J Math Chem (2010) 47:626–635
DOI 10.1007/s10910-009-9590-9

ORIGINAL PAPER

Protein folding invariants

Victor A. Nicholson

Received: 26 April 2009 / Accepted: 11 August 2009 / Published online: 19 September 2009

Abstract This article discusses the topological invariants associated with an augmented ribbon model of single domain protein. The model is a triple (S, G, J) in S^3 where S is a 2-manifold with boundary, G is a circle-with-chords, and J is an arc. The surfaces satisfy an embedding condition called laundry. The invariants are necessary and sufficient conditions for two triples to be equivalent by ambient isotopy. The model describes the native state, the unfolded state, and a unique folding pathway as a single mathematical entity. This may help illuminate some of the remarkable properties of protein. Twist transitions are introduced that allow the surface to pass through itself. A new arithmetic involving the complex numbers is used to represent variable linking numbers.

Keywords Surface · Linking · Protein

1 Introduction

There may be many different folding pathways for a polypeptide chain involving transient substructures. An augmented ribbon model [1] can marry the native state to a unique folding pathway that does not include transient structures. This suggests how the backbone of a protein must eventually move to reach the native state. In this paper there is no discussion of energy or kinetics. The interest is in describing the relation between protein structure, a mathematical model and a mathematical pathway.

V. A. Nicholson (✉)
Department of Mathematical Sciences, Kent State University, Kent, OH 44240, USA
e-mail: vnichols@kent.edu

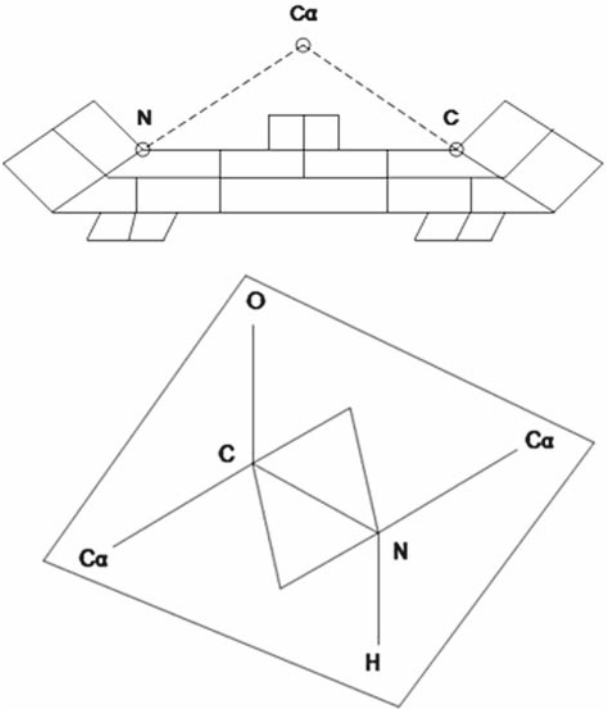

Fig. 1 The surface (S, G, J) at a single residue. The connecting patch within a peptide unit is illustrated in the *lower* figure

2 An augmented ribbon model is constructed from the native state

The small globular protein Crambin is used as an example to illustrate the method. An augmented ribbon model (S, G, J) is constructed for Crambin. The model consists of a surface S, a graph G, and an arc J. The upper figure in Fig. 1 illustrates the surface (S, G, J) at a single residue. The alpha carbon, nitrogen, and carbonyl carbon atoms are the vertices of a triangle. The edges of this triangle are extended by 0.9 Å through the C and N atoms adding a strip along the bottom edge of the triangle about 0.5 Å wide. The strip is composed of three planar patches. Two connecting patches are shown at the ends of the strip that would connect to adjacent residue strips, if present. The dihedral angles φ and ψ are the angles between the strip and the connecting patches at N and C, respectively. The lower figure in Fig. 1 illustrates that each connecting patch lies in the "plane" of a peptide unit. The line through the center of the patches extends the length of the protein backbone and forms the arc J. Each of the three patches in the strip is shown with a tab attached where a connection could appear. Figure 2 illustrates the five patches used to form a connection for a hydrogen bond. The oxygen and hydrogen atoms are used to locate the connection. Disulfide bonds and connections associated with side chain interactions are made at the second patch and use the alpha carbon atom.

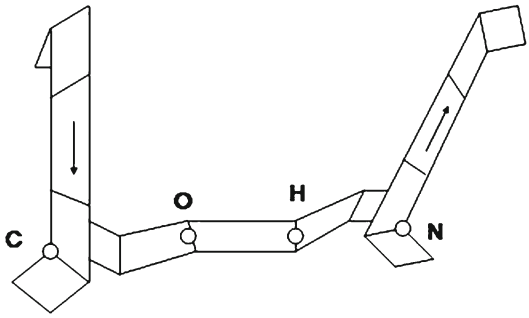

Fig. 2 The oxygen and hydrogen atoms are used to form the connection for a hydrogen bond

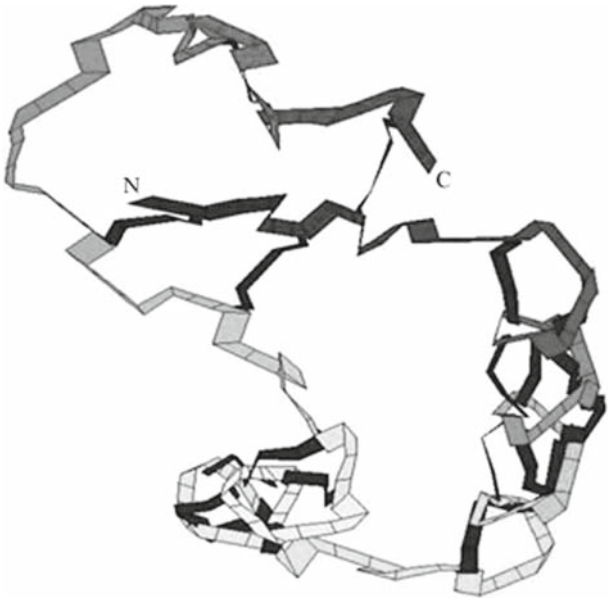

Fig. 3 The surface (S, G, J) as computed from a PDB file for Crambin (1CRN). Only hydrogen bonding is shown

Figure 3 shows the surface (S, G, J) as computed from a PDB file for Crambin (1CRN) with hydrogens added. Only hydrogen bonding is shown. The surface S is the collection of patches and the graph G is formed by the lines through the center of the patches. Figure 4 shows the graph G as a circle-with-chords in rectangular position. The arc J is the line across the top and is oriented from left to right to match that of the primary sequence from the N-terminal residue to the C-terminal residue as seen in Fig. 3. The graph edges that pass through the connections are the chords which hang below. Figure 1 does not show a close-up connection that connects the two ends of the backbone. The close-up connection forms the sides and bottom of Fig. 4 completing the circle in the graph G.

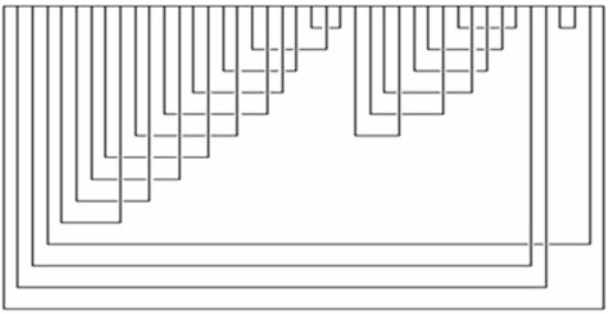

Fig. 4 The circle-with-chords graph G in rectangular position

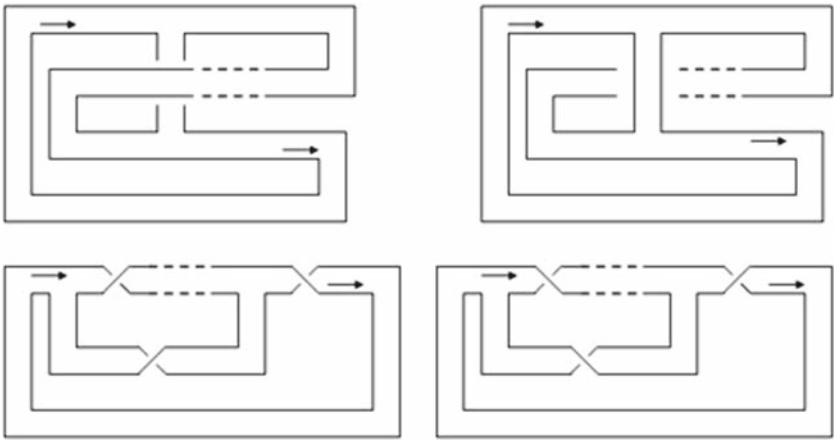

Fig. 5 Equivalent twisted surfaces are illustrated below each figure. The twists have opposite signs due to the ribbon passing over in the *left* figure and under in the *right* figure

3 A folding pathway as a motion carrying a twisted surface to an augmented ribbon model

Connections in the PDB are defined by ladders formed by nearby bonds. The essential feature of an augmented ribbon model is that the connections are ribbons. It is important to recognize that motion in mathematics (ambient isotopy) moves not only the object but the space that surrounds it. As the backbone ribbon moves the connections maintain their mathematical reality and properties. Unlike bonds, the connecting ribbons may shorten or lengthen and twist as the backbone moves.

Consider a simpler example. Figure 5 illustrates a connection between parallel strands. Straightening the strip between the two arrows (the backbone) while holding down the ends creates an equivalent *twisted surface* below each figure. Connections with an odd number of twists arise from parallel strands. The twists have opposite signs due to the backbone passing over in the left figure and under in the right figure. This example illustrates how the twist in a single connection in a beta sheet can indicate whether the backbone has moved over or under the sheet.

Consider a model of the folding of an single domain protein that gives the positions of the backbone atoms and computes the backbone ribbon as above. As the protein folds each of the bonds in the native state eventually appear and remain. Suppose that we simply have a movie of a moving ribbon. The usual notion of a folding pathway is one that carries a flat ribbon to the native state. In the view presented here folding begins with the twisted surface. The usual motion is preceded by one that untwists the backbone of the twisted surface causing the connections to wrap around the horizontal arc. Consider either twisted surface in Fig. 5. Apply a motion that untwists the backbone and then apply a motion F that shortens the connections and folds the ribbon. In each case F takes the plane ribbon and moves it to a backward letter "S". This gives an augmented ribbon model as described above.

If a pathway is to begin with a twisted surface how do we get a twisted surface? Note that the graph G in Fig. 4 is not tangled. Protein have such strong topological properties that a twisted surface can be computed using a system of equations directly from the native state. The next sections describe the role of these properties.

4 The linking matrix

Imagine a drawing of the projection into a plane of two disjoint oriented cycles R and T in space. At each point where the projected cycles intersect there is a crossing sign as shown in top of Fig. 6. The linking number of R and T is one-half the sum of the crossing signs. The linking number is an invariant that cannot be changed by simply moving the cycles in space. Note that in Fig. 4 each chord in G defines a cycle consisting of the chord and an arc in J between the chord's endpoints. The cycles are numbered using their first endpoint in the order of the orientation of J. Each cycle is oriented counter clockwise. For a cycle T in the surface S in space there is a cycle (or cycles) T′ obtained by lifting T off the surface in both directions. If T follows a Mobius band in the surface then T′ is a single cycle. If T follows an annulus then T′ consists of two cycles. For any cycle R, which may be T, the linking matrix value L(R, T) is one-half the sum of the crossing signs of R and T′. This lifted T′ is difficult to use in software. Instead, T′ can be represented by a cycle (or cycles) in the surface itself. To see this, imagine walking along the surface in the direction of T's orientation. Move the portion of the cycle above to your right and the portion below to your left. Figure 7 illustrates that T′ crosses the surface at the ends, Ta and Tb, of its connection. Otherwise, T′ lies in the boundary of the backbone ribbon. The cycle R is shown going down the middle of the surface at the ends of its connection, Ra and Rb. The crossing signs used to compute the linking matrix are indicated for each of the cases. These crossing signs are not computed because they always sum to zero. To see this, consider the possible configurations of the two intervals [Ra, Rb] and [Ta, Tb] on the arc J. If the intervals are disjoint there are no crossings. If the intervals overlap, there are crossings at the middle two. In the two cases Ra, Ta, Rb, Tb, and Ta, Ra, Tb, Rb the middle pairs cancel. If the intervals are nested, there are again crossings at the middle two. In the two cases Ra, Ta,Tb, Rb, and Ta, Ra, Rb, Tb the middle pairs cancel. In our approach, R and T′ are approximated on disjoint offset cubic grids. It is convenient to

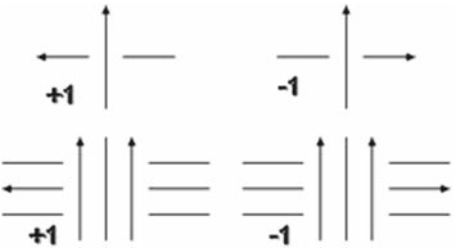

Fig. 6 Crossing signs are shown above. The two places where the boundary and the center curve cross in the surface is illustrated in the *lower* figure

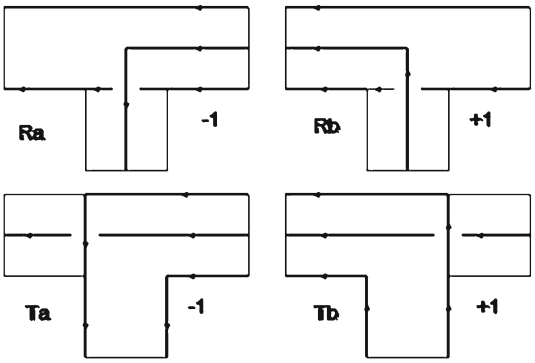

Fig. 7 The crossing signs of the cycle R in the center and T′ in the boundary always cancel

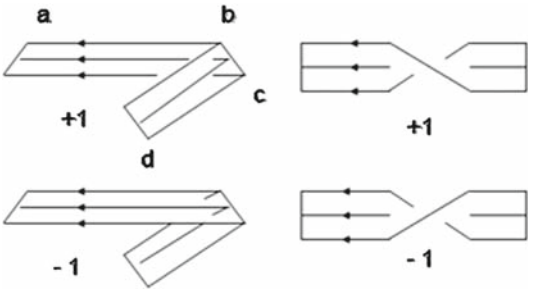

Fig. 8 The *upper* figures illustrate a plus one twist. The *lower* figures illustrate a minus one twist. A twist transition is a change from one to the other. The change in the linking number is plus or minus two

have all crossings occur at right angles when computing linking numbers. The linking matrix is an invariant of (S, G, J).

5 Twist transition and the variable linking matrix

Twist is illustrated as it usually appears on the right side of Fig. 8. Twists as they appear in Fig. 3 are illustrated on the left. When cycles R and T overlap each other on the arc J a part of T′ appears as the two edges in the boundary and a part of R appears down the middle. The cycles are oriented in parallel. There are two crossings at a twist. The twist is one-half the sum of these crossing numbers. A *twist transition* occurs when the surface passes through itself by a change from one twist to the other. That is, a transition between the upper and lower figures on the left in Fig. 8. This changes the crossing signs of the two cycles and the linking number by plus or minus two.

In order to deal with twist transition we introduce a new arithmetic. We define addition for a special collection of integer sequences. The notation r:2:s refers to the sequence of integers that begins with r, increases in steps of two, and ends with s, where $r \leq s$. Let Seq denote the set of all such sequences. Define addition of two sequences A and B in Seq as the set of all possible sums, $A + B = \{a + b : a \in A \text{ and } b \in B\}$. The set Seq is closed under addition. Let C denote the subset of complex numbers $a + ei$ where a and e are integers and $0 \leq e$. The set C is also closed under addition. The identification $a + ei = a - e : 2 : a + e$ is a one-to-one correspondence that preserves addition (a semigroup isomorphism). To see this, note that for any sequence r:2:s, $s = r + 2e$, for some $0 \leq e$. Let $a = r + e$. Then the sequence is $a - e : 2 : a + e = a + ei$. Note that a is the average value and e is the error. The sum of sequences $A = a - b : 2 : a + b$ and $B = c - d : 2 : c + d$ is the sequence $(a - b) + (c - d) : 2 : (a + b) + (c + d)$ or $(a + c) - (b + d) : 2 : (a + c) + (b + d)$. Which corresponds to the sum of the complex numbers a + bi and c + di. Thus, we able to add sequences by simply adding the complex numbers.

Each twist in Fig. 8 is shown on the left as it typically appears in the protein model at the ψ dihedral angle N–Ca–C–N. This is the dihedral angle a–b–c–d in Fig. 8. The angle is considered to be positive for a counter clockwise rotation of the front pair, a–b, when the back pair, c–d, is held fixed. The angles are measured from $-180°$ to $180°$. The sign of this dihedral angle is the opposite of the twist sign. The angle ψ can be close to zero in a *cis* conformation. In order to allow for error in the sign of the dihedral angle in the native state a twist transition is allowed at proline, alanine, and glycine residues. When the linking matrix L is computed the signs of the ψ dihedral angles are also computed. A positive dihedral angle means that the twist could change from minus one to one. Suppose the linking number L(R, T) = t. Suppose p is the largest number of possible twist transitions from minus one to one and m is the largest number of possible transitions from one to minus one. The minimum linking value would be $t - 2m$, the maximum would be $t + 2p$, and the possible values would be the integer sequence $t - 2m : 2 : t + 2p$. The average value $a = t + p - m$. The error is $e = (t + 2p) - (t + p - m) = p + m$. The sequence $a - e : 2 : a + e$ is the same sequence and corresponds to the complex number $a + ei$. The variable linking number VL(R, T) is defined to be $a + ei$. A variable linking number is simply the sequence of possible values. Since the primary sequence of amino acids is invariant the location of the above residues (and the patches defining the ψ angles) does not change. This makes the variable linking matrix, VL, invariant. The error reflects the property of protein that some regions are flexible and others rigid.

6 Turn

Turn is defined for each edge in the arc J by surrounding the edge by a box on the surface as illustrated in Fig. 9. The intersections of the boundary of the box with J are labeled “a” and those with the chords labeled “b”. The turn is determined by the permutation, (aabb) or (abab), obtained when traveling around the box. The turn is zero in the first case and one in the second. For the protein model, turn is computed by considering the way the patches attach to each other along the backbone. Turn is

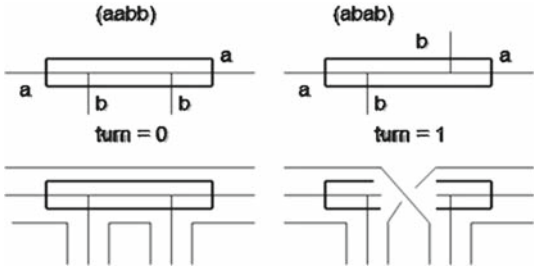

Fig. 9 An edge is surrounded by a box on the surface. The intersections of the boundary of the box with the arc are labeled "a" and those with the chords labeled "b". The turn is determined by the two possible permutations obtained when traveling around the box. In rectangular position turn is the twist modulo two

an invariant that is not affected by the motion of the surface in space. Also, turn is not affected by a twist transition.

7 The laundry embedding theorem

The central mathematical feature of protein is probably that the amino acids appear as an ordered sequence along the backbone. This property was used to orient the arc J. Note that each chord in Fig. 4 meets any other chord at most once and it meets subsequent chords in the same order as their order along the arc J. This is particularly evident in the two staircases formed by the helices. This property is referred to as the laundry condition. The strings of a tennis racquet are a common example. The graph G was fitted onto a cuboctahedral lattice [2] and software was used to move the graph in space so that it coincides with the rectangular graph in Fig. 4. The motion consisted of the two stages discussed earlier, straightening the backbone arc J and then unwrapping the connections. This shows that the graph G (and by definition, the surface S) is laundry embedded. Being laundry embedded is an invariant. This is an important mathematical property because the invariants considered earlier become not only necessary but sufficient. Among the family of laundry surfaces, they are a complete invariant. The laundry embedding theorem [1]: Two laundry surfaces are equivalent if they have the same graph, linking matrix, and turns. The equivalence is ambient isotopy, that is, one surface can be moved to the other. This means that we needed to move the graph G to rectangular position but that we do not need to move the surface to rectangular position.

8 The equations for laundry embeddings

When the embedding is laundry the invariants (graph, linking matrix, and turns) give rise to a system of equations that always has a solution defining the twisted surface. A complete description of the twisted surface is given by the sequence of chord endpoints, of twists along the arc, the chord twists, and the crossing signs. The endpoints of the 21 chords form the sequence: 1, 2, 3, 4, 5, 6, 7, 8, 5, 9, 6, 10, 7, 11, 8, 12, 9, 13, 10, 11, 12, 14, 13, 14, 15, 16, 17, 15, 18, 19, 16, 20, 17, 18, 19, 20, 3, 2, 21, 21, 4, 1. The twists on the 41 edges: 0, 0, −1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 2, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 12, 0, 0, 3, 2, 0. The twists on the chords: 0, 0, 0,

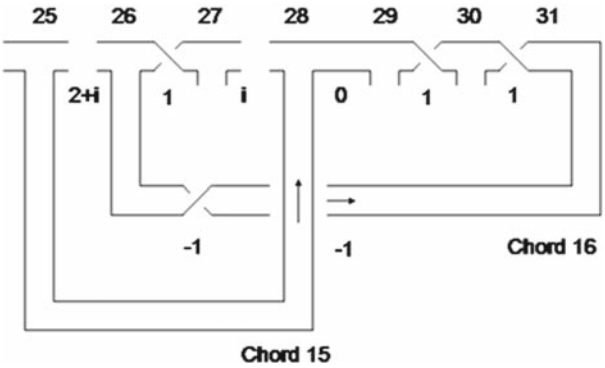

Fig. 10 The variable twists shown are i = {−1, 1} and 2 + i = {1, 3}. The twists in chord 15 and chord 16 are 0 and −1, respectively. The linking values VL(15, 15) = 3 + 2i and VL(16, 16) = 2 + i are the sums of the twists on the arc and the twists in the chords. The crossing number of chords 15 and 16 is −1. The variable linking value VL(15, 16) = i is the sum of the common twists on the arc and the crossing number

0, −1, −1, −1, −1, −1, −1, −1, −1, −1, 0, 0, −1, −1, −1, −1, −1, 0. The crossing signs are all minus one. The equations have additional solutions that correspond to flipping the chords around the arc like the pages in a book. These are equivalent embeddings, but with additional twisting. We have chosen a solution with minimal twisting.

Considering the signs of the ψ angles at the proline, alanine, and glycine residues in Crambin yields the variable twists on the 41 edges: 0, 0, −1, −1 + i, 1, 1, 1, i, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1 + i, 1 + i, 2 + i, 1, i, 0, 1, 1, 0, i, 1, 1, 1, 11 + i, 0, −1 + 3i, 3 + 2i, −1 + i, 0. Each twist given above is one of the possible variable twist values. Note that the error in twist does not affect the chord twists or the crossing signs.

Since the invariants are complete they can be recovered from any example. Figure 10 illustrates the twisted surface near chords 15 and 16. Cycle 15 is oriented counter clockwise along chord 15 beginning at point 25 and then back along the arc from point 28 to 25. The linking number of chord 15 with itself, VL(15, 15) = 3 + 2i, is the sum of the twists along the arc and the twist in the chord: (2 + i) + 1 + i + 0. Similarly, cycle 16 is oriented counter clockwise along chord 16 beginning at point 26 and then back along the arc from point 31 to 26. The linking number of chord 16 with itself is VL(16, 16) = 2 + i. The crossing number for these two cycles is minus one as is indicated by the two arrows. The linking number for these two cycles, VL(15, 16) = i, is the sum of the twists along the arc that are common to the two cycles and the crossing sign: 1 + i − 1. For the twisted surface the turn of an edge is the twist modulo two. Note that when using variable twist, say 1 + i = {0, 2}, that the turn is zero not one because the average is not a possible value. These relations between the twists and crossings to the linking numbers and turns define equations that can be solved in either direction.

9 Uniqueness of the twisted surface

Changing the crossing sign of chords 15 and 16 in Fig. 10 or the twist in chord 16 would not change the fact that the graph G is laundry embedded. The surface would be

a different laundry surface so it must have different invariants. If we were to shorten the chords and fold the structure it could not return to the model of Crambin. Conversely, starting with Crambin and following two different routes back to a twisted surface must yield an equivalent surface.

10 Multiple models for a protein

Since the model is a surface there can be only one connection at a patch. Only patch two is available for connections representing side chain interactions at a residue. This side chain may be involved with as many as ten others [2]. Multiple surfaces can include connections associated with disulfide bonds and residue packing.

The fact that proteins fold rather than tie themselves like knots suggests that bonds which would create embeddings that are not laundry may not be common. If the graph is not laundry embedded it may mean that a chord meets another chord more than once or it meets some subsequent chord in the wrong order. Again, multiple models using different sets of chords may be a solution.

References

1. V.A. Nicholson, J. Math. Chem. **40**, 105–117 (2006)
2. G. Raghunathan, R.L. Jernigan, Protein Sci. **6**, 2072–2083 (1997)