins. The approach is based on a combination of graph- and knot-theoretical characterizations of Cartesian projections of the space curve describing the protein backbone. The proposed technique reduces the essential shape features to a topologically based code formed by a sequence of knot symbols and polynomials. These polynomials are topological invariants that describe the overcrossing and knotting patterns of curves derived from the molecular space curve. These descriptors are algorithmically computed. The procedure is applied to describe the structure of the carboxy terminal fragment of the L7/L12 chloroplast ribosomal protein (CTF L7/L12) and the potato carboxypeptidase A inhibitor protein (PCI), which has a set of three disulfide bridges. In the former case, we describe the protein's shape features in terms of its α-helices, and a backbone simplified by considering helices without internal structure. An extension of the methodology to describe disulfide bridges is discussed and applied to PCI. Changes in the knot-theoretical characterization due to possible uncertainties in the resolution of the X-ray structure, as well as the inclusion of low-frequency motions of the backbone, are also discussed.

Keywords: protein structure, folding analysis, knot-theory, Jones polynomials, CTF L7/L12 ribosomal protein, carboxypeptidase inhibitor protein

INTRODUCTION

Many of the essential features of a protein fold emerge from a molecular graphics display of its α-carbon backbone. At a qualitative level, visual inspection at the screen allows for recognition of systematic structural patterns.[1–10] Shape descriptors have also been developed to convey in a more concise way the various structural features.[11–22] Many shape descriptors, such as the winding, writhing, twisting and linking numbers, are global in nature.[11,12] These appproaches do not provide a characterization that is sensitive to changes in the local features of the molecular backbone, which are usually rather important. On the other hand, global shape descriptors (some based on graph theory) have been developed to represent protein structural motifs,[14–17] and to search for them within large databases.[19–21] The study of changes in the folded state of biomacromolecules requires the elaboration of mathematically well-defined local descriptors that might help in sensing protein changes due to dynamical fluctuations and, more specifically, folding–unfolding processes.

In this work we apply and further develop a method for characterizing the shape of the tertiary structure of proteins. The scheme is based on a graph- and knot-theoretical approach to examining Cartesian projections of the space curve associated with the protein backbone.[23] Following this approach, the shape features can be reduced to a discrete number of knot symbols and polynomials related to the overcrossing and knotting patterns of curves derived from the molecular space curve. These descriptors can be computed algorithmically, without visual inspection, and displayed with the molecular model, thereby complementing the information provided by the latter mode of inspection.

In the next section we give a brief explanation of the derivation of knots from the molecular space curves representing the protein backbone. The knots are discerned with topological invariants; we have used the Jones polynomials as in Reference 23. The recognition of some basic structural patterns in terms of the knots is discussed. An algorithm for the characterization of protein structures is presented that involves, on the one hand, a description consisting of the α-helices and, on the other hand, a description consisting

Address reprint requests to Dr. Tapia at the Department of Physical Chemistry, University of Uppsala, Box 532, S 751 21 Uppsala, Sweden.
Received 3 October 1990; accepted 6 November 1990

of the backbone, with the helices replaced by straight-line segments. The procedure is applied to the analysis of the structure of the carboxy terminal fragment (CTF) of the L7/L12 ribosomal protein from chloroplast. The latter is of interest because it is involved in polypeptide synthesis in bacteria.[24–26] This example is discussed thoroughly as an illustration of the procedure. The method provides a distinctive description of shape features of the three α-helices in the CTF L7/L12 protein, as well as its backbone structure, simplified by considering only the helical axes. A similar analysis is then performed for the potato carboxypeptidase A inhibitor (PCI).[27] This protein contains three disulfide bridges. We discuss the extension of the methodology to consider explicitly disulfide bridges or metal bonds. This generalization leads to the occurrence of linked, rather than knotted, structures. Finally, we comment on the application of this method to the study of atomic motions in proteins. The potential usefulness of the technique for recognizing invariant shape patterns in dynamical studies is discussed.

KNOT-THEORETICAL CHARACTERIZATION OF MOLECULAR SPACE CURVES

The protein backbone is approximately described by the sequence of α-carbon atoms of the amino acid residues. In this representation some groups are not included. Accordingly, our analysis will be relevant to only some large-scale structural features of the molecule.

The parametric space curve $\mathbf{r}(t)$ of α-carbon atoms, a molecular space curve, is given as[28]

$$\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k},\ 0 \leq t \leq 1 \qquad (1)$$

where $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ indicate the three unit vectors of an orthogonal Cartesian framework taken as a reference. The line described by Equation (1) is a sequence of straight-line segments. The scalar functions in Equation (1) need not be differentiable; as for now, we shall require them to be single valued, bounded and continuous. In most of the cases for biomacromolecules, $\mathbf{r}(t)$ will be an open curve (i.e., it will not be a loop). The parametric form (1) represents well the orientation of the backbone, where $\mathbf{r}(0)$ and $\mathbf{r}(1)$ are the N-terminal and C-terminal ends, respectively.

In Reference 23 a number of techniques were given to characterize the essential shape features of Equation (1). In all cases, the procedures were related to the occurrence of crossings in the planar projections of Equation (1) along some preferential viewing direction. These "crossings" result when one section of the space curve passes over another when viewed from some direction in space. We shall refer to these as "overcrossings," thus reserving the word "crossing" for those found when the curve is projected to a plane (an actual crossing). In a "degenerate projection" two or more crossings may be projected on one another. In a "regular projection" all projected crossings are separated.

If the two end points $\mathbf{r}(0)$ and $\mathbf{r}(1)$ of the mathematical curve $\mathbf{r}(t)$ are joined, we obtain an object that is topologically a loop and possibly a knot.[29–31] The usefulness of modern knot theory in chemical applications was recognized by Walba,[32] and since then it has been studied at length in the literature.[23,33–41] In this work, we follow the notations of Reference 38 and the approach of Reference 23, where a rule to derive knots from the molecular space curves was presented.

In what follows, we shall analyze examples of protein backbones from X-ray coordinates given in the Protein Data Bank (PDB). The preferential viewing directions will be taken as the three Cartesian projections to the xy, yz and zx planes defined by the PDB coordinate frame.

A knot-theoretical description allows one to describe some topological features that remain invariant for various placements and deformations of the backbone (as long as it does not break or rejoin somewhere). As the curve is oriented, it can be characterized by the handedness[29] of its overcrossings. Based on the arrangement of overcrossings, it is possible to assign polynomials to each knot. These polynomials are topological invariants, i.e., they do not change as a function of the arrangement or deformation of the string. A number of polynomials are available to characterize the knots.[29–34] The Jones polynomials are used here;[23] they are easy to compute and discriminating enough for our needs.[23,38] In Table 1 we list the results for the Jones polynomials we shall find in the following sections of this work.

To derive a knot we proceed as follows:[23]

(1) Consider a projection of the molecular space curve to one of the preferential viewing planes. We shall assume that all crossings of the resulting planar curve are nondegenerate. This can always be achieved by a small distortion that makes the placement of $\mathbf{r}(t)$ a regular one.
(2) Attach to $\mathbf{r}(0)$ and $\mathbf{r}(1)$ straight-line segments perpendicular to the viewing plane and pointing away from the viewer. Make these segments long enough so that they will reach beyond the most distant point of the original space curve.
(3) Join the far ends of the line segments by another straight-line segment parallel to the viewing plane.

These operations produce a loop from the original space curve $\mathbf{r}(t)$ that is a simple loop or a knot K_0 characterized by a polynomial $V(K_0)$. For proteins, this closing corresponds to formally joining the C- and N- termini. In the detailed, local analysis that follows we will disregard all crossings produced by the closing of the loop, as they introduce information not present in the original curve.

In practice, performing operations 1–3 on the protein backbone usually produces an unknotted loop. The simple loop (called an "*unknot*") is the trivial knot; its Jones polynomial is given by $V(K_0) = 1$. However, one can derive nontrivial results from the above, by constructing a family

Table 1. Jones polynomials for the knots and links discussed in this work

Knot symbol	Jones polynomial
0_1	1
3_1	$-t^4+t^3+t$
3_1^*	$-t^{-4}+t^{-3}+t^{-1}$
4_1	$t^2-t+1-t^{-1}+t^{-2}$
$4_1^{2}{}^*$	$t^{9/2}+t^{5/2}-t^{3/2}+t^{1/2}$
5_2^*	$-t^{-6}+t^{-5}-t^{-4}+2t^{-3}-t^{-2}+t^{-1}$

of loops from the original one, after performing switches in the original overcrossings.

Consider the projection of the loop K_0 to a plane. Suppose that the planar curve has n crossings, $n \geq 1$, resulting from n regular overcrossings in the original loop. Because the curve of K_0 is oriented, all n crossings of the projection can be characterized by the numbers $C_j = \pm 1$ (crossing types) representing right- and lefthanded crossings, respectively. This information is collected into a vector

$$\mathbf{C} = (C_1, C_2, \ldots, C_n) \tag{2}$$

One can associate a family of possible knots with the same two-dimensional (2D) projection by suitably modifying some or all of the n C_j numbers. This generates a set of polynomials that are determined from the original loop K_0 and that provide a characterization. Consider the n-dimensional vector (the so-called switching vector)

$$\mathbf{v} = (v_1, v_2, \ldots, v_n) \tag{3}$$

with elements

$$v_n = \pm 1 \tag{4}$$

One can generate a new vector of crossing types $\mathbf{C}^v$ from the reference vector $\mathbf{C}$, as follows

$$\mathbf{C}^v = (v_1C_1, v_2C_2, \ldots, v_nC_n) \tag{5}$$

By taking all of the 2^n possible n-dimensional vectors $\mathbf{v}$ of form (3), the crossing vectors $\mathbf{C}^v$ of all possible knots (and links) compatible with the given 2D projection (with crossing information suppressed) will be generated. The family of knots obtained is denoted by $\{K_b\}$, the corresponding family of Jones polynomials is $\{V_{K_b}(t)\}$, with t the polynomial variable. Note that topologically equivalent knots may be obtained by two or more different choices of $\mathbf{v}$ and $\mathbf{C}^v$ vectors; some choices of $\mathbf{v}$ vectors may be inconsistent with the 2D projection in the sense that they may not lead to knot. We will consider a subset of single switches. This set is defined by the following vectors:

$$\mathbf{v}_0 = (1,1,1, \ldots, 1)$$

$$\mathbf{v}_1 = (-1,1,1, \ldots, 1)$$

$$\mathbf{v}_2 = (1,-1,1, \ldots, 1)$$

$$\vdots$$

$$\mathbf{v}_n = (1,1,1, \ldots, -1)$$

The Jones polynomials of the knots $\{K_n\}$, obtained by performing the switches specified by vectors $\mathbf{v}_i$,[1-4,7] are in most instances different from the actual knot K_0. Hence, they provide a more detailed characterization of the projection. In the following sections we shall use the complete set of knots $\{K_n\}$ as a shape descriptor, following a formal vector notation (knot vector $\mathbf{K}$):

$$\mathbf{K} = (K_0, K_1, K_2, \ldots, K_n) \tag{6}$$

One may take the family of the corresponding Jones polynomials $\{V_{K_n}(t)\}$ for characterization. Alternatively, one may select just one polynomial, according to some criteria, for a concise nonvisual descriptor.[23]

CROSSING PATTERNS AND STRUCTURAL HIERARCHIES

The tertiary structure of proteins is expressed by the relative location of the basic elements defining the secondary structure: α-helices, β-strands, β-sheets and connecting loops.[1,2,9] In this section we shall discuss how the occurrence of some of these features is characterized in terms of the knots $\{K_n\}$ discussed above.

Let us first analyze the occurrence of an α-helix. Figure 1 shows three views of the same lefthanded α-helix. In view 1, which is perpendicular to the helical axis, there are no real overcrossings. Accordingly, if we transform the helix into a loop by joining its terminal points, the resulting characteristic knot vector $\mathbf{K}$ from Equation 6 is

$$\mathbf{K}(\textit{view 1}) = (0_1) \tag{7a}$$

where 0_1 identifies the unknot (0 crossings, Table 1). If the angle between the helical axis and the viewing direction is slightly changed, then crossings within the helix appear (view 2 in Figure 1). Although a number of crossings show up, switching any one of them leaves the loop an unknot. The shape description is then given as

$$\mathbf{K}(\textit{view 2}) = (0_1, 0_1, 0_1, \ldots) \tag{7b}$$

where each new unknot corresponds to altering a full turn of the original helix.

The first nontrivial description appears if the helix is turned further toward the viewing direction (view 3 in Figure 1). Note that in this case two turns generate three crossings. The knots derived from this projection are illustrated in Figure 2, which displays two consecutive turns in the helix. This case is characterized as

$$\mathbf{K}(\textit{view 3}) = (0_1, 0_1, 4_1, 3_1, 0_1, 4_1, 3_1, \ldots) \tag{7c}$$

where 4_1 (four crossings) and 3_1 (three crossings) represent the figure-eight and the lefthanded trefoil knots, respectively (Table 1). Note the recurrent triplet $(0_1, 4_1, 3_1)$, corresponding to crossings between consecutive turns. A righthanded helix would be distinguished from a lefthanded one at the arrangement of view 3. If the helix is righthanded, the

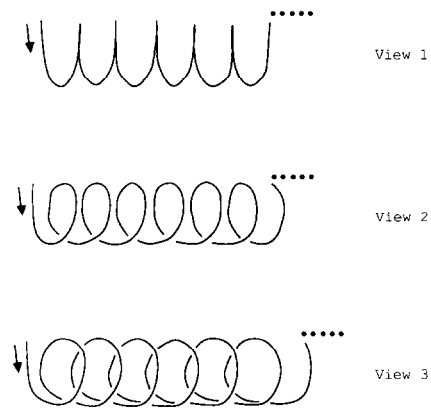

Figure 1. Crossing patterns for a lefthanded α-helix. The three different views represent the changes in crossing pattern due to a change in orientation. The characterization of the views in terms of knots is discussed in the text; dots indicate the continuation of the helix

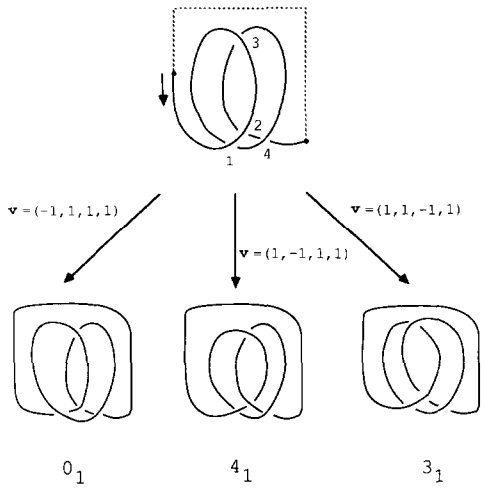

Figure 2. Knots found by performing a switch of a single overcrossing in two turns of a lefthanded α-helix. The first three switching vectors are indicated in the figure. The three knots obtained are the unknot, the figure-eight knot and the lefthanded trefoil knot. See Table 1 for the corresponding Jones polynomials

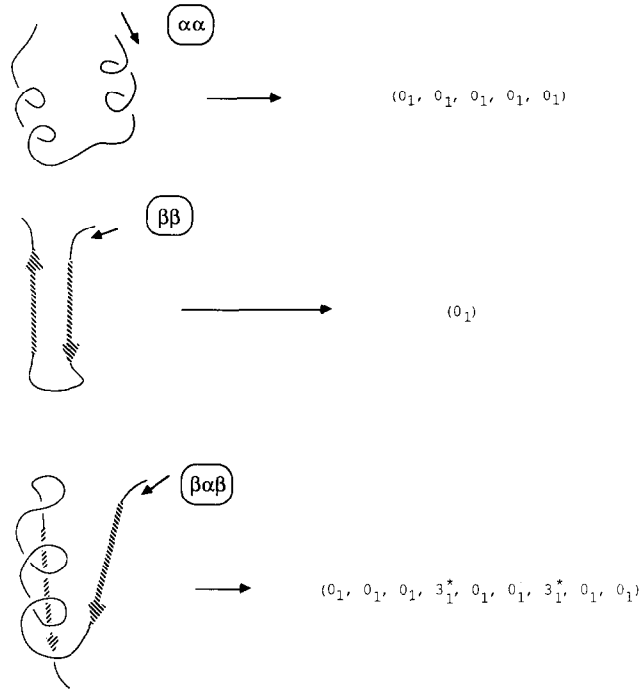

Figure 3. Characterization of the αα, ββ and βαβ structures by knot vectors. The results correspond to one view, from an arbitrarily chosen direction

righthanded trefoil knot 3_1* will appear in Equation (7c) instead of the lefthanded trefoil knot 3_1. The figure-eight knot, being achiral, does not change with the handedness of the helix.

This analysis shows that the relative orientation of a helix with respect to the viewing direction is distinctly characterized by the sequence of derived knots. Similar conclusions can be found for other characteristic structures.

Figure 3 compares three arrays of consecutive secondary structures: two α-helices, two β-strands and a βαβ sequence.[9] We have used the simplest views. The two α-helices are easily recognizable here, as they lead to a series of unknots:

$$\mathbf{K}(\alpha\alpha) = (0_1,0_1,0_1,0_1,0_1) \quad (8a)$$

The two β-strands lead to a single unknot, as in this view they do not cross:

$$\mathbf{K}(\beta\beta) = (0_1) \quad (8b)$$

The third drawing in Figure 3 shows a βαβ sequence, where the second β-strand overcrosses completely the α-helix with two turns. The result of this case is distinct from that of the other two, and it involves a trefoil knot (righthanded because the helix is righthanded):

$$\mathbf{K}(\beta\alpha\beta) = (0_1,0_1,0_1,3_1^*,0_1,0_1,3_1^*,0_1,0_1) \quad (8c)$$

The periodicity of the structure is present in the sequence $(0_1,3_1^*,0_1)$, which elicits the crossing of the β-strand over a turn of the α-helix.

These examples suggest that the more complex the crossing pattern, the more complicated the knots involved. This behavior is exemplified in the analysis of the packing of β-strands to form a β-sheet. Figure 4 shows the sequence of

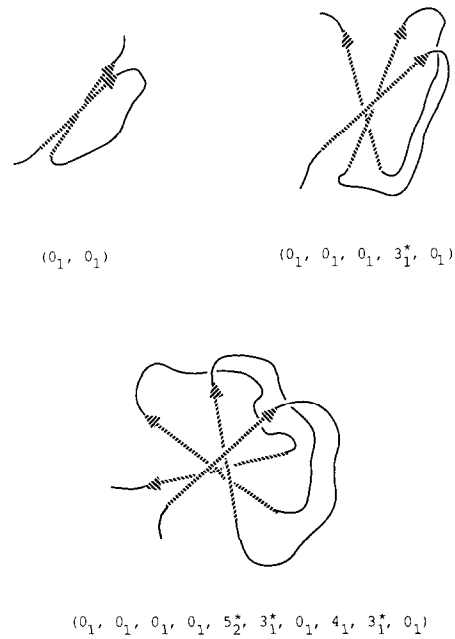

Figure 4. Characterization by means of knot vectors for the packing of an increasing number of β-sheets. The results correspond to one view, from an arbitrarily chosen direction

two, three and four β-strands. The number of crossings is one, four and nine, respectively. Note that the overcrossing of β-strands may be degenerate (i.e., several sections of the curve can cross over the same point). To avoid this problem we have tilted the views slightly in the cases of the packed

three and four strands, to make the placement of the string regular. The results are

$$\mathbf{K}(\beta\beta) = (0_1,0_1) \quad (9a)$$

$$\mathbf{K}(\beta\beta\beta) = (0_1,0_1,0_1,3_1^*,0_1) \quad (9b)$$

$$\mathbf{K}(\beta\beta\beta\beta) = (0_1,0_1,0_1,0_1,5_2^*,3_1^*,0_1,4_1,3_1^*,0_1) \quad (9c)$$

Note that a crossing of four strands leads to a knot with five crossings. More complicated knots are found for more extended sheets.

For some applications it is important to characterize the relative location in three-dimensional space of the α-helices and β-sheets. To describe these features exclusively it is better to omit the crossings due to the turns within the helices. This can be achieved, as discussed in Reference 23, by replacing the α-helices by their main axes. (See also References 26 and 42 regarding the computation of the helical axes.) This approach will be discussed in subsequent sections using some illustrative examples. A polyhedral representation of novel chiral features of α-helical domains[22] has also been formulated in terms of knots, where the details of individual helices are disregarded.

The occurrence of other elements, such as disulfide bridges or metal bonds, in the molecular backbone needs to be treated in a separate fashion. Because a bridge links two branches of the backbone, the curve is no longer a simple string; it exhibits bifurcations. The extension of the knot-theoretical characterization to systems with these properties is nevertheless possible. This subject is discussed in a later section.

CHARACTERIZATION OF CTF L7/L12

The example we discuss here is CTF L7/L12. This is a small protein with 69 amino acid residues, exhibiting three right-handed α-helices and three β-strands.[43,44] The CTF L7/L12 protein has recently been studied in the context of molecular dynamics. Low-amplitude motions have revealed that the backbone of this protein can move its subdomains as quasi-rigid solids characteristic of the folding pattern.[26,44]

We have performed the analysis of the protein on two levels. First we have characterized and compared the three helices. Second we have determined the relative location of helices and sheets by replacing the helices with their main axes.

The characterization can be done automatically by a computer program or can be deduced from the graphical display of a molecular space curve. In our case, the graphical information was obtained by using the mdFRODO graphical package developed at Uppsala.[45]

The three helices will be called *A*, *B* and *C*, according to the order of their occurrence along the oriented backbone. As an illustration, Figure 5 shows schematically the three Cartesian views of the helix *C*, as obtained from the PDB coordinates. The helix is defined by the sequence of residues from 53 to 63. Figure 5 shows an extended segment spanning residues 49–63. It is clear that only one view (the *xz* projection) will lead to nontrivial knots.

Figure 6 summarizes the results of the characterization of the *xz* projection of helix *C*. The upper drawing shows the original helix transformed into a loop by formally joining

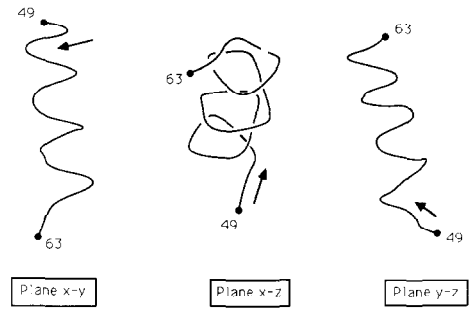

Figure 5. Different Cartesian projections for the backbone of the α-helix of CTF L7/L12. The axes x, y and z are taken from the Protein Data Bank files. Numbers on the backbone indicate the residues

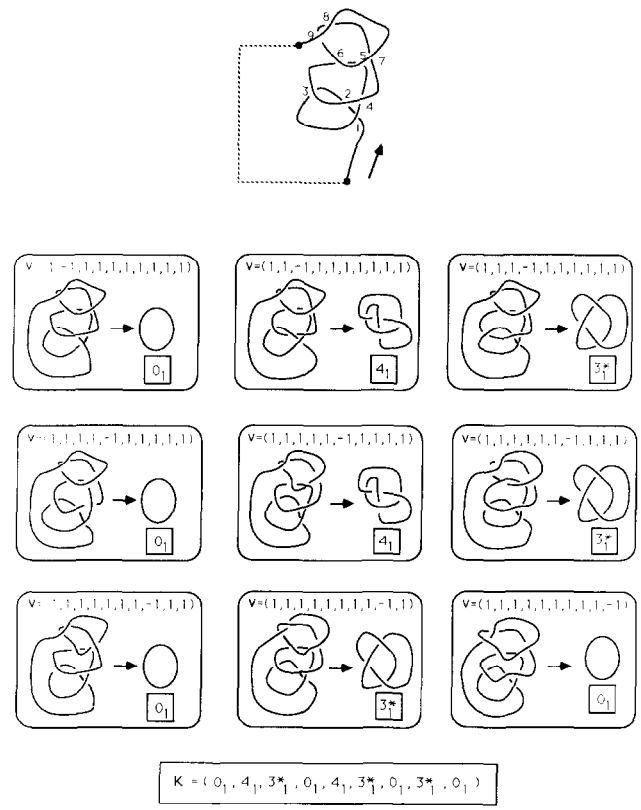

Figure 6. Derivation of knots from the loop associated with the projection of the α-helix of the CTF L7/L12 to the x-z plane. The diagrams indicate the knots obtained after performing a single switch for every one of the nine overcrossings present in the projection. The notation follows that of Table 1. The knot vector **K** *at the bottom of the figure summarizes the findings*

the 49 and 63 termini. The curve exhibits nine overcrossings. The nine graphs below show the results of the nine single-switches performed over the curve. In each graph we have indicated the curve and its corresponding knot obtained after the switch. At the bottom of the figure, the characterization of this view of the helix is summarized by its corresponding knot vector **K**.

Table 2. Characterization of CTF in terms of its series of asssociated knots. (See Tables 1 and 3.) Disulfide bridges have been omitted. The descriptor K′ corresponds to simplifying the overcrossing pattern by considering the uncertainty in the resolution

Projection	K	K′
x-y	$(0_1,0_1,0_1,0_1,0_1,0_1)$	$(0_1,0_1,0_1,0_1)$
y-z	$(0_1,0_1,0_1,3_1{*},0_1,0_1)$	$(0_1,0_1,0_1,0_1)$
x-z	$(0_1,0_1,0_1,0_1,0_1,0_1,0_1,0_1,0_1)$	$(0_1,0_1,0_1,0_1,0_1,0_1)$

This procedure can be applied to each helix, along each of the three preferential viewing directions. Table 2 summarizes the results by providing the knot vectors (first three rows). From this table we can deduce several structural features of the packing of the three helices:

(1) All of the helices have a trivial knot characterization in the *xy* projection. According to the results in the previous section, this implies that the helical axes are approximately perpendicular to the *z*-axis of the PDB reference framework.

(2) Helices *A* and *C* have a trivial knot descriptor in the *yz* projection, whereas helix *B* has a trivial knot descriptor in the *xz* projection. This implies that helices *A* and *C* are approximately parallel with the *y*-axis, whereas the axis of helix *B* is perpendicular to the axes of *A* and *C*.

(3) A comparison of the results for the actual shape descriptors allows one to make quantitative assessments of the parallel character of helical axes. It is also possible to decide the degree to which the actual helix deviates from an ideal helix lying on a cylindrical surface.

   (a) The table reveals only four crossings for helix B. Moreover, we find the characteristic sequence $(0_1,4_1,3_1{*})$ of two consecutive turns of a right-handed α-helix. This result indicates that helix *B* is either a rather short one with loosely packed turns, or else its axis is not closely aligned with the *x*-axis. (Note that the closer the alignment, the more crossings between turns of the helix.)

   (b) The characterization of helix *C* shows nine crossings and two $(0_14_1,3_1{*})$ sequences. This fact indicates that the helix is longer, or else that its alignment with the *y*-axis is better than the alignment of helix *B* with the *x*-axis.

   (c) Helix *A* shows thirteen crossings, but only one $(0_1,4_1,3_1{*})$ sequence. Accordingly, it is likely that the axis of helix *A* is better aligned with the *y*-axis.

In Color Plate 1 we present the *xz* view of the full structure of the CTF L7/L12 protein. The figure reveals all of the features that we have characterized by knot vectors in Table 3. Helix *A* is seen along a direction nearly coincident with its axis. (Note the hole.) Helix *B* is perpendicular to the viewing direction and the axis of helix *C* forms an angle of approximately 45° with the viewing direction. Note that the helix *B* is the shortest, with only 2 turns.

The comparison of the results shown in the graphical display with those characterized by the abstract shape descriptor indicates that most of the important features can be recovered by our analysis.

The characterization of the essential shape features of the CTF L7/L12 protein can be completed by computing the knot vector for the full backbone, where now the α-helices have been replaced by straight-line segments. Color Plates 2–4 show the three preferential views of this protein backbone, under the above simplification. The characterization in terms of knots is given as the last row in Table 3. Note that the projections on the *xy* and *xz* planes produce a number of crossings between β-strands and loops (β-turns) with the axes of the α-helices; however, all of these crossings lead to unknots. On the other hand, the *yz* projection presents some switches leading to righthanded trefoil knots. If we compare this with the result (9b), we can argue that only the projection along the *x*-axis (on the *yz* plane) is suitable for the recognition of the "sheet structure," formed by the two β-strands and the axes of some of the helices. A comparison of the views in Color Plates 2–4 clearly shows that this is indeed the case.

The analysis performed here illustrates how this methodology[23] can be applied to obtain a mathematical characterization of the tertiary structure of a protein. The procedure can thus be applied in a hierarchical manner, describing first the shape features of the isolated helices, and then the overall backbone structure with the helices replaced by straight-line segments.

Of course, the features studied here correspond to the rigid structure. Some of these may disappear if the uncertainty in the resolution, or small motions, are included.

Table 3. Characterization of CTF in terms of its series of associated knots. Knots correspond to single switches of overcrossings; the first knot stands for the original backbone structure after joining the endpoints

Plane: *x-y*	*x-z*	*y-z*
Helix *A* (0_1)	$(0_1,0_1,0_1,3{*}_1,3{*}_1,0_1,3{*}_1,0_1,$ $4_1,3{*}_1,0_1,3{*}_1,0_1,3{*}_1)$	(0_1)
Helix *B* (0_1)	(0_1)	$(0_1,0_1,4_1,3{*}_1,0_1)$
Helix *C* (0_1)	$(0_1,0_1,4_1,3{*}_1,0_1,4_1,3{*}_1,0_1,$ $3{*}_1,0_1)$	(0_1)
Skeleton $(0_1,0_1,0_1,0_1,0_1)$	$(0_1,0_1,0_1,0_1)$	$(0_1,0_1,0_1,0_1,0_1,3{*}_1,3{*}_1)$

CHARACTERIZATION OF PCI

As a second example we have chosen the inhibitor protein of the carboxypeptidase from potatoes.[27] This is a small protein having only 38 amino acid residues. Moreover, it is irregular in that it exhibits neither recognizable long helices nor sheets. Its structure is stabilized by the presence of three disulfide bridges, a feature that has not been considered yet in our knot-based analysis.[23]

The knot-theoretical characterization of this small irregular protein may lead to patterns comparable to the ones discussed above. However, the occurrence of disulfide bridges makes it necessary to modify the present approach. (The Jones polynomials are not defined if there are bifurcations along the string.) We concentrate first only on the backbone structure. Thereafter, we discuss the modifications that are necessary to take the bridges properly into account.

Figure 7 displays the three preferential views of the backbone structure of PCI, as obtained from the PDB coordinates (without disulfide bridges). The graphs show a rather featureless folding pattern.

Table 2 provides the characterization of these three views in terms of the knot vector **K**. Note that the projections to the xy and xz planes do not lead to a knot, though there are many crossings. This sequence of unknots could be taken in principle as two views of an α-helix. However, it cannot be an α-helix because the characterization of the projection to the yz plane includes a righthanded trefoil knot. As discussed above, this feature is not present for a single α-helix. On the other hand, the characterization of the yz view is quite similar to that of a $\beta\alpha\beta$ sequence (Equation (8c)) or a $\beta\beta\beta$ crossing (Equation (9b)). As a matter of fact, the characterization of a $\beta\alpha\beta$ sequence with a formal "single-turn" α-helix is identical to that of the $\beta\beta\beta$ sequence. Considering the larger number of crossings (leading to unknots) for the other orthogonal views, it would seem more reasonable to classify the backbone of PCI as akin to a β-single turn-$\alpha\beta$ structure. Accordingly, though the protein is irregular, one could argue the incipient formation of a recognizable motif in it.

Note that some of the overcrossings in the given views of this protein occur by a very small margin (Figure 7). That is, some sections of the backbone overcross each other at a given view but would not do so if the viewing direction were slightly tilted. This fact should be taken properly into account in any realistic shape characterization. The backbone structure has an inherent uncertainty, resulting from the experimental uncertainty in the nuclear positions as given by the X-ray structure. Even a good resolution in the geometry (say, 1.5 Å) leaves a significant margin to override an overcrossing. This uncertainty can be included in our analysis by contrasting the shape description based on the original X-ray structure with the description in which marginal overcrossings are neglected. As an illustration, we have estimated the latter for PCI. The results appear as the knot vector **K′** in the last column of Table 2. Note that the description becomes simpler, because the number of crossings is reduced and the trefoil knot disappears. This reveals that the regularities observed before are the results of small-scale geometric features.

The joint description in terms of the full knot vector **K** and the vector **K′**, for the structure without marginal overcrossings, is a more complete one. This approximation should be used in all practical cases, to quantitatively estimate the relevance of observed overcrossings.

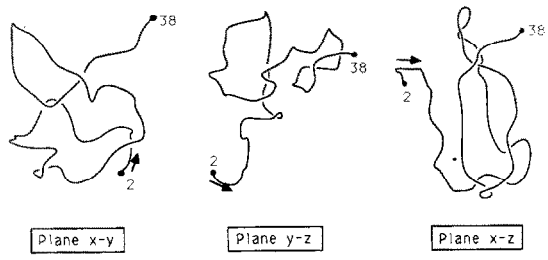

Figure 7. Projections of the backbone of PCI, without the disulfide bridges

EXTENSION TO DISULFIDE BRIDGES OR METAL BONDS

To apply knot theory to this problem, it is necessary to replace the bridges between two segments of the space curve $\mathbf{r}(t)$ by other objects. One possibility is to replace each of the bridgeheads of a disulfide bridge (or metal bond) by two lefthanded crossings. The procedure is illustrated in Figure 8. The upper part of the figure shows the original bridge-

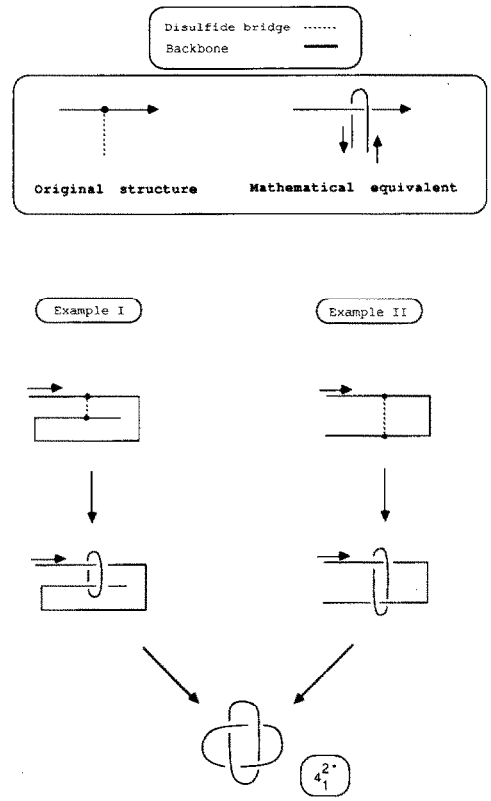

Figure 8. Derivation of a link from a disulfide bridge between main chain atoms. The upper part of the figure indicates the replacement of the bridge by two looped lefthanded crossings. Two examples of disulfide bridge are provided below. Example 1 shows a bridge between parallel chains and example 2 the analogous case for antiparallel chains

head, and its replacement by a turn around the main direction of the curve (backbone). Now, the two turns for the two bridgeheads of a disulfide bond can be joined to form a loop.

The bottom part of Figure 8 illustrates the two possible ways of joining the bridgeheads. On the lefthand side, one finds a bridge between two parallel sections of the curve (according to the curve's orientation). The result is the replacement of the bridge by a loop enclosing the two branches of the curve. On the righthand side, the analogous situation is shown for a bridge between two antiparallel sections of the curve. In this case, the bridge is replaced by a loop twisted around the two segments. Note, however, that when the termini of the main chain are closed to form the loop then, depending on signs of other crossings, the descriptions of the parallel and antiparallel bridges may lead to the same link. An example is shown in Figure 8, where both cases are characterized by the link 4_1^{2*} (Table 1).

A description of systems with these types of bonds becomes a bit more involved than that of ordinary proteins, because the number of knots and links that appear is larger than those found in the previous examples. Nevertheless, we have shown that their occurrence can be handled within the context of a knot-theoretical characterization of space curves.

CONCLUSION

In this work we have developed further a methodology for characterizing the shape features of protein backbones in a discrete fashion. We have shown the knot-theoretical characterization of a number of essential structures occurring in proteins, and explored its application to the case of some small proteins. Our goal is to fully implement this procedure with a graphical display of the backbone as a complementary tool for unbiased assessment of molecular shape. The comparison of homologous structures of different proteins, as well as the study of shape invariances along conformational paths (folding), are among the potentially interesting applications of this analytical tool.

Concerning the practical use of this procedure to analyze large proteins, it is clear that the number of overcrossings (hence, the number of knot types) may grow for large molecules, particularly along certain viewing directions. Thus, it becomes necessary to assess the relevance of the various crossings. One possibility is to neglect the marginal overcrossings, i.e., those which do not survive a small tilt in the observational direction. This approach takes into account the uncertainty in the input nuclear geometries.

Still, there could exist crossings that, surviving the above selection, might be omitted. This would be the case, for instance, of overcrossings between remote segments of the curve. To handle this situation, one could introduce a cutoff measure based on distances between residues.[23] Only the overcrossings between segments with distances below the given threshold would be retained for the analysis. The cutoff value could take into account the hydrophobic or hydrophilic nature of the residues involved in the crossing.

Finally, there is the problem of the choosing preferential viewing directions. This problem may become critical in following a folding pattern of a protein. As discussed in Reference 23, this difficulty can be circumvented by using a projection onto a sphere, which takes into acount all viewing directions. Applications along all these lines of research are under development and will be communicated elsewhere.

ACKNOWLEDGMENTS

Arteca would like to express gratitude for financial help from the University of Uppsala that made possible his stay at the Department of Physical Chemistry, where part of this work was completed. Tapia acknowledges financial support from NFR. Financial support of the Topology Program (PGM) from NSERC, in the form of both strategic and operating grants, is gratefully acknowledged.

REFERENCES

1 Richardson, J.S. *Adv. Protein Chem.* 1981, **34**, 167
2 Richardson, J.S. *Methods Enzymol.* 1985, **115**, 359
3 Carson, M. and Bugg, C.E. *J. Mol. Graphics* 1986, **4**, 121
4 Carson, M. *J. Mol. Graphics* 1987, **5**, 103
5 Lesk, A.M. and Hardman, K.D. *Science* 1982, **216**, 539
6 Lesk, A.M. and Hardman, K.D. *Methods Enzymol.* 1985, **115**, 381
7 Dearden, T. *J. Comp. Chem.* 1989, **10**, 529
8 Richardson, J.S. *Methods Enzymol.* 1985, **115**, 341
9 Jaenicke, R. *Prog. Biophys. Mol. Biol.* 1987, **49**, 117
10 Kikuchi, T., Némethy, G. and Scheraga, H.A. *J. Comp. Chem.* 1986, **7**, 67
11 Delbrück, M. *Proc. Symp. Appl. Math.* 1962, **14**, 55
12 Fuller, F.B. *Proc. Symp. Appl. Math.* 1962, **14**, 64
13 Fuller, F.B. *Proc. Natl. Acad. Sci. USA* 1971, **68**, 815
14 Le Bret, M. *Biopolymers* 1979, **18**, 1709
15 De Santis, P., Morosetti, S. and Palleschi, A. *Biopolymers* 1983, **22**, 37
16 Hao, M.-H. and Olson, W.K. *Biopolymers* 1989, **28**, 873
17 Mitchell E.M., Artymiuk, P.J., Rice, D.W. and Willett, P. *J. Mol. Biol.* 1990, **212**, 151
18 Liebman, M.N., Venanzi, C.A. and Weinstein, H. *Biopolymers* 1985, **24**, 1721
19 Rawlings, C.J., Taylor, W.R., Nyakairu, J., Fox, J. and Sternberg, M.J.E. *J. Mol. Graphics* 1985, **3**, 151
20 Richards, F.M. and Kundot, C.E. *Protein Struct. Function Genetics* 1988, **3**, 71
21 Abagyan, R.A. and Maiorov, V.N. *J. Biomol. Struct. Dynam.* 1988, **5**, 1267
22 Maggiora, G.M., Mezey, P.G., Mao, B. and Chou, K.C. *Biopolymers*, in press
23 Arteca, G.A. and Mezey, P.G. *J. Mol. Graphics* 1990, **8**, 66
24 Möller, W. in *The Ribosomal Components Involved in EF-G- and EF-Tu-Dependent GTP Hydrolysis*, (M. Nomura, A. Tissieres, and P. Lengyel, Eds.) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1974
25 Petterson, I. and Kurland, C.G. *Proc. Natl. Acad. Sci. USA* 1980, **77**, 4007

26 Tapia, O., Nilsson, O., Campillo, M., Åqvist, J. and Horjales, E. in *Structure and Methods: DNA Protein Complexes and Proteins* (R.H. Sarma and M.H. Sarma, Eds.) Adenine Press, New York, 1990, vol. 2, p. 147
27 Weinstein, H., Liebman, M.N. and Venanzi, C.A. in *New Methods in Drug Research* (A. Makriyannis, Ed.) Prous Science, Barcelona, vol. 1, chap. 14, 1985
28 Lass, H. *Vector and Tensor Analysis*. McGraw-Hill-Kogakusha, Tokyo, 1950
29 See, for example, Crowell, R.H. and Fox, R.H. *Introduction to Knot Theory*. Springer-Verlag, Berlin, 1977
30 Dowker, C.H. and Thistlethwaite, M. *Comp. Rend. Acad. Sci.* (Canada) 1982, **2**, 129; Topology and Its Applicat. 1982, **16**, 19
31 Thistlethwaite, M. *London Math. Soc. Lecture Notes* 1985, **93**, 1
32 Walba, D.M. Stereochemical Topology. in *Chemical Applications of Topology and Graph Theory* (R.B. King, Ed.) Elsevier, Amsterdam, 1983
33 Jones, V.F.R. *Bull. Am. Math. Soc.* (NS) 1985, **12**, 103
34 Freyd, P., Yetter, D., Hoste, J., Lickorish, W.B.R., Millett, K. and Ocneanu, A. *Bull. Am. Math. Soc.* (NS) 1985, **12**, 239
35 Walba, D.M. *Tetrahedron* 1985, **41**, 3161
36 Wasserman, S.A. and Cozzarelli, N.R. *Science* 1986, **240**, 110
37 Connolly, M.L., Kuntz, I.D. and Crippen, G.M. *Biopolymers* 1980, **19**, 1167
38 Mezey, P.G. *J. Am. Chem. Soc.* 1986, **108**, 3976
39 Millett, K.C. *J. Comp. Chem.* 1987, **8**, 536
40 Simon, J. *J. Comput. Chem.* 1987, **9**, 718
41 Sumners, D.W. *J. Math. Chem.* 1987, **1**, 1
42 Åqvist, J. *Comp. Chem.* 1986, **10**, 97
43 Leijonmarck, M., Liljas, A. and Subramanian, A.R. *Biochem. Int.* 1980, **8**, 69
44 Åqvist, J., van Gunsteren, W.F., Leijonmark, M. and Tapia, O. *J. Mol. Biol.* 1985, **183**, 461
45 Nilsson, O. *J. Mol. Graphics* 1990, **8**, 192