Journal of Mathematical Chemistry 11(1992)47–63 

# AN AUGMENTED RIBBON MODEL OF PROTEIN STRUCTURE

Victor A. NICHOLSON
*Department of Mathematics and Computer Science, Kent State University, Kent, OH 44242, USA*

and

Gerald M. MAGGIORA
*Computational Chemistry, The Upjohn Company, Kalamazoo, MI 49001, USA*

## Abstract

In this paper, work on a new model, *an augmented ribbon model*, for describing a number of features of primary, secondary, and super-secondary protein structure in a qualitative but mathematically rigorous way is discussed. The structural features that can be treated by the model include connectivity, directionality, chirality, orientation, and proximity which, in many cases, are difficult to deal with using more traditional structural representations (e.g. "wireframe", "ball-and-stick", or "space-filling CPK"). In practice, the information encoded in the augmented ribbon model is represented by a labelled, directed graph (digraph) which provides an efficient means for storing and analyzing the information on computers. This opens the way to computer-based analyses of proteins and for the application of similarity methods that have been shown to be quite useful in treating small molecules.

## 1. Introduction

A protein's function is known to be closely tied to its three-dimensional structure [1–3]: denatured (i.e. unfolded) proteins are non-functional. Proteins exhibit a wide variety of beautiful and complicated structural forms [4], and the representation of their various structural features inherent in these forms is an important problem, especially as mathematical and computational methods for studying protein structure depend crucially upon representation.

Typically, protein structures are represented by "wire-frame", "ball-and-stick", or "space-filling CPK" models that provide significant structural detail, but which may obscure other structurally important features such as chirality (vide infra). Although chemical graphs have been used to represent the structure of small molecules, the use of similar graph representations in proteins has been limited, one reason being that chemical graphs cannot deal with the stereochemical features of proteins. Topologically chiral features of proteins, which are described by non-planar graphs have, however, been investigated for proteins with multiple disulfide bonds [5] and for secondary-structural elements (viz. $\alpha$-helices and $\beta$-sheets) held together by hydrogen bonds [6].

In the present work, an *augmented ribbon model* (ARM) of protein structure is developed which provides a powerful means for representing many structural features of proteins, such as connectivity, directionality, chirality, orientation, and proximity, in a clear and mathematically rigorous way. Importantly, the ARM can be represented by a labelled, directed graph, or digraph, which extends the familiar chemical graph of a molecule. Such a representation makes it possible to use a number of currently available computational methods, such as those based on molecular similarity [7], to analyze protein structures and their interrelationships in very powerful and mathematically rigorous new ways.

## 2. Basics of protein structure

Proteins are made up of "chains" of amino acids linked together by planar peptide bonds, $-CO=NH-$, as illustrated in fig. 1(a). The chains are folded into a limited number of three-dimensional patterns [8], which do not possess knots or links [9]. The sidechain $R_i$ of each amino-acid residue defines its "character" and determines whether it is likely to be located on the surface or within the core of the protein. The present work focusses on globular proteins, structurally the most well-characterized class, which possess an "oily" inner core of predominantly hydrophobic amino acids surrounded by a polar outer shell of predominantly hydrophilic amino acids. It is the latter type of amino acids that imbues this class of proteins with aqueous solubility.

The backbone of a protein chain is often represented as a two-dimensional twisted "ribbon", as shown in fig. 1(b). In the polypeptide chain depicted in fig. 1(b), the two four-atom sets $C-N-C_\alpha-C$ and $N-C_\alpha-C-N$ define the $\phi$ and $\psi$ torsional angles, respectively. Specifically, the angle between the projection C–N and $C_\alpha-C$ bonds onto a plane perpendicular to the $N-C_\alpha$ bond defines $\phi$, with a corresponding definition for $\psi$. The angle is considered to be positive for a counter-clockwise rotation of the "front" pair, C–N, when the "back" pair, $C_\alpha-C$, is held fixed. The angles are usually measured from $-180°$ to $+180°$. Since the peptide bond is planar, the $\phi-\psi$ angles determine the geometry of the chain fold.

Protein structure is considered to be hierarchical with five basic levels of structural organization, viz. primary structure, secondary structure, super-secondary structure, tertiary structure, and quaternary structure. *Primary structure* is defined by the linear sequence of amino acids, where each amino acid is symbolized by a three-letter code, e.g. $H_2N-\cdots-Phe-Gly-Lys-\cdots-COOH$. The sequence has a natural order, beginning with the residue containing the free amine (i.e. the N-terminal residue) and ending with the residue containing the free carboxyl group (i.e. C-terminal residue).

*Secondary structure* is characterized by quasi-regular structures such as helices and extended strands. As is true for primary structure, secondary structure can also be described by ordered sequences whose constituents in this case are secondary-

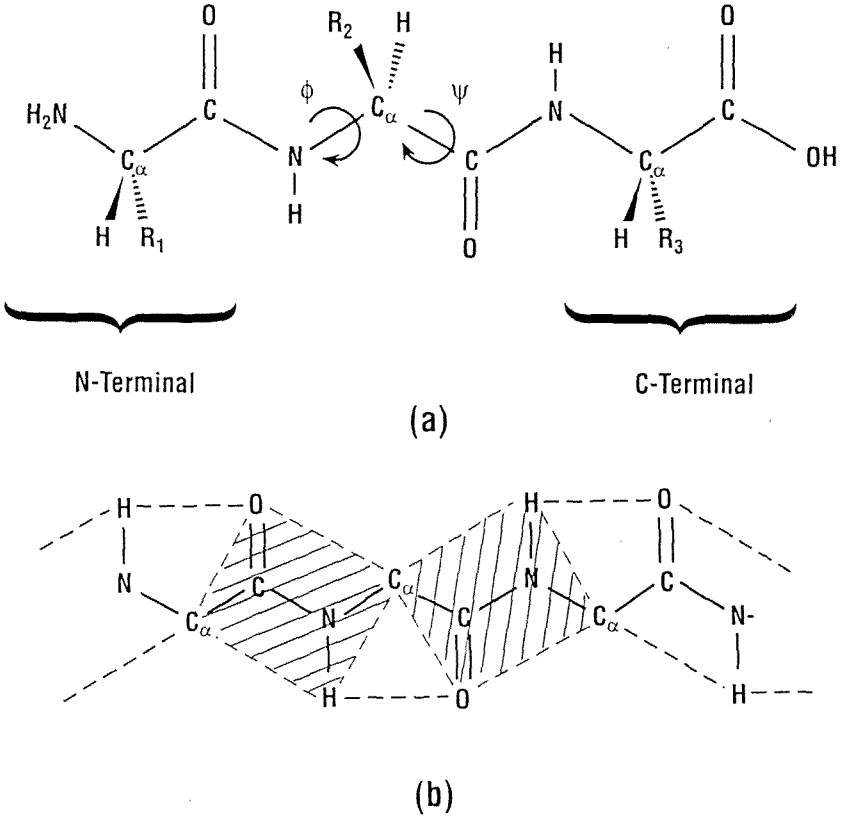


Fig. 1. (a) Extended polypeptide chain depicting the N– (amino) and C– (carboxy) terminal residues. By convention, the chain runs from its N-terminal residue $R_1$ to its C-terminal residue $R_3$. The $\phi-\psi$ rotation angles are also indicated. (b) A twisted-ribbon model of the polypeptide backbone. The edges of the "ribbon" are indicated by dashed lines, and the shaded areas represent the planes of respective peptide bonds.

structural elements, e.g. $-\alpha-\beta-\alpha-\beta-\alpha$, $-\alpha-\alpha-\alpha-\alpha-$, or $-\beta-\beta-\beta-\beta-\beta-$, where helices are designated by "α" and extended strands by "β". Only a limited number of such sequence patterns have been observed to date [8]. The secondary-structural elements are connected by either sharp "turns" in the backbone, called β-turns, which reverse the direction of the chain by approximately 180°, or by non-regular segments of the chain that are generally referred to as coils.

Secondary-structural elements pack together into higher-order structures called *super-secondary structures*, which give rise to the tertiary-fold of a protein. Specification of the complete three-dimensional structure of an individual protein chain and the structure of all its sidechains is given by *tertiary structure*. In this work, our emphasis will be on features of the chain fold, and the specifics of sidechain geometry will be neglected. Finally, the packing of individual protein subunits into higher-order

oligomeric structures, e.g. dimers, trimers, tetramers, . . . , defines a protein's *quaternary structure*. This level of structure will not be discussed further in the present work.

Although proteins are highly complex structures which may possess several thousands of atoms, details of their underlying molecular architectures can be described in ways that simplify and clarify important structural features that are obscured by more elaborate "wire-frame", "ball-and-stick", or "space-filling CPK" type models. For example, the protein backbone can be portrayed as a "ribbon", as shown in fig. 1(b), which provides a powerful means for describing many features, both obvious and subtle, of protein structure. This particular type of representation is the basis for the topological anaylsis of protein structure presented here.

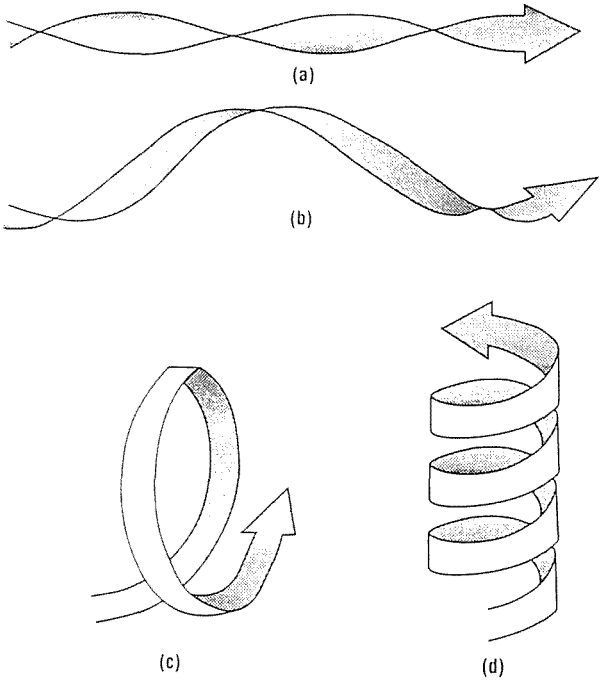


Fig. 2. Portrayal of the polypeptide backbons as a "twisted ribbon". (a) An extended chain with three right-hand half twists. (b) An "intermediate-twist" structure formed by slowly pushing the two ends of the twisted ribbon shown in (a) towards each other. (c) Ribbon representation of the right-handed crossover found in βXβ motifs (see fig. 8(a)). This "structure" can be formed by continuing the process described in (b). (d) Ribbon representation of a right-handed helix.

Chirality is a fundamental feature of proteins which is present in many forms at all levels of structure [10]. The lowest level of chirality arises from the asymmetric α-carbon atom, $C_\alpha$, found in each amino acid residue. As seen in fig. 1(a), the L-configuration is specified by the clockwise order of $R_i \rightarrow N \rightarrow C_\alpha$, as viewed from the hydrogen atom H attached to $C_\alpha$. All naturally occurring amino acids found in

proteins, except glycine, possess this configuration. Secondary-structural elements also possess a handedness; when the backbone is represented as a ribbon (see fig. 1(b)), helices and extended strands exhibit right-handed twists, the twist being considerably more pronounced in the former case, as depicted in fig. 2 (see ref. [10] for further discussion). Additional types of chirality are present in super-secondary structures, and will be touched upon in greater detail later.

A mathematical model for characterizing protein structure should be able to treat a number of the salient features, such as chirality, orientation, and proximity, which are present at the different levels of structural organization. As will be seen in detail in subsequent sections, the ARM advanced here is capable of representing a significant number of these features. For example, it can represent the connectivity and directionality of the backbone chain, it can represent the "twisted ribbon" character of the backbone without having the ribbon lose its twist or having adjacent twists in opposite directions cancel out – in other words, the chiral information residing in the twist will be preserved. Orientation is also an important feature of super-secondary structure: that is, "right" and "left", "clockwise" and "counter-clockwise", and "right-handed" and "left-handed". Proximity relationships such as encountered when two chains or secondary-structural elements pass within a given distance of one another can also be modeled. These features are important for describing, in a qualitative fashion, the way packed secondary-structural elements relate to one another. Links occur in large proteins when proximity connections are added [9]; small knots are not, however, found in proteins. The ARM proposed here allows links and knots to be formed only by using several connections and thus, is consistent with these observations.

## 3. Mathematical prerequisites

In this section, we define the mathematical features of the model. A more detailed discussion of the mathematical terms can be found in the book by Stillwell [11].

An *n-ball* is defined as the set of points in *n-space*, an $n$-dimensional Euclidean space, whose distance from the origin is less than or equal to one. The boundary of an $n$-ball is an *n-sphere* consisting of those points whose distance from the origin is equal to one, while the interior of an $n$-ball consists of those points whose distance from the origin is less than one; $n$-balls of dimension $n = 0, 1, 2$, and 3 are, respectively, the empty set, two points, a circle, and a sphere.

Two subsets of Euclidean space are *homeomorphic* if there exists a one-to-one function $f$ from one subset to the other such that both $f$ and its inverse $f^{-1}$ are continuous: $f$ is called a *homeomorphism*. The continuity of $f$ and $f^{-1}$ prevents cutting and glueing but does not, in general, preserve lengths or angles. Sets that are homeomorphic are considered to be *topologically equivalent*. The interior of an $n$-ball is topologically equivalent to $n$-dimensional Euclidean space.

An *n-cell* is any subset homeomorphic to an $n$-ball. It is convenient to describe mathematical objects as *cell complexes* using $n$-balls as building blocks. A cell complex is a union of sets $X_0, X_1, X_2, \ldots$, subject to the conditions that (1) all cells possess disjoint interiors, (2) each $X_i$ is the union of $X_{i-1}$ and of $i$-cells, and (3) the boundary of each $i$-cell in $X_i$ is contained in $X_{i-1}$. A cell complex is built up by adding cells of increasing dimensions. For example, the set $X_0$ consists of isolated vertices ("0-cells") to which edges or arcs ("1-cells") are added to form a mathematical graph $X_1$.

The cell complex shown in fig. 3 will be referred to as the *basic annulus*. The graph with vertices *a, b, c, d, e,* and *f* is referred to as the subgraph of the basic annulus. The subgraph contains a cycle *S*. The basic annulus is the union of

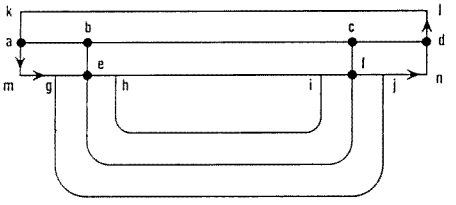


Fig. 3. A sketch of the basic annulus illustrating the main cell and the handle. The sketch shows the subgraph with vertices *a, b, c, d, e,* and *f*, the cycle $S = bcfeb$ in the subgraph, the oriented ends *km* and *nl*, and the oriented side arcs *mg* and *jn*.

two 2-cells which intersect on their boundaries along the arcs *geh* and *ifj*. The cell containing the arc *bc* is referred to as the main cell. The cell containing the arc *ef* is called the handle. The basic annulus has two oriented ends, *kam* and *ndl*. The basic annulus has two oriented side arcs *mg* and *jn*. A local $x$, $y$, and $z$ coordinate system is specified at the point *a*. The $x$-axis is along *am*, the $y$-axis is along *ab*, and the $z$-axis is determined by the right-hand rule or cross product.

Each of the three embedded complexes shown in fig. 4 will be referred to as a *twisted annulus*, which is formed from the basic annulus by twisting the cells. A twisted annulus has one of three forms *U, R,* and *L* corresponding, respectively, to an untwisted, right-hand twisted or left-hand twisted main cell. Each handle has a left-hand twist which makes the boundaries of the twisted annuli topologically distinct (note that either right- or left-handed twists may be used, the choice is arbitrary). The boundaries of the *U, R,* and *L* twisted annuli possess one component, two unlinked components, and two linked components, respectively. The twisted annuli are said to be attached end-to-end if their intersection is an end of each and the vertices *k, a,* and *m* of one coincide with the vertices *l, d,* and *n*, respectively, of the other (see fig. 3). The purpose of the handle is to fix the twist in the main cell, which prevents untwisting of the ribbon. Also, it should be noted that a pair

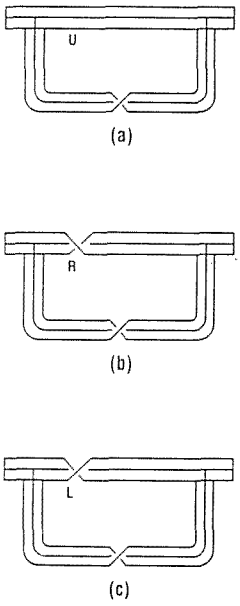


Fig. 4. The three forms of a twisted annulus in standard position are illustrated. From the top, the figures show the (a) untwisted, (b) right-handed twisted, and (c) left-handed twisted main cells. Each handle has a left-handed twist (see text for further details).

of adjacent annuli of type $R$ and $L$ attached end-to-end is not equivalent to a single annulus of type $U$ because one has two handles and the other only one. Thus, the handles prevent neighboring twists from canceling out.

Euclidean space can have one of two possible *orientations*. The "directions" are specific to the dimension of the space, and are determined by the ordering of the coordinate axes up to an even permutation. The orientations are the familiar *left* and *right* in *1*-space, *clockwise* and *counter-clockwise* in *2*-space, and *left-* and *right-handed* in *3*-space. Cells in each dimension greater than zero can be oriented. An oriented 1-cell has the appearance of a vector; an oriented 2-cell can be visualized as a small circle oriented either clockwise or counter-clockwise; an oriented 3-cell can be visualized as having a corkscrew that turns with a right- or left-handed twist.

The twisted annulus $U$ is homeomorphic to a Möbius band. If considered only as a surface, it would be non-orientable and one-sided. However, turning it over does not preserve the orientation of the side arcs or ends. Thus, the local coordinates of a twisted annulus are well-defined at the point $a$.

The drawing of the handle on a twisted annulus can be suppressed by specifying a standard position. A twisted annulus is in standard position if (1) the boundary of the main cell, when projected into the $xy$-plane, has no crossings in the case of the $U$ annulus and exactly one crossing point in the cases of the $R$ and $L$ annuli,

and (2) the ends and the subgraph lie in the $xy$-plane. The annuli in fig. 4 are in standard position. In the following, a twisted annulus in standard position will be presented by a drawing of only its main cell, with a letter indicating its type.

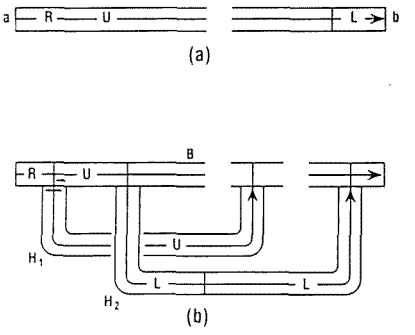


Fig. 5. A twisted ribbon is illustrated in (a). The arrow indicates the orientation of the spanning arc $ab$. The letters $U$, $R$, and $L$ refer to the forms of the twisted annuli (shown without handles) that comprise the ribbon. A twisted surface is illustrated in (b) as a union of twisted ribbons. The connections $H_1$ and $H_2$ are shown attached to the backbone $B$ so that the orientations of the ends and side arcs cancel.

A twisted strip is an ordered union of twisted annuli attached end-to-end, with the orientations of the ends canceling. A twisted strip is shown in fig. 5(a). A twisted strip is determined by a three-letter sequence on the letters $U$, $R$, and $L$. The arc $ab$ shown in fig. 5(a) is referred to as the spanning arc. The spanning arc is contained in the union of the subgraphs of the twisted annuli. The ends of the first and last twisted annulus that lie in the boundary of the strip are the first and last ends of the strip, respectively.

A twisted surface is illustrated in fig. 5(b). A twisted surface is a union of twisted strips $B, H_1, \ldots, H_k$ satisfying: (1) each $H_i$ is attached to $B$ at its ends and along side arcs in $B$, (2) the $H_i$ are pairwise disjoint, (3) the first end of $H_i$ is attached to a twisted annulus which precedes (in the order of $B$) a twisted annulus to which the last end of $H_i$ is attached, (4) the orientation of the ends of the $H_i$ cancels the orientation of the side arcs of the cells of $B$, and (5) the endpoints of the spanning arcs of the $H_i$ meet $B$ at the ends of annuli in $B$. The strip $B$ is referred to as the backbone and the $H_i$ are referred to as connections. The subgraph of a twisted surface is the union of the spanning arcs of its strips together with the edges $am$ (see fig. 3) of the annuli in the backbone that connect the spanning arcs.

Figure 6 shows two equivalent embeddings of a twisted surface. Figure 6(b) is obtained from fig. 6(a) by sliding the edge $ab$ completely around the loop as indicated by the arrows. This introduces a full twist as shown. Handle sliding is prevented in a twisted surface by the subgraph. The number of edges incident at

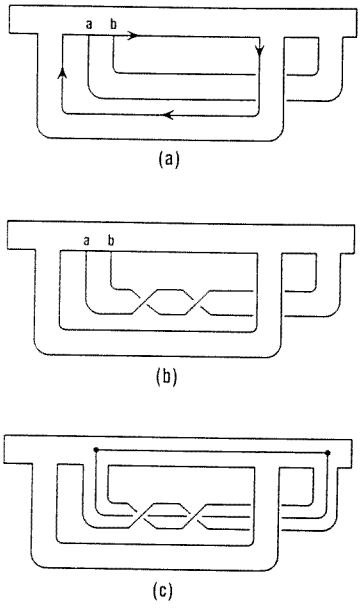


Fig. 6. Topologically equivalent embeddings of a surface are portrayed. (b) is obtained from (a) by sliding the handle attached at *ab* around the loop in the direction of the arrows. (c) shows the cycle in the subgraph that becomes linked with a boundary curve making the two twisted surface embeddings inequivalent because the subgraph is defined as part of the twisted surface.

a vertex will change when two vertices coincide during the slide. Note that the boundary component of fig. 6(b) links a cycle in the subgraph, as shown in fig. 6(c).

The backbone strip in a twisted surface representation of a protein corresponds to the ribbon shown in fig. 1(b). Connections in the twisted surface can be made at the α-carbons along the backbone. Connections can represent disulfide bridges or hydrogen bonds between sidechains in the protein. Corresponding connections are made in the twisted surface at the α-carbons where the sidechains are attached to the protein's backbone. Proximity connections can also be made between the α-carbons in close proximity. Connections representing hydrogen bonds can be made between oxygen and hydrogen atoms on the ribbon's edge.

The backbone strip becomes a twisted surface when the connections are added. The connections are also strips. This encodes twisting in the connections and serves to preserve the four possible orientations of the backbone at the connection.These orientations are shown in fig. 7. The four orientations are defined by changing the signs of two of the coordinates $x$, $y$, or $z$ in the local coordinate system. Or, equivalently, the four combinations may be obtained by choosing one of parallel or anti-parallel and one of top or bottom. None of these orientations are interchangeable.

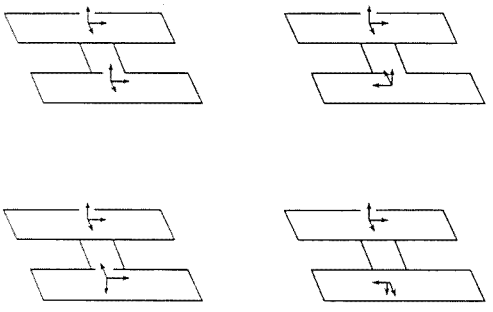

Fig. 7. Two horizontal segments of the backbone are shown with a vertical connection between them. The figures show the four possible orientations of the two local coordinate systems at the ends of the connection.

Suppose $M$ is a twisted surface in three-dimensional space and $P_1, \ldots, P_t$ are the images of the spanning arcs of the connections in $M$ under the projection into the $xy$-plane. Then $M$ is said to be in standard position if (1) the spanning arc of the backbone lies in the $x$-axis, (2) each $P_i$ is an arc in the bottom half of the $xy$-plane, (3) any pair of $P_i$ intersect in at most one point, (4) no $P_i$ meets the spanning arc of the backbone at an interior point of $P_i$, (5) each twisted annulus is in standard position, and (6) if $P_i$ crosses $P_j$ and $P_k$ with $i < j < k$, then $P_i$ first crosses $P_j$ and then $P_k$.

An *ambient isotopy* of three-dimensional space is a continuous family of homeomorphisms $h_t$ of the space onto itself, $0 \leq t \leq 1$, with $h_0$ being the identity function. An ambient isotopy defines a continuous movement of an object and its surrounding space in a way that preserves topological properties. The condition that the initial function ($t = 0$) be the identity prevents a reversal of orientation, which can occur when Euclidean space is mapped onto its mirror image. In the following, we shall assume that all twisted surfaces are ambient isotopic in $3$-space to a twisted surface in standard position. Continuously deforming a ribbon by an ambient isotopy keeps the ribbon topologically equivalent to its initial state at any point in time. This prevents the ribbon from coming apart, passing through itself, or reversing its orientation (i.e. "handedness"). The spanning arcs of a twisted surface in standard position project into the planes as arcs and can cross each other at most once. These conditions prevent small knots and links in structures (vide supra). Large knots or links can be formed in the subgraph of a twisted surface by using several connections.

## 4. Examples

Examples are given below to illustrate some of the known features of protein structure that are modeled by the twisted surface. Connections between two strands such as are found in the $\beta-\alpha-\beta$ crossovers shown in figs. 8(a) and 8(b) form a turn

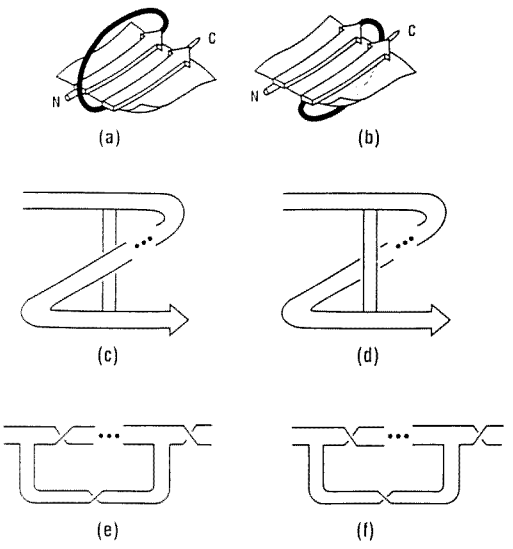


Fig. 8. Two adjacent parallel β-strands within a β-sheet are shown in (a) and (b). The backbone (heavy dark line) can pass either over, as in (a), or under, as in (b), in connecting the two strands. (c) and (d) illustrate the corresponding twisted surfaces with a single vertical connection joining two points along the backbone. The corresponding twisted surfaces in standard position are shown in (e) and (f) with their opposite twists.

which is either up or down. The right- and left-handed crossover connections are depicted in ARM form in figs. 8(c) and 8(d), respectively. The corresponding twisted surfaces in standard position are depicted in figs. 8(e) and 8(f), respectively. Note that each of the corresponding twists in the two surfaces are reversed. The two types of crossovers are distinguished by the different twists. The reader can create the twisted surface by making the crossover with paper strips and then straightening the backbone while holding down the ends of the backbone. A check for the number of twists, but not their handedness, can be made by following the boundary lines in the figures.

A loop connection is shown in fig. 9(a). When the backbone is straightened, the connection winds around as shown in fig. 9(b). A full twist in the backbone moves the connection below into the standard position as shown in fig. 9(c).

A schematical drawing of variant 3 neurotoxin [12] appears in fig. 10(a). A fictionalized structure is shown in fig. 10(b). The second structure is the same except that the backbone is shown crossing over instead of under itself. The corresponding twisted surfaces are shown in figs. 11(a) and 11(b). *The additional twists in fig. 11(b) illustrate the capability of the twisted surface to distinguish*

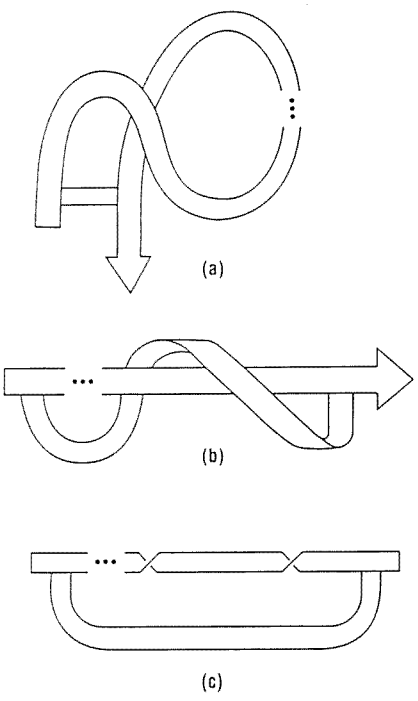


Fig. 9. Sketches illustrating how a loop connection is encoded. (a) illustrates the loop, (b) an intermediate step showing the connection winding around the backbone, and (c) the twisted surface in standard position with a full left-hand twist.

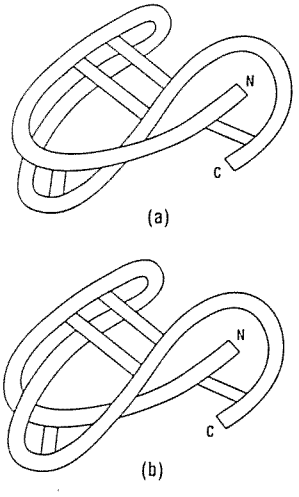


Fig. 10. (a) A simplified portrayal of variant 3 neurotoxin. (b) A fictionalized structure of the neurotoxin. The backbone ("ribbon") in (b) crosses over (instead of under) itself when viewed from the standard directionality of the polypeptide chain (see fig. 1(a)).

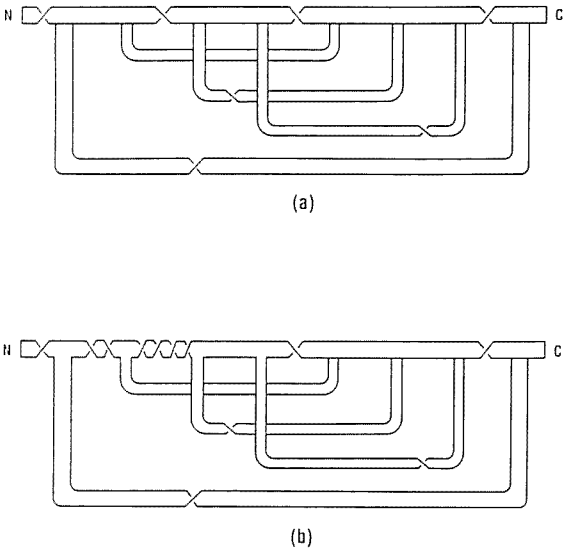


Fig. 11. (a) The twisted surfaces in fig. 10 depicted in standard position. The additional twists in (b) show that twisted surfaces can distinguish between over and under crossings. The similarity between the two surfaces is indicated by the amount of twisting that is identical between (a) and (b).

*between over and under crossings. If the model were constructed from lines instead of strips, the twisting would not appear and figs. 11(a) and 11(b) would be indistinguishable.*

In the above examples, the backbone of the protein is shown as a flat strip, which illustrates *free twists*. These twists occur in the model because of the three-dimensional structure of the protein. Free twists are allowed to cancel each other. *Fixed twists* in the backbone are modeled by twisted annuli and do not cancel. The next example illustrates how both appear in a protein structure.

A schematic drawing of crambin [13], which possess two α-helices and a small two-stranded β-sheet appears in fig. 12(a). When a ribbon is formed into a helix and then straightened, the twists in the ribbon become apparent. Figure 12(b) shows the twisted surface in a flattened state with connections made to the backbone strip. The connections correspond to three disulfide bridges and two strand–strand "proximity connections". The fifteen dashes on the backbone strip represent fixed twists. The twisted strip in its flattened state is easily straightened while the surface remains planar. Free twists are introduced into the backbone strip when the connections are moved below into standard positions. The sequence of fixed and free twists and their positions relative to the connections are shown in fig. 13. The free twists are denoted by primes. Each connection is untwisted. Figure 12 shows the ARM for crambin.

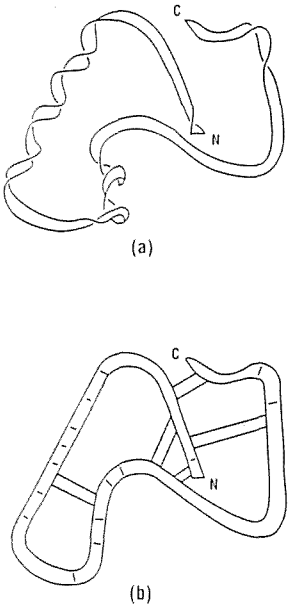


Fig. 12. (a) A sketch of the backbone of the protein crambin. (b) shows the backbone flattened with connections added. Fixed twists in (a) are indicated by transverse dashes along the ribbon surface in (b). Connections corresponding to the three disulfide bridges are shaded.

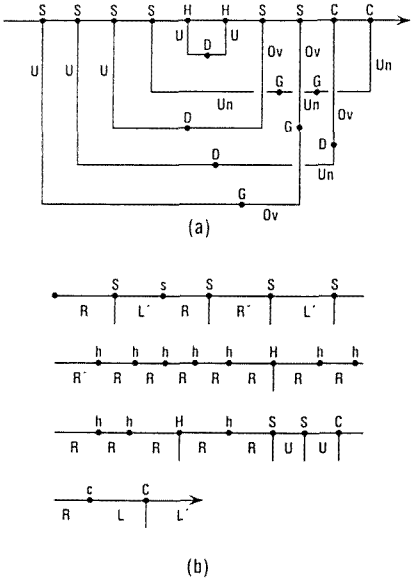


Fig. 13. (a) A labeled graph corresponding to the ARM of crambin shown in fig. 12(a). (b) Additional detail of the backbone labeled digraph – vertical lines correspond to the locations of the connections shown in (a).

## 5. Labeled digraph representation of the augmented ribbon model

The ARMs of proteins can be represented by labeled digraphs, which are similar to chemical graphs but contain considerably more structural information (vide supra). An ARM can be constructed as follows. The edges of the graph denote twists or crossings. Edges labeled $R'$, $L'$, and $U'$ together with their unprimed versions denote free and fixed twists, respectively. Free twists between two connections in the backbone or within a connection are placed before any fixed twists. Edges labeled *Ov* or *Un* denote over or under crossings of connections, respectively. Vertices are inserted to make the edges. Vertices in the backbone are labeled according to the conformational state in which they appear, such as helix *H*, strand *S*, turn *T*, and coil *C*. Vertices in connections can be labeled to distinguish between disulfide bridges *D*, hydrogen bonds *G*, or proximity connections *P*. Connections should have at least one vertex to indicate their type.

The model is flexible with regard to the amount and the type of information that is included. Connections are made at α-carbons, so some α-carbons appear as vertices. Vertices representing additional α-carbons can be added to the backbone. Vertices may be labeled by their residue type rather than conformation state. Chemical graphs for important sidechains under investigation can be attached to the α-carbons. The detail can extend to the chemical graph of the entire protein. On the other hand, very little detail may be included. An α-helix might be represented by a single vertex or the twisting in a section of coil ignored. Details of active sites can be included while the remaining structure is given only general form. Recently, Kaden et al. [14] have developed an interesting graph-theoretical representation of proteins which provides an alternative to the current approach but does not contain stereochemical information, as is the case in the work presented here.

## 6. Similarity

The ARM allows similarity methods to be applied to proteins just as for other organic compounds [15]. The twisted surfaces shown in figs, 11(a) and 11(b) illustrate how the ARM can be used to match similar protein structures. Most of the twists are identical in the two surfaces. The two strip models will have matching edges at these identical twists. The unmatched twists measure the difference in the two structures caused by the difference in the ribbon crossing over instead of under itself.

A data file of ARMs is created in the same way as a molecular data file. Some changes may be made to allow for more flexibility in the labeling of the atoms and bonds. Also, the prescribed order of the backbone needs to be followed. A program that determines the similarity of two protein structures is being adapted from one used to study small molecules [12].

## 7. Non-physical "topological" folding/unfolding of proteins

No attempt is made with the ARM to simulate the dynamics of real proteins in solution. The examples presented here were "unfolded by hand", as described above. We are currently developing software that should unfold proteins from a data base of known structures such as the Brookhaven Protein Data Bank [16]. It is instructive to shorten the connections and see a twisted surface fold back up into the original structure. The ARM explicitly shows the twisting necessary to attain the original conformation.

## 8. Summary

An augmented ribbon model of protein structure has been developed that provides a qualitative but mathematically rigorous description of many structural features such as connectivity, directionality, chirality, orientation, and proximity found at all levels of the protein structure hierarchy. Since ARMs can be represented by labeled digraphs, powerful, computable mathematical methods can be used to investigate protein structural features and their relationships that previously were difficult to analyze rigorously. Thus, the possibility is increased that new structural relationships may be discovered which heretofore would have been difficult, if not impossible, to uncover without the ARM.

## Acknowledgements

This work was supported in part by grants from the Upjohn Company. V.N. would like to thank Kent State University for sanctioning the study leave at the Upjohn Company. It is a pleasure to thank B. Mao and S.L. Narasimhan for many helpful discussions on the structure of proteins.

## References

[1] G.E. Schulz and R.H. Schirmer, *Principles of Protein Structure* (Springer, New York, 1979).
[2] B. Robson and J. Garnier, *Introduction to Proteins and Protein Engineering* (Elsevier, Amsterdam, 1986).
[3] G.M. Maggiora, B. Mao, K.C. Chou and S.L. Narasimhan, in: *Methods of Biochemical Analysis*, Vol. 35, ed. C.H. Suelter (Wiley, New York, 1990), p. 1.
[4] J.S. Richardson, Adv. Protein Chem. 34(1981)167.
[5] B. Mao, J. Amer. Chem. Soc. 111(1989)6132.
[6] B. Mao, K.C. Chou and G.M. Maggiora, Eur. J. Biochem. 188(1990)361.
[7] G.M. Maggiora and M.A. Johnson, in: *Concepts and Applications of Molecular Similarity*, ed. G.M. Maggiora and M.A. Johnson (Wiley, New York, 1990), p. 1.
[8] G.M. Maggiora, S.L. Narasimhan, C.A. Granatir and J.B. Moon, in: *NATO Advanced Research Workshop on Computational Models of Organic Chemistry*, ed. Formishino and Arnaut (Kluwer, Dordrecht, 1991), p. 137.

[9] M.L. Connolly, I.D. Kuntz and G.M. Crippen, Biopolymers 19(1980)1167.
[10] G.M. Maggiora, B. Mao and K.C. Chou, in: *New Directions in Molecular Chirality*, ed. P.G. Mezey (Kluwer, Dordrecht, 1990), p. 93.
[11] J. Stillwell, *Classical Topology and Combinatorial Group Theory* (Springer, New York, 1980).
[12] R.J. Almassy, J.C. Fontecilla-Camps, F.L. Suddath and C.E. Bugg, J. Mol. Biol. 170(1983)497.
[13] W.A. Hendrickson and M.M. Teeter, Nature 290(1981)107.
[14] F. Kaden, I. Koch and J. Selbig, J. Theor. Biol. 147(1990)85.
[15] V. Nicholson, C.-C. Tsai, M. Johnson and M. Naim, in: *Graph Theory and Topology in Chemistry*, ed. R.B. King and D.H. Rouvray (Elsevier, Amsterdam, 1987), p. 226.
[16] R. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, J. Mol. Biol. 112(1977)535.