Protein Science (1993), *2*, 41–54. Cambridge University Press. Printed in the USA.

Disulfide bonding patterns and protein topologies

CRAIG J. BENHAM AND M. SALEET JAFRI

Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York 10029

(RECEIVED August 4, 1992; REVISED MANUSCRIPT RECEIVED September 10, 1992)

Abstract

This paper examines the topological properties of protein disulfide bonding patterns. First, a description of these patterns in terms of partially directed graphs is developed. The topologically distinct disulfide bonding patterns available to a polypeptide chain containing n disulfide bonds are enumerated, and their symmetry and reducibility properties are examined. The theoretical probabilities are calculated that a randomly chosen pattern of n bonds will have any combination of symmetry and reducibility properties, given that all patterns have equal probability of being chosen. Next, the National Biomedical Research Foundation protein sequence and Brookhaven National Laboratories protein structure (PDB) databases are examined, and the occurrences of disulfide bonding patterns in them are determined. The frequencies of symmetric and/or reducible patterns are found to exceed theoretical predictions based on equiprobable pattern selection. Kauzmann's model, in which disulfide bonds form during random encounters as the chain assumes random coil conformations, finds that bonds are more likely to form with near neighbor cysteines than with remote cysteines. The observed frequencies of occurrence of disulfide patterns are found here to be virtually uncorrelated with the predictions of this alternative random bonding model. These results strongly suggest that disulfide bond pattern formation is not the result of random factors, but instead is a directed process.

Finally, the PDB structure database is examined to determine the extrinsic topologies of polypeptides containing disulfide bonds. A complete survey of all structures in the database found no instances in which two loops formed by disulfide bonds within the same polypeptide chain are topologically linked. Similarly, no instances are found in which two loops present on different polypeptide chains in a structure are catenated. Further, no examples of topologically knotted loops occur. In contrast, pseudolinking has been found to be a relatively frequent event. These results show a complete avoidance of nontrivial topological entanglements that is unlikely to be the result of chance events. A hypothesis is presented to account for some of these observations.

Keywords: covalent bond topology; entanglements; knots; protein structure

Topology is the branch of mathematics that studies those properties of shape that remain invariant under continuous deformations. Topological properties naturally subdivide into two types—those that derive from the intrinsic structure of the object under study, and those that relate to how that structure is embedded in space. For example, a closed circle has a different intrinsic topological structure than a finite line segment. One can convert a circle into a line segment only by introducing a cut, which is a discontinuous deformation. As these two structures have different intrinsic topologies, one naturally might expect them also to have different ranges of possible realizations in space. All embeddings of a finite linear segment in three-dimensional space are topologically equivalent in the sense that any one can be converted to any other by a continuous deformation. In particular, a segment cannot be topologically knotted, because any candidate knot can be undone without recourse to cutting. One need only pass the ends of the segment back through whatever loops have been formed, which is a continuous deformation. It follows that all geometric shapes having the topological structure of finite line segments are topologically equivalent, both intrinsically and in all spatial embeddings. In contrast, a closed circular curve can be knotted. Different knot types cannot be interconverted without introducing transient cuts. Two circular curves having distinct knot types differ only in the way they are embedded in space. Both have the same intrinsic topology, that of a closed circle.

The pattern of covalent connections among amino acid residues imparts topological structure to a polypeptide chain. (Small loops, such as those occurring in aromatic

Reprint requests to: Craig J. Benham, Department of Biomathematical Sciences, Box 1023, Mount Sinai School of Medicine, 1 Gustave Levy Place, New York, New York 10029.

rings, fused rings, and similar local structures, commonly are disregarded because their topologies show no variability.) Although a polypeptide chain is synthesized as a linear polymer, it need not always have the trivial intrinsic topology of a line segment. The formation of covalent disulfide bonds between cysteine residues within a polypeptide chain produces circular loops of covalent bonds (Thornton, 1981). These covalent self-associations impart nontrivial intrinsic topology to the polypeptide. Molecules containing such covalent loops also may have nontrivial embedded topologies. Possible examples include knotted loops, interlinked pairs of loops on the same polymeric backbone, catenanes between loops on different backbones, as well as other forms of entanglement (Crippen, 1974, 1975). As this paper treats only topological properties, loop penetrations that are not topological in character are not considered, although these also may be important in practice (Connolly et al., 1980; Klapper & Klapper, 1980).

The topological state of a molecule constrains its geometry in specific and potentially important ways (Meirovitch & Scheraga, 1981a,b; Kikuchi et al., 1986, 1989). A protein can fold only into those conformations that are consistent with its topology. This limits the portion of conformation space that a molecule containing disulfide bonds may sample. The change in entropy consequent on this restriction can stabilize the conformation, as demonstrated by the increase in denaturation temperature observed when a disulfide bond in introduced (Johnson et al., 1978). Moreover, the folding pathway of a protein may involve the transient or permanent formation of specific disulfide bonds that constrain the molecule in a way that directs it toward its correct final conformation (Creighton & Goldenberg, 1984; Scheraga et al., 1984; Weissman & Kim, 1991).

Disulfide bonding patterns and intrinsic topologies

Consider the distinct disulfide bonding patterns (i.e., states of connectivity) available to a polypeptide containing M cysteine residues. The backbone of this polymeric chain consists of the sequence of residues covalently connected through peptide bonds, which are oriented in the N → C direction. Covalent disulfide bonds may form between pairs of cysteines, with any single cysteine residue participating in at most one such bond. These disulfide bonds possess a chemical symmetry that does not endow them with a natural orientation.

A disulfide bonding pattern has the mathematical structure of a partially directed graph. The vertices of this graph are the C- and N-termini of the chain, plus each of the cysteine residues that participates in a disulfide bond. The edges of this graph are the covalent connections between these vertices. The polypeptide backbone of the molecule is comprised of directed edges, each oriented according to its N → C chemical direction, forming a unique, directed, unbranched tree that spans every vertex. Because disulfide bonds are unoriented, the edges corresponding to them are undirected. The end vertices have order one, and all others have order three. (The order of a vertex, also called its valence by graph theorists, is the number of edges that are connected to it.) The three edges impinging on an interior vertex have distinct properties: one edge is directed into the vertex, one is directed away from the vertex, and the edge corresponding to the disulfide bond is undirected. This formulation differs from earlier graph-theoretic treatments of disulfide bonding patterns in that here the direction corresponding to the chemical orientation of the polymeric backbone is included. Earlier approaches used undirected graphs only (Walba, 1985; Mao, 1989).

Disulfide bonding patterns may be depicted by drawing the polymer backbone as a straight line, oriented left to right in the N → C direction, with the disulfide bonds shown as interconnections between the vertices corresponding to the pairs of cysteine residues involved. When not indicated by arrows, the backbone orientation always is chosen to be left to right as described. When necessary the vertices may be numbered in the order they are encountered as the backbone is traversed in the direction assigned by its orientation. For example, the three different patterns containing two disulfide bonds are shown in Figure 1. Because we are concerned with topological properties relating to connectivity, not at present with metric properties, the numbers of residues in each part of the polymer chain are not relevant.

An alternative representation of a pattern labels the disulfide bonds alphabetically in the order they are first encountered, starting from the N-terminus. The pair of cysteines connected by a particular bond are given its alphabetic label. An n-bond pattern is specified by giving the sequence of letters associated with the bonded cysteines, as they are encountered when the chain is traversed

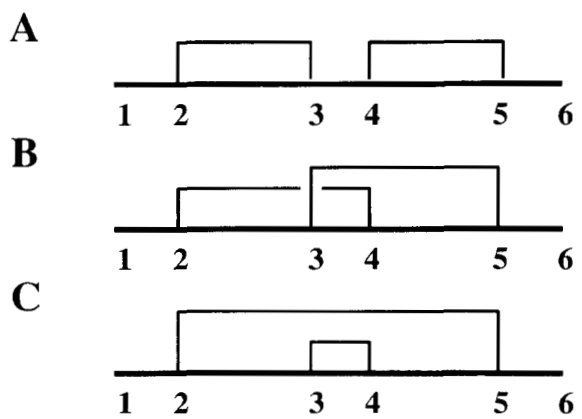

Fig. 1. The three different disulfide bond patterns in polypeptides containing two such bonds. All three patterns are symmetric, whereas only pattern A is reducible.

starting from the N-terminus. Thus, each pattern containing n disulfide bonds determines a sequence of length $2n$ whose entries are the first n letters of the alphabet, each of which appears twice, with new letters appearing in alphabetic order. In this notation the three two-bond disulfide patterns are *aabb*, *abab*, and *abba*.

In this paper the *pattern* associated with a state of disulfide bonding of a polypeptide chain is a partially directed graph of the type shown in Figure 1, having a unique directed spanning tree corresponding to the backbone. The *graph* associated with the pattern is the simple collection of edges and vertices shown, with no orientation and no distinction between different types of edges.

Two patterns have the same topological structure if one can be transformed into the other by a continuous deformation. This transition must preserve the directed nature of the polypeptide chain connections. Therefore its action on the directed backbone spanning tree is unique. In particular, it associates corresponding vertices in the order they are encountered along the chain. It maps directed edges to their corresponding directed edges, and disulfide bonds to disulfide bonds. It follows that two patterns are topologically equivalent exactly when all their disulfide bonds connect corresponding pairs of vertices. That is, only identical patterns are topologically equivalent. Two patterns are topologically distinct if no continuous transformation between them exists. This means that their interconversion requires the formation, disruption or rearrangement of disulfide bonds. Distinct patterns are always topologically nonequivalent.

It is important to note that the topological properties of patterns are not the same as the topological properties of their underlying graphs. Two graphs have the same intrinsic topology (i.e., are isomorphic) when there is a way of numbering the vertices of each so that corresponding edges join pairs of vertices having the same numbers in both graphs (Roberts, 1984). In graphs the numbering of vertices may be chosen arbitrarily and is not determined by a directed spanning tree (i.e., polypeptide backbone), as was the case for patterns. Thus, two topologically distinct patterns may have isomorphic underlying graphs. For example, two graphs that are mirror images are isomorphic, although asymmetric patterns in which the disulfide bonds occur in mirror image order are not topologically equivalent because the mirror image mapping does not preserve the backbone orientation. Another example of distinct patterns having isomorphic graphs is shown in Figure 2.

Disulfide bonding patterns have specific attributes that could be important for protein structure. One such property is symmetry. A pattern is symmetric if it and its mirror image both have the same disulfide bonding connections. Alternatively, the pattern is symmetric if its alphabetic representation reads the same when labels are assigned in the N → C direction as when they are assigned in the opposite direction. For example, all of the two-bond patterns are symmetric, although patterns with three or more disulfide bonds may be asymmetric, as is the case for both patterns shown in Figure 2. The second important property is reducibility. A reducible pattern is one in which a single cut somewhere along the backbone can separate the pattern into two nontrivial subpatterns. That is, some disulfide bonds occur entirely to the left of the cut point and others entirely to the right, but no disulfide bonds span the cut point. The pattern in Figure 1A comprised of two disjoint loops is reducible, whereas both of the other patterns are irreducible. A third intrinsic topological property of a disulfide bonding pattern is nonplanarity. A pattern is nonplanar if its graph cannot be drawn in a plane in a way in which no edges cross (Crippen, 1974). A pattern is nonplanar exactly when it contains the (sub-)pattern *abcdbcda*. (This topological definition of nonplanarity differs from that used by Kikuchi et al. [1986, 1989].)

In the following sections formulas are derived expressing the numbers of distinct (hence topologically nonequivalent) disulfide bond patterns, as well as the numbers of these that have all combinations of symmetry and reducibility properties. Intrinsic nonplanarity will not be considered in detail here, as it is less likely to be of practical importance in protein structure.

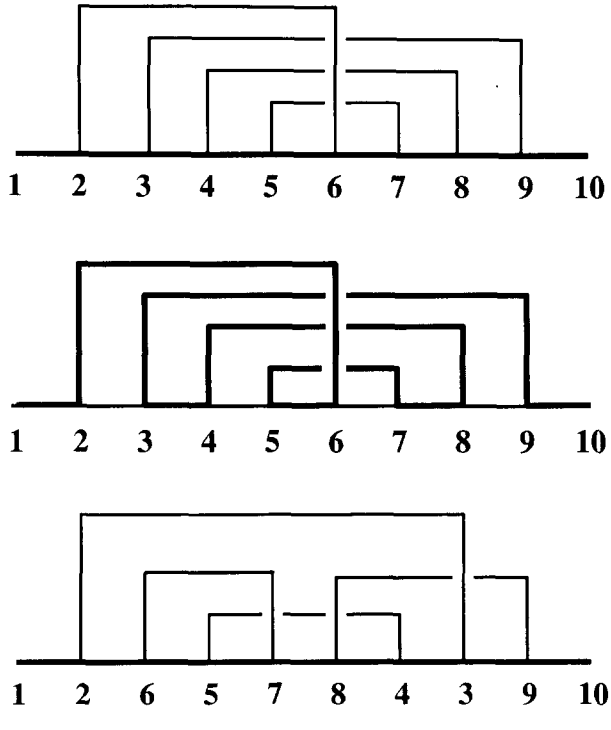

Fig. 2. An example of two different patterns whose underlying graphs are isomorphic. The top graph is the original pattern, where all edges now are regarded as undirected. If the vertices of this graph are visited along the path shown in the middle graph, and then this path is drawn as a straight line, the graph at the bottom results. Here the vertices retain their original numbering for clarity.

The number of disulfide bonding patterns

Consider a polypeptide chain containing M cysteine residues in which n disulfide bonds are formed, so $M \geq 2n$. The number of ways of choosing the $2n$ cysteines participating in the disulfide bonding is ${}_MC_{2n} = M!/(2n)!\,(M-2n)!$. Now suppose that the participating cysteines have been specified. The number of distinct patterns containing n disulfide bonds may be found by the following procedure (Cantor & Schimmel, 1980). Consider the participating cysteine nearest the N-terminus. There are $2n-1$ other cysteines to which it may be attached by a disulfide bond. Specify to which of these that bond is made. This leaves $2n-2$ cysteines whose disulfide connections remain to be determined. Of these, choose the unattached cysteine closest to the N-terminus. There are $2n-3$ possible choices for which other cysteine forms the disulfide bond with this one. Specify to which of these candidates that bond is to be made. Continue this process until all $2n$ cysteines have been connected. At the first step there were $2n-1$ choices, at the second $2n-3$, at the third $2n-5$, etc. The total number of choices is the product of all the odd numbers from 1 to $2n$:

$$P(n) = \prod_{i=1}^{n} (2i-1) = \frac{(2n)!}{2^n n!}. \qquad (1)$$

These equations give the number of different patterns containing n disulfide bonds. The factorial form of this expression was first presented by Kauzmann (1959). As noted above, all of these possibilities are topologically distinct as patterns, although some of their underlying graphs may be isomorphic.

The number of arrangements of n disulfide bonds among M cysteine residues on a polypeptide chain therefore is

$$\alpha(M,n) = {}_MC_{2n}P(n) = \frac{M!}{2^n n!\,(M-2n)!}, \qquad M \geq 2n. \qquad (2)$$

This expression was derived by Sela and Lifson (1959). Hereafter we will not consider cysteines that do not participate in disulfide bonds.

(In mathematics, an algebraic structure can be given to the set of patterns by defining a multiplication operation on them. However, it is not known whether the resulting construct, called the full connection monoid on $2n$ points [Kaufmann & Vogel, 1992], is relevant to protein structure.)

The number of symmetric patterns

The patterns involving n disulfide bonds may be classified according to whether or not they possess symmetry. This attribute may reflect (or dictate) a folding pattern having approximately symmetric regions or other regularities. Numbering the $2n$ bonded cysteines starting at the N-terminus, a pattern is symmetric if, whenever cysteines i and j are bonded, then so are cysteines $2n-i+1$ and $2n-j+1$. We note that this symmetry relates only to the topological pattern of disulfide bonding, not to metric properties such as the lengths of the polypeptide chain spanned by the bonds.

The number $S(n)$ of symmetric disulfide bonding patterns may be found as follows. All patterns containing either one or two disulfide bonds are symmetric, so $S(1) = 1$ and $S(2) = 3$. For the general case, we first enumerate those symmetric patterns in which a disulfide bond connects the first cysteine to the last (i.e., $2n$th) cysteine, as shown in Figure 3A. This is a symmetric arrangement of that bond. There remain $n-1$ other bonds to specify. For the entire pattern to be symmetric, these other bonds must be arranged in a symmetric manner. As there are $S(n-1)$ ways in which this can be done, this gives the number of symmetric patterns of this first type. Alternatively, suppose the pattern has a disulfide bond connecting the first cysteine to the jth cysteine, $j \neq 2n$. There are $(2n-2)$ choices for the cysteine to which this connection is made: only 1 and $2n$ are excluded. For the entire pattern to be symmetric, the $2n$th cysteine must be connected to the $2n-j+1$st cysteine, as shown in Figure 3B. Also, the remaining $(n-2)$ disulfide bonds must be arranged in a symmetric manner, which can be done in $S(n-2)$ ways. Hence the total number of symmetric patterns of this type is $(2n-2)S(n-2)$. Putting these results together, the

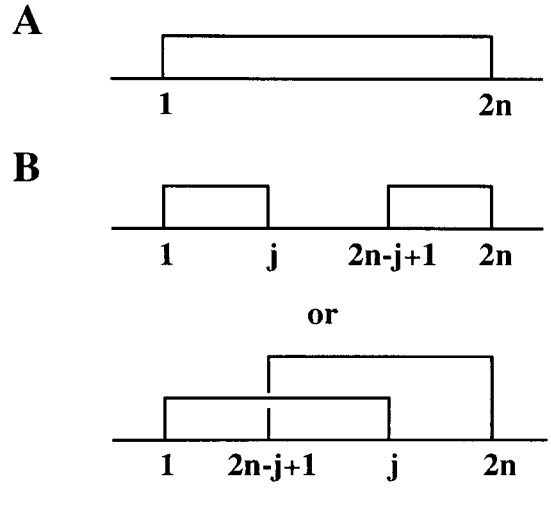

Fig. 3. The two cases encountered in the derivation of the recursion relation for $S(n)$, as described in the text. In the first case (**A**) a disulfide bond joins the first and last ($2n$th) cysteines, whereas in the second case (**B**) the first cysteine bonds to some cysteine other than the last. The disulfide bond shown in the first case is symmetric. However, in the second case the symmetry condition requires the presence of a mirror image disulfide bond as shown.

total number of symmetric disulfide bonding patterns is shown to obey the following recursion relation:

$$S(1) = 1,$$

$$S(2) = 3,$$

$$S(n) = S(n-1) + 2(n-1)S(n-2), \qquad n \geq 3. \quad (3)$$

This recursion relation may be solved explicitly, yielding the following closed form expression:

$$S(n) = \sum_{k=0}^{[n/2]} \frac{{}_nP_{2k}}{k!}. \quad (4)$$

(Here ${}_iP_j = i!/(i-j)!$ is the permutation of i objects taken j at a time, which is the number of different ways of choosing j objects, in order and without replacement, from a collection of i objects. Throughout this paper square brackets in equations denote the greatest integer function.)

The number of reducible patterns

A disulfide bonding pattern is reducible if it consists of two nonoverlapping, nontrivial subpatterns. In other words, if there is a site on the polypeptide backbone where a single cut will decompose the pattern into two subpatterns, then the pattern is reducible.

Recursion relations enumerating the reducible and irreducible patterns are derived as follows. A pattern containing n disulfide bonds is reducible exactly when it has at least one interior cut point, as described above. Traversing the sequence starting from the N-terminus, suppose the first such cut point that is encountered has i disulfide bonds on its N-terminal side and $n-i$ bonds on its C-terminal side, $1 \leq i < n$. Then the subpattern consisting of the i bonds on the N-terminal side must be irreducible, because this is the first cut site encountered. The subpattern comprised of the $n-i$ bonds on the C-terminal side of the cut can have any form, reducible or irreducible. So there are $P(n-i)$ choices for this pattern. Therefore the number of ways in which an n bond pattern can be chosen whose first cut site occurs as stated is the product $I(i)P(n-i)$, where $I(i)$ denotes the number of irreducible patterns with i bonds. For a pattern to be reducible it must have a cut point of this type at some position for which $1 \leq i \leq n-1$, so the total number $R(n)$ of reducible patterns is the sum

$$R(n) = \sum_{i=1}^{n-1} I(i)P(n-i),$$

$$I(n) = P(n) - R(n). \quad (5)$$

A similar calculation derives the recursion relation giving the number $S_r(n)$ of n-bond patterns that are both symmetric and reducible. Again, suppose the first cut point occurs after i bonds, so the number of choices for the subpattern of these initial bonds is $I(i)$. Because the complete pattern is symmetric as well as reducible, the last i bonds must be the mirror images of the first ones. It follows that $n \geq 2i$, and that the subpattern of the middle $n-2i$ bonds, if any, is all that remains to be determined. For the entire pattern to be symmetric, the subpattern of the middle $n-2i$ bonds must be symmetric. Hence there are $S(n-2i)$ choices for this structure. It follows that the number of symmetric, reducible patterns in which the first cut occurs after i bonds is the product $I(i)S(n-2i)$, so the total number of patterns that are both symmetric and reducible is

$$S_r(n) = \sum_{i=1}^{[n/2]} I(i)S(n-2i), \qquad n \geq 2. \quad (6)$$

The above results determine the number $A(n)$ of nonsymmetric patterns on n disulfide bonds to be

$$A(n) = P(n) - S(n). \quad (7)$$

Similarly, the number of patterns that are symmetric and irreducible is

$$S_i(n) = S(n) - S_r(n). \quad (8)$$

The number of nonsymmetric, reducible patterns is

$$A_r(n) = R(n) - S_r(n), \quad (9)$$

and the number of patterns that are both nonsymmetric and irreducible is

$$A_i(n) = A(n) - A_r(n). \quad (10)$$

Table 1 displays the numbers of patterns $P(n)$ containing n disulfide bonds, $1 \leq n \leq 12$, together with the numbers of these patterns that are symmetric, reducible, or both. From these values the numbers of patterns with all other combinations of symmetry and reducibility properties may be calculated according to the above equations.

Table 2 shows the fractions of patterns with given symmetry and reducibility properties for the cases $1 \leq n \leq 12$. These are the probabilities that a randomly chosen pattern of n disulfide bonds has the given attribute(s), provided every pattern is equally likely to be chosen. One sees that the fractions of patterns that are asymmetric or irreducible or both grow with n, while the fractions with all other combinations of attributes decrease. The probability of symmetry decreases rapidly as n grows, while the probability of reducibility decreases more slowly.

Table 1. *Number P(n) of patterns of n disulfide bonds, together with the numbers of these patterns possessing specific symmetry and reducibility properties*[a]

n	$P(n)$	$S(n)$	$R(n)$	$S_r(n)$
1	1	1	0	0
2	3	3	1	1
3	15	7	5	1
4	105	25	31	5
5	945	81	239	9
6	10,395	331	2,233	41
7	135,135	1,303	24,725	105
8	2,027,025	5,937	318,631	485
9	34,459,425	26,785	4,707,359	1,609
10	654,729,075	133,651	78,691,633	7,777
11	13,749,310,575	669,351	1,471,482,725	31,425
12	316,234,143,225	3,609,673	30,469,552,111	160,965

[a] These quantities were calculated using the methods described in the text.

Observed protein topologies

In this section we describe the results of database surveys evaluating the intrinsic and embedded topological properties of known polypeptide disulfide bonding patterns. The intrinsic topologies are given by the corresponding disulfide bonding patterns, whereas the embedded topological properties considered include knotting of loops and interlinking of pairs of loops. Intrinsic topologies are determined by disulfide bond connections alone, whereas the evaluation of embedded topologies requires knowledge of the structure of the protein.

Table 2. *Fractions of n-bond patterns having specific symmetry and reducibility properties*[a]

n	$p_s(n)$	$p_r(n)$	$p_{sr}(n)$	$p_{ar}(n)$	$p_{ai}(n)$
1	1.000000	0.000000	0.000000	0.000000	0.000000
2	1.000000	0.333333	0.333333	0.000000	0.000000
3	0.466667	0.333333	0.066667	0.266667	0.266667
4	0.238095	0.295238	0.047619	0.247619	0.514286
5	0.085714	0.252910	0.009524	0.243386	0.670899
6	0.031842	0.214815	0.003944	0.210871	0.757287
7	0.009642	0.182965	0.000777	0.182188	0.808170
8	0.002929	0.157191	0.000239	0.156952	0.840119
9	0.000777	0.136606	0.000047	0.136559	0.862664
10	0.000204	0.120190	0.000012	0.120178	0.879618
11	0.000049	0.107022	0.000002	0.107020	0.892931
12	0.000011	0.096351	0.000001	0.096351	0.903638

[a] In terms of the quantities calculated in Equations 1–10, these fractins are: $p_s(n) = S(n)/P(n)$, $p_r(n) = R(n)/P(n)$, $p_{sr}(n) = S_r(n)/P(n)$, $p_{ar}(n) = A_r(n)/P(n)$, and $p_{ai}(n) = A_i(n)/P(n)$. These fractions also give the probability that a randomly selected pattern has the corresponding set of attributes, provided all patterns have equal probabilities of selection. Here the subscript s stands for symmetric, a for asymmetric, r for reducible, and i for irreducible.

Intrinsic topologies—Disulfide bond patterns

Information regarding known disulfide bond patterns in proteins has been culled from two databases. The Brookhaven National Laboratories protein structural database (PDB) contains atomic coordinates for the structures of approximately 600 molecules (Berstein et al., 1977). Most of these structures have been found by crystallography, although some are theoretical predictions. In several cases a single database entry contains information on multiple subunits of the molecule, or on an additional molecule such as a bound inhibitor. A total of 259 protein molecules in the structural database were found to have disulfide bonds. This total includes duplicate entries, successive refinements of the same molecule, and entries for identical molecules from closely related species. Some structures are reported only for fragments of molecules or for molecules that have been altered by mutations affecting the number of cysteines present. In developing the population of observed structures examined here, theoretically predicted structures, mutated molecules, and fragments were removed from further consideration, as the information in the database does not specify the disulfide bonding pattern of the actual complete molecule in these cases. When duplicate and closely related entries also are deleted, a population of 62 distinct, complete polypeptide molecules containing disulfide bonds remains (listed in the kinemage file). The numbers of occurrences in this database of each type of observed disulfide bonding pattern are given in the fourth column of Table 3 below.

The National Biomedical Research Foundation (NBRF) protein sequence database (Barker et al., 1986) contains many thousands of entries, only some of which report disulfide bonding information. However, the absence of this information for a given molecule does not necessarily imply that it lacks disulfide bonds. In the small number of cases where disulfide bonding is reported, the accuracy of the pattern is not always known. Some entries rate bonds as certain, probable, or possible, whereas others give alternative possible disulfide bonding patterns. In some cases bonding patterns have been inferred by homology with other molecules. The disulfide bonding information derived from this database, although more plentiful than that found from the PDB structure database, must be regarded as being less reliable.

A total of 455 complete polypeptide chains in the NBRF sequence database were found to have intrachain disulfide bonds. This figure excludes fragmentary molecules and cases where considerable uncertainty regarding the disulfide connections was reported. Deletion of repeat entries and closely related molecules resulted in a population of 186 distinct polypeptides containing disulfide bonds. Column 3 of Table 3 reports the occurrences of each type of observed pattern in this population.

When the populations culled from the structure and sequence databases were amalgamated and duplicate entries

were deleted, an aggregate population of 208 distinct polypeptides containing disulfide bonds resulted. All occurrences of each type of disulfide bonding pattern in this aggregate population were determined. The results are given in column 5 of Table 3. Column 2 in this table gives the reducibility and symmetry properties of each observed pattern.

Table 4 shows the observed frequencies of disulfide bonding patterns having specific reducibility and symmetry attributes. The number of distinct occurrences of a given pattern is evaluated from the data of Table 3, separately for each database and also for the aggregate population. Also shown is the theoretical probability of each type of attribute, calculated using the expressions in the previous sections, assuming that each pattern of n bonds has equal probability of forming. These data show that, in cases where more than three disulfide bonds are present, symmetric patterns occur with frequencies that greatly exceed what would be predicted from random, equiprobable bonding. When $n \geq 6$, this frequency is an order of magnitude greater than random. This disparity is greatest for patterns that are both symmetric and reducible, which are overrepresented for all values of n. When $n \geq 6$, the prevalence of this type of pattern is two orders of magnitude greater than would arise with random bonding. In contrast, patterns that are irreducible are underrepresented at all values of n. As shown in Table 2, the probabilities of a randomly chosen pattern being asymmetric and/or irreducible all grow with n, whereas the probabilities of every other type of pattern decrease. However, the observed frequencies of asymmetric and/or irreducible patterns are much smaller than would be predicted if all patterns were equally likely to be chosen. These results clearly show that disulfide bonding patterns do not arise by random, equiprobable selection among all possibilities.

Table 5 shows the observed frequencies of occurrence of irreducible subpatterns as components of larger, reducible patterns. These data are derived from the aggregate population culled from both databases. It demonstrates that reducible structures arise predominantly through the catenation of short, irreducible subpatterns. Only three occurrences are seen of irreducible components containing more than three disulfide bonds, whereas 8 of the 10 possible irreducible three-bond patterns occur. Moreover, the data in Tables 3 and 4 show that polypeptide chains containing large numbers of disulfide bonds are found to occur predominantly, indeed for $n > 8$ exclusively, in reducible patterns. This suggests that such proteins are constructed from repeated iterations of subpatterns chosen from a small number of alternatives having few disulfide bonds.

The data presented above demonstrate that equiprobable random choice is unlikely to account for the observed distribution of disulfide bonding patterns. An alternative random bonding hypothesis has been formulated by Kauzmann (1959). He calculated theoretical frequencies of patterns, assuming bond formation occurred as cysteine pairs encounter each other during random coil fluctuations of a polymer chain containing equally spaced cysteines. According to the statistical theory of random coil polymers, the number of available configurations is reduced when a connection is made between two sites on the chain (viz. by a disulfide bond) that constrains the intervening segment to form a loop. Moreover, the number of available configurations becomes smaller as the separation along the chain between the connected sites increases. Accordingly, cysteine pairs that are near on the chain are more likely to encounter and bond than are remote ones. Therefore the distribution of patterns computed using this statistical-mechanical approach favors isolated bonds between neighboring cysteines. Figure 4 plots the theoretical probabilities of formation for all three-bond patterns as calculated by Kauzmann against the observed frequencies found here. Points plotted as circles give the data for all three-bond patterns, whereas points plotted as stars are for irreducible three-bond patterns that appear as components of larger, reducible patterns. In this diagram an observation that agrees with

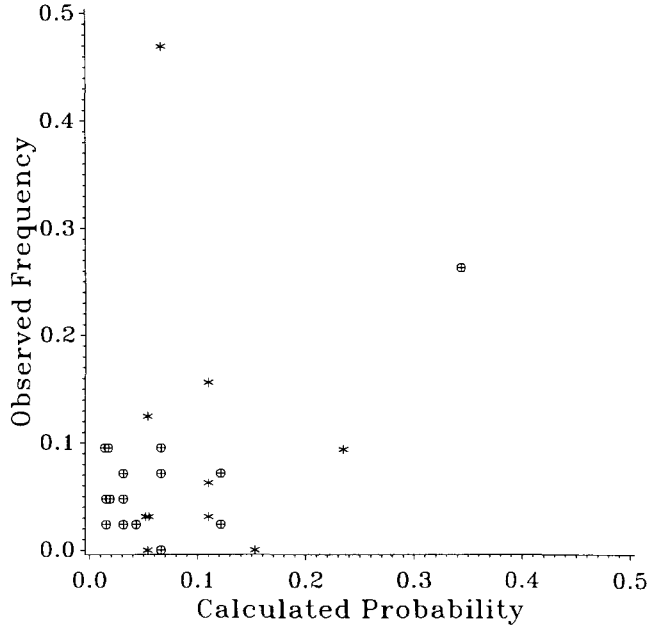

Fig. 4. Predicted and observed frequencies of all three-bond patterns are plotted. The theoretically predicted calculations assume Kauzmann's random bonding scheme, as described in the text. The observed frequencies are derived from the data presented in Tables 3 and 4. The circles plot all three-bond patterns, and the stars are for three-bond subpatterns that occur as irreducible components of larger reducible patterns. In the three cases where points superimposed, a small offset was introduced to show them both. Here a case where theory and observation agree would appear as a point falling on the diagonal line. One sees that the observed frequencies are virtually uncorrelated with the predictions from this model.

Table 3. *Numbers of occurrences of all intrachain disulfide bonding patterns in the NBRF and PDB databases, and in the composite population derived from both*[a]

n	Pattern	Attributes	NBRF	PDB	Composite
1	*aa*	*s, i*	51	20	66
2	*aabb*	*s, r*	19	8	20
	abab	*s, i*	8	1	8
	abba	*s, i*	5	0	5
3	*aabbcc*	*s, r*	11	4	11
	ababcc	*a, r*	4	1	4
	abcabc	*s, i*	4	1	4
	abacbc	*s, i*	1	0	1
	abccba	*s, i*	1	1	1
	abbacc	*a, r*	3	0	3
	aabccb	*a, r*	1	1	1
	abcacb	*a, i*	2	1	2
	abbcac	*a, i*	1	1	1
	abcbac	*a, i*	3	2	4
	abccab	*s, i*	3	0	3
	abaccb	*a, i*	2	0	2
	abcbca	*s, i*	2	1	2
	aabcbc	*a, r*	1	3	3
	abbcca	*s, i*	0	0	0
4	*aabbccdd*	*s, r*	1	1	1
	aabccdbd	*a, r*	3	1	3
	aabcddbc	*a, r*	1	0	1
	abbacddc	*s, r*	1	0	1
	ababcdcd	*s, r*	2	0	2
	aabcdbdc	*a, r*	1	0	1
	abcdcdba	*s, i*	2	2	2
	ababccdd	*a, r*	1	1	1
	abcddabc	*s, i*	1	1	1
	abcddcab	*a, i*	1	0	1
	abcddcba	*s, i*	1	0	1
	abcadbcd	*s, i*	1	0	1
	abcdbcda	*s, i*	1	1	1
	abcbdcda	*s, i*	1	0	1
	aabcbcdd	*s, r*	1	0	1
	abbccadd	*a, r*	0	1	1
5	*aabbccddee*	*s, r*	4	0	4
	ababccddee	*a, r*	1	0	1
	aabbcddece	*a, r*	1	0	1
	abbcaddece	*a, i*	2	1	2
	abccdbeead	*a, i*	1	0	1
	abccdebead	*a, i*	1	0	1
	abbacddece	*a, r*	1	1	1
	aabccbdeed	*a, r*	1	0	1
	abaccbddee	*a, r*	1	1	1
	abcdbedcea	*s, i*	1	0	1
	abcdadebce	*a, i*	1	0	1
	abbaccddee	*a, r*	1	0	1
	abcbdeecda	*a, i*	1	0	1
	abcdedabce	*a, i*	1	0	1
	abbcdaeedc	*a, i*	0	1	1
6	*aabbccddeeff*	*s, r*	1	0	1
	abacddbefcfe	*a, i*	1	0	1
	abbccadeeffd	*s, r*	1	0	1
	abbacdceedff	*a, r*	1	0	1
	abcbcadefefd	*s, r*	1	0	1
	abbcdaeefdfc	*a, i*	1	1	1
	abbccddeeffa	*s, i*	1	0	1
	abcbcadeefdf	*a, r*	1	0	1
	ababccddeeff	*a, r*	1	0	1

(*continued*)

Table 3. *Continued*

n	Pattern	Attributes	NBRF	PDB	Composite
7	*aabcdcdbefgfge*	*a, r*	1	0	1
	abcbdefgfgecad	*a, i*	1	1	1
	abccdadeeffggb	*a, i*	1	0	1
	abcdcefagfgedb	*a, i*	1	1	1
	abcddbefeggfca	*a, i*	1	0	1
8	*aabbccddeeffgghh*	*s, r*	1	0	1
	aabcddebeffgghhc	*a, r*	1	0	1
	abbccdeffgghheda	*a, i*	0	1	1
9	*aabbcdedecfghghfii*	*a, r*	1	0	1
11	*aabcdcbdeffeghghijjikk*	*a, r*	1	0	1
12	*aabbcdedecfghghfijjikkll*	*a, r*	1	0	1
14	$[abab]_7$	*s, r*	1	0	1
15	$[abab]_4cc[dede]_3$	*a, r*	1	0	1
	abbacdedecffghhgijklmnmnkoojli	*a, r*	1	0	1
16	$[abcabcdd]_4$	*a, r*	1	1	1
17	*abbacdcdefgfgheihijklkljmnonomppqq*	*a, r*	1	0	1
	$aa[bcbc]_8$	*a, r*	1	0	1
	$[abab]_2cc[defefd]_3ghhigi$	*a, r*	1	0	1
28	$[abab]_{14}$	*s, r*	1	0	1

[a] The symmetry (*s* or *a*) and reducibility (*r* or *i*) properties of each observed pattern are given.

theory yields a point that falls on the 45° diagonal line. One sees that the fit between the observed data and the predictions of this model is not good. Indeed, the plotted distribution of points shows virtually no correlation with the diagonal line. Kauzmann also calculated that the structure containing 17 disulfide bonds in the pattern $aa[bcbc]_8$ should occur with a frequency six orders of magnitude less than that of the pattern $[aa]_{17}$, in which all 17 disulfide bonds are disjoint. However, the former pattern, deemed highly improbable in Kauzmann's analysis, has been observed, whereas the latter pattern has not. These results strongly suggest that random encounters between cysteines of the type proposed by Kauzmann also are not the determinants of complete disulfide bond patterns.

A subsequent refinement of Kauzmann's model includes the effects of internal constraints, such as the presence of previously formed disulfide bonds (Chan & Dill, 1990). By extending this approach to average over all orders of formation one may be able to compute probabilities of patterns using a random coil model in a way that accounts for conformational freedom. This refinement will be considered elsewhere.

Two alternative scenarios by which random processes might dictate disulfide bond patterns have been shown here not to agree with observations. These are the equiprobable patterns model and the random encounters model of Kauzmann (1959). This suggests that disulfide bond patterns do not arise through random events, a conclusion also reached by Sela and Lifson (1959).

The results presented here regarding distributions of bonding patterns cannot be analyzed for statistical significance because the sample of protein molecules whose disulfide bond structures are known cannot be regarded as representative of all such structures. The patterns in the NBRF sequence database often have uncertainties associated with them and hence are not entirely reliable. Although the PDB structure database contains more exact data, the sample it provides is small and inherently biased by its limitation to crystallizable proteins.

This sample contains a single occurrence of a topologically nonplanar pattern: the scorpion neurotoxin protein (1SN3) has the pattern *abcdbcda*. This attribute had not been noted previously to occur in any known protein structure.

Embedded topologies—Linkages

A polypeptide chain containing two or more disulfide bonds in principle can assume conformations in which the resulting loops interlink. However, true topological linking is only possible if the disulfide bonds involved span disjoint portions of the polypeptide backbone. For example, only the reducible two-bonded pattern in Figure 1A above can experience topological linking. If the loops involved share a portion of the chain in common, as occurs

Table 4. *Observed frequencies of n-bond patterns with specific symmetry (s or a) and reducibility (r or i) properties, as derived from the NBRF and PDB databases, and from the aggregation of both*[a]

n	Attributes	Theory	NBRF	PDB	Combined
2	*s*	1.0	1.0	1.0	1.0
	r	0.333333	0.59375	0.888889	0.606061
	s, r	0.333333	0.59375	0.888889	0.606061
	s, i	0.666667	0.40625	0.111111	0.393939
3	*s*	0.466667	0.5641	0.4375	0.52341
	r	0.333333	0.51282	0.5625	0.52381
	s, r	0.066667	0.282051	0.25	0.261905
	s, i	0.4	0.282051	0.1875	0.261905
	a, r	0.266667	0.230769	0.3125	0.261905
	a, i	0.266667	0.205128	0.25	0.214286
4	*s*	0.238095	0.631579	0.625	0.6
	r	0.295238	0.578947	0.5	0.6
	s, r	0.047619	0.263158	0.125	0.25
	s, i	0.190476	0.368421	0.5	0.35
	a, r	0.247619	0.315789	0.375	0.35
	a, i	0.514286	0.052632	0.0	0.05
5	*s*	0.085714	0.277778	0.0	0.263158
	r	0.252910	0.555556	0.5	0.526316
	s, r	0.009524	0.222222	0.0	0.210526
	s, i	0.076190	0.055556	0.0	0.052632
	a, r	0.243386	0.333333	0.5	0.315789
	a, i	0.670899	0.388889	0.5	0.421053
6	*s*	0.031842	0.444444	0.0	0.444444
	r	0.214815	0.666667	0.0	0.666667
	s, r	0.003944	0.333333	0.0	0.333333
	s, i	0.027898	0.111111	0.0	0.111111
	a, r	0.210871	0.333333	0.0	0.333333
	a, i	0.757287	0.222222	1.0	0.222222
>6	*s*	<0.01	0.166667	0.0	0.157895
	r	<0.2	0.777778	0.25	0.736842
	s, r	<0.001	0.166667	0.0	0.157895
	s, i	<0.01	0.0	0.0	0.0
	a, r	<0.2	0.611111	0.25	0.578947
	a, i	>0.8	0.277778	0.75	0.263158

[a] Theoretical expected frequency is shown in each case, calculated assuming every pattern has an equal probability of occurrence.

Table 5. *Numbers of occurrences of irreducible patterns as components of longer, reducible patterns in the aggregate sample culled from both the NBRF and PDB databases*

n	Pattern	Occurrences
1	*aa*	75
2	*abab*	48
	abba	15
3	*abcbca*	15
	abbcac	5
	abaccb	2
	abcabc	4
	abcbac	1
	abbcca	3
	abccab	1
	abcacb	1
5	*abcbcdaede*	1
7	*abcdefefcggbda*	1
	abccdadeeffggb	1

for both irreducible two-bonded patterns, then rotations of one cysteine about this common region can alter apparent links, as is shown in Figure 5. Because this rotation is a continuous deformation, apparent linkage of nondisjoint bonds is not topological in character. Nontopological loop interpenetrations of this type, called pseudolinks, are known to occur in proteins (Kinemage 1; Kikuchi et al., 1986; Le Nguyen et al., 1990). However, because pseudolinkage is not a topological condition, its consideration will be deferred to a later time.

Topological linkage between two disjoint loops of a polypeptide can be determined from its molecular structure by evaluation of the linking number $\mathcal{L}$ of the loops. $\mathcal{L}$ is an integer topological invariant that measures the extent of interlinking of two disjoint closed loops in space. Unlinked pairs of loops always have $\mathcal{L} = 0$, whereas a non-zero value of $\mathcal{L}$ demonstrates topological linkage of the loops involved.

The linking number $\mathcal{L}$ of two disjoint loops may be calculated using the following Gaussian integral (Rolfsen, 1976). Let s_1 (resp. s_2) denote the contour length param-

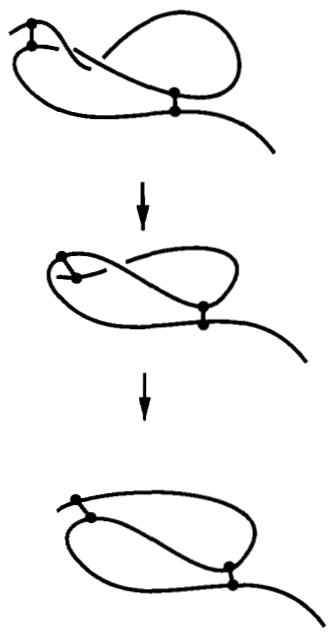

Fig. 5. Pseudolinking of nondisjoint loops is not a topological constraint. Such pseudolinks can be either induced or reversed by continuous deformations.

eter of loop 1 (resp. loop 2). That is, position in each loop is uniquely determined by measuring the distance s_i along that loop from some starting location, with $0 \le s_i \le L_i$, i = 1, 2. Let $\mathbf{r}(s_1, s_2)$ denote the vector joining the point s_1 on loop 1 to the point s_2 on loop 2. Denote by $\mathbf{e}(s_1, s_2)$ the unit vector $\mathbf{e} = \mathbf{r}/|\mathbf{r}|$. Finally, let $\mathbf{T}_1(s_1)$ (resp. $\mathbf{T}_2(s_2)$) denote the unit tangent vector to the loop at the point s_1 (resp. s_2). Then the linking number associated to these two loops is

$$\mathfrak{L} = \frac{1}{4\pi} \int_0^{L_1} \int_0^{L_2} \frac{\mathbf{e} \cdot \mathbf{T}_1 \times \mathbf{T}_2}{|\mathbf{r}^2|} \, ds_2 ds_1. \qquad (11)$$

In the calculations whose results are reported here, the loops formed by disulfide bonding were determined from the molecular structure using a virtual bond approach. The a-carbon positions of all the amino acid residues in that part of the polypeptide chain spanned by the disulfide bond were found, and the covalent peptide bonds were regarded as straight line connections between these a-carbons. The disulfide bond was considered to be a straight line connecting the α-carbons of the participating cysteines. In this way the loop formed by a disulfide bond spanning n residues (including the bonded cysteines) was modeled as a polygonal curve in space comprised of n straight segments. If loop *i* contains n_i segments, then the integral expressing the linking number decomposes into the sum of $n_1 \times n_2$ expressions,

$$\mathfrak{L} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathfrak{L}_{ij}, \qquad (12)$$

where each $\mathfrak{L}_{ij}$ has the form of the integral of Equation 11 above, evaluated over the ith segment of loop 1 and the jth segment of loop 2. Because these segments are regarded as straight virtual bonds, the tangent vectors $\mathbf{T}_1$ and $\mathbf{T}_2$ in each of the integrals $\mathfrak{L}_{ij}$ are constant. This virtual bond simplification has no effect on the computed results, as it does not alter the linkage of the loops.

These integrals may be evaluated in any of several ways. The double integral over straight segments can be solved analytically, and the resulting algebraic expression evaluated for every pair of segments. Because this expression is quite complex, in practice it is simpler to use a numerical routine to solve the integrals. Although this is less accurate than the algebraic method, it is simpler to program. Because $\mathfrak{L}$ is an integer-valued invariant, the level of precision needed in these calculations is only that required to distinguish neighboring integers. For this reason the slight degradation of accuracy consequent on performing numerical integrations is entirely inconsequential.

A computer program was written that searches the PDB structure database for all polypeptide chains having disjoint intrachain disulfide bonds. The linking number associated to each pair of disjoint loops in the chain was calculated by the procedure described above. Linking of loops was detected as a non-zero integer value of $\mathfrak{L}$. This procedure was carried through for all molecules in the database, including fragments, precursors, duplicate entries, and related molecules. A total of 209 database entries were found to have one or more pairs of intrachain disjoint loops. Some polypeptide chains contained large numbers of such loop pairs. The most extreme case was wheat germ agglutinin (3WGA), each subunit of which has 16 disulfide bonds in the pattern $[abcabcdd]_4$. This arrangement has 108 pairs of disjoint, intrachain loops. The calculation of linking number described above was performed on every pair of disjoint intrachain loops found. A total of 1,616 loop pairs were examined this way. In every case the loops involved were found not to be linked. True topological linking does not occur in any of the proteins currently represented in the PDB structure database.

A similar procedure was used to probe for catenation of loops on distinct polypeptide chains. All cases were examined where multiple chains containing disulfide bonds were reported in the same database entry. One would not expect catenation to occur in cases where the chains associate after their disulfide bonds are formed, such as between an enzyme and its inhibitor. However, in other cases catenation could occur. For example, cleavage of a precursor would result in catenated disulfide bonds if these bonds were linked in the precursor molecule. A total of 174 database entries were found that contained more than one disulfide-containing polypeptide chain. All pairs of loops on disjoint chains were examined for catenation by calculating their linking numbers using the process described above. In addition, all cases were examined where one of the loops arose by the formation of two *interchain* disulfide bonds. A total of 2,487 pairs of disjoint loops were examined in this way, and no instances of catenation were found in this population.

At present it is difficult to assess with any precision the probability that this lack of true linkage occurs by chance. Crippen (1975) has estimated the probability of linkage between disjoint loops in a random coil structure to be approximately 0.15. He simulated the chain as a self-avoiding random walk on a lattice, with the residues connected by disulfide bonds constrained to occupy neighboring lattice points. If this estimate were applied to the present sample, about 250 cases of intrachain linking would be expected, whereas none are found. Estimates of linking probabilities based on random chain statistics may not be applicable if the specific interactions involved with protein folding are not approximated by random fluctuations. For the reasons described in an earlier section it is not possible to generate a meaningful random sample of protein configurations from which linking probabilities could be estimated. Yet the complete absence of true linking in the present sample is striking.

It is important to note that neither topological link formation nor catenation require the protein chain to be

threaded through a preexisting loop. An alternative mechanism that can give rise to true linking is shown schematically in Figure 6. There an early stage of folding creates a helical interwind, which results in linkage of loops when the disulfide bonds form later.

In contrast to true linking, pseudolinking is known to occur in proteins with reasonably high frequency. Kikuchi et al. (1986) found four examples of pseudolinked proteins among 18 structures examined. Other cases of pseudolinking also have been reported (Le Nguyen et al., 1990). For pseudolinking to occur after disulfide bond formation, as Figure 5 shows may happen, threading of the polypeptide chain through a preexisting loop would be required. The close-packing that occurs in folded proteins may render such a motion sterically or energetically forbidden. Instead, pseudolinks probably arise primarily at the time of bond formation by a mechanism involving preexisting interwinding analogous to that shown in Figure 6 for topological linkage. Calculations of the relative stability of substructures within bovine pancreatic trypsin inhibitor indicate that their formation can occur prior to and independent of disulfide bonding, in support of the present claim (Chou et al., 1985).

The relatively frequent occurrence of pseudolinks in a small sample of examined structures makes the complete absence of true links in a much larger sample appear unlikely to result from chance. This impression is reinforced by the statistics of the relevant types of bond pairs. As shown in Table 3, there are more examples of reducible two-bond molecules, the pattern needed for true linkage, than there are of both types of irreducible two-bond patterns combined. Similar prevalences are found when one considers the irreducible components of reducible molecules. Thus, the absence of true links and the frequency of pseudolinks is not simply a consequence of the statistics of occurrence of disjoint versus nondisjoint bond pairs. These observations suggest that the rarity of true linking (none found in any known structure) may not be the result of chance factors operating in protein folding. Any speculations concerning why topological linking is disfavored must also account for the fact that nontopological pseudolinking is not disfavored. It remains to elucidate precisely why this would be true. The avoidance of topological linking becomes especially puzzling when one considers that the loop penetrations involved with pseudolinking may be virtually permanently fixed in the structure by steric or energetic constraints. Although they are not topological, these penetrations may be effectively as restrictive and permanent as if they were.

Fig. 6. Schematic illustration of a mechanism by which interlinked loops can occur through disulfide bonding that does not require threading the chain through preexisting loops.

We note that other types of entanglement than true linkage can occur between disjoint loops. Two examples are shown in Figure 7. The two disjoint loops shown in Figure 7A form a structure known in mathematics as the Whitehead tangle. These loops are topologically entangled because they cannot be separated without cleavage. However they are not linked as their linking number is zero. Similarly, Figure 7B shows the entanglement of three loops in a Borromean ring structure (Rolfsen, 1976). Here no pair of loops is linked: the removal of any single loop leaves the remaining two loops unentangled. But the three rings together are topologically entangled. These and other, higher forms of multiloop entanglement have not been evaluated in this examination of protein topology.

Embedded topologies—Knotting

In principle, a loop formed by disulfide bond closure could be knotted. The conformational intertwining that determines knotting must occur prior to loop closure by disulfide bond formation.

The PDB structural database has been searched for knotted polypeptides. A total of 103 different protein structures were examined, many of which did not contain disulfide bonds. (This was done because other constraints, such as β-sheet formation, can produce loops.) No examples of topological knotting were found in this case-by-case search. To date all structures that were reported to be knotted have proven to be pseudolinked (Le Nguyen et al., 1990). Again, nontrivial topology of this type has not been detected in the sample of molecules surveyed.

Crippen (1975) has estimated the probability of knot formation in a polypeptide loop modeled as a closed, self-avoiding random walk on a cubical lattice. He performed a Monte Carlo calculation to generate a sample of configurations and examined their topological properties. He found that the probability of knotting increases with

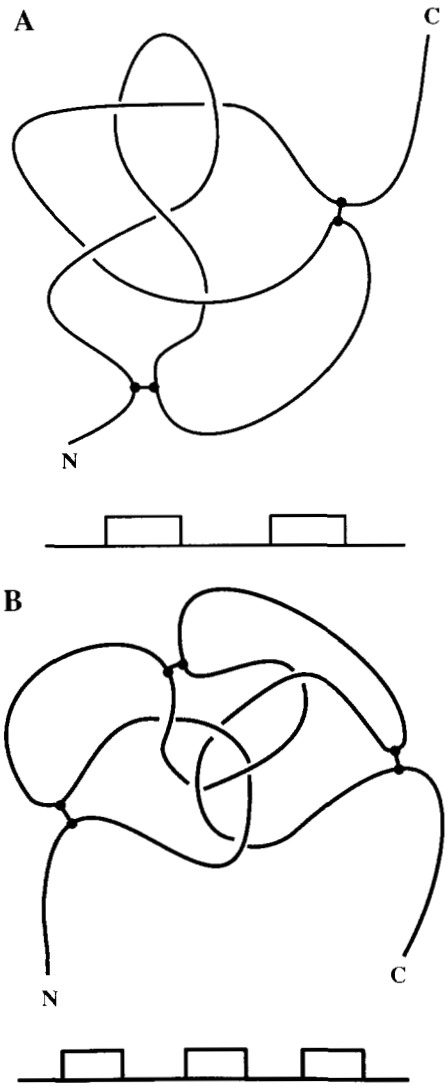

Fig. 7. Two types of nonlinked topological entanglement. The occurrence of these structures in proteins has not been evaluated.

chain length. Although the relative frequency of knotted structures found in this way was small (not exceeding 4%), Crippen speculated that it would approach unity were chain length to increase without limit.

An automated search procedure for knotting is presently being developed. Its results will be presented in a future paper.

Discussion

This paper examines the occurrence of intrinsic and embedded topologies in proteins, arising from the formation of covalent loops caused by disulfide bonding. It has been shown that symmetric and/or reducible patterns are highly overrepresented in the database, relative to amounts predicted were every pattern equally probable. For $n \geq 5$ the overabundances observed in this sample of structures are positively correlated: the conditional probabilities satisfy $p(S|R) > p(S)$, and $p(R|S) > p(R)$. Two models for random bonding were examined—the equiprobable patterns model and Kauzmann's random coil model. The predictions of both models regarding relative frequencies of patterns were shown to deviate sharply from the observed frequencies. These results strongly suggest that specific, nonrandom effects are involved in disulfide bond pattern selection.

The PDB structural database was examined for two types of nontrivial embedded topology—knots and interlinked loops. No examples of either type of nontrivial topology were found in this survey. This result suggests that specific mechanisms may exist for the avoidance of such extrinsic topological entanglements.

A hypothetical protein folding scenario can be proposed to explain the trends shown by these observations. Suppose that folding happens first at particular parts of a protein, called early folding regions (EFRs). If disulfide bonds were confined primarily or exclusively to such regions, and rarely or never joined two EFRs, the distributions of patterns and embedded topologies that result would have some of their observed properties. First, long proteins would be expected to contain several EFRs. If disulfide bonds were formed, their propensity to occur within EFRs would favor reducible patterns. Moreover, if long molecules were constructed from repetitions of one or a small number of EFRs, then an enrichment of symmetric patterns also would be seen. Regarding topological linking, this scenario would disfavor or preclude the folding pattern shown schematically in Figure 6 if the central, interwound portion of that structure is regarded as an EFR. The disulfide bonds connect that EFR to other parts of the molecule, which has been hypothesized to be disfavored or forbidden. For this conjectural explanation to hold, it is necessary that disulfide bonding occur at relatively early stages of protein folding, approximately simultaneous with EFR folding and prior to the interaction of EFRs with other parts of the molecule.

This hypothesis is consistent with the observed high frequencies of reducible and/or symmetric patterns and the paucity of nontrivial embedded topologies. However, the hypothesized tendency toward early disulfide bond formation is unlikely to completely preclude structures with any form of nontrivial embedded topology. This observed absence, if it remains correct as new structures are reported (and especially solution structures found by NMR), may require a more deterministic mechanism for the exclusion of topologically nontrivial conformations.

Investigations of other attributes of disulfide bonding patterns are presently in progress. These include automated searches of the database for knotted loops and for pseudolinked pairs of nondisjoint loops. The metric properties of disulfide bond patterns (i.e., attributes related to the positions of the participating cysteines along the chain) also are being examined at present. The plausibility of the

above hypothetical folding mechanism is being examined by evaluating the frequency with which disulfide bonds are seen to connect distinct domains in a multidomain protein. (We note, however, that the EFRs hypothesized above might not always, or even often, coincide with domains.) The results of these investigations will be reported in future contributions.

We note that the considerations relating to topological structure that have been developed here specifically for disulfide bonding also relate to other types of loop formation. Specifically, hydrogen bonding, either to form β-sheets in proteins or secondary structures in RNA molecules, produces loops whose topologies are subject to the same considerations as those presented here (Connolly et al., 1980; Richardson, 1985; Mao et al., 1990). However, there are four important differences between these cases and that of disulfide bonding. First, hydrogen bonds can form simultaneous associations between a given site and more than one other. RNA molecules can form triple-stranded helices, whereas runs of contiguous amino acids can form β-sheet associations with more than one other site. This creates the possibility of more complex topological structures, such as sheets closing to form barrels. Second, the susceptible sites in these cases are not equivalent. Local amino acid sequences vary in their propensities to form β-sheets, and secondary structure formation in single-stranded nucleic acids requires some degree of local sequence complementarity. This stands in contrast to disulfide bonding, where in principle all cysteines are approximately equivalent. Third, there is a directionality to these types of hydrogen-bonded self associations that is not present in disulfide bonding. β-Sheets can form in either of two orientations, parallel or antiparallel, whereas duplex formation in nucleic acids requires antiparallel orientations of the portions of the sequences involved. Finally, because these self associations involve relatively weak hydrogen bonds, the possibility exists that at equilibrium a population will occur in a distribution of configurations. Thus, dynamics and fluctuations may be important at equilibrium in these cases. In contrast, the disulfide bonds present in a fully folded polypeptide are thought to be permanent.

Acknowledgments

The authors gratefully acknowledge fruitful discussions and useful suggestions from Drs. George Némethy, Charles DiLisi (who first suggested this problem to C.J.B.), and Michael Waterman. The authors are especially grateful to Adrian Mogos for his able assistance in compiling the database information. This work was supported in part by grant DMB 88-96284 from the National Science Foundation.

References

Barker, W.C., Hunt, L.T., George, D., Yeh, L.S., Chen, H.R., Blomquist, M., Seibel-Ross, E., Elzanowski, A., Hong, M.K., Ferrick, D., Blair, J., Chen, S.L., & Ledley, R.S. (1986). *Protein Sequence Database*. National Biomedical Research Foundation, Washington, D.C.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rogers, J., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Cantor, C.R. & Schimmel, P. (1980). *Biophysical Chemistry*, Vol. 1, p. 292. W.H. Freeman, San Francisco, California.

Chan, H.S. & Dill, K.A. (1990). The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys. 92*, 3118–3135.

Chou, K.-C., Némethy, G., Pottle, M., & Scheraga, H. (1985). Folding of the twisted β-sheet in bovine pancreatic trypsin inhibitor. *Biochemistry 24*, 7948–7953.

Connolly, M., Kuntz, I., & Crippen, G. (1980). Linked and threaded loops in proteins. *Biopolymers 19*, 1167–1182.

Creighton, T. & Goldenberg, D. (1984). Kinetic role of a meta-stable native-like two-disulphide species in the folding transition of bovine pancreatic trypsin inhibitor. *J. Mol. Biol. 179*, 497–526.

Crippen, G. (1974). Topology of globular proteins. *J. Theor. Biol. 45*, 327–338.

Crippen, G. (1975). Topology of globular proteins. II. *J. Theor. Biol. 51*, 495–500.

Johnson, R., Adams, P., & Rupley, J. (1978). Thermodynamics of protein cross-links. *Biochemistry 17*, 1479–1484.

Kaufmann, L. & Vogel, P. (1992). Link polynomials and a graphical calculus. *J. Knot Theory 1*, 59–104.

Kauzmann, W. (1959). Relative probabilities of isomers in cystine-containing randomly coiled polypeptides. In *Sulfur in Proteins* (Benesch, R., Benesch, R.E., Boyer, P., Klotz, I., Middlebrook, W.R., Szent-Gyorgyi, A., & Schwarz, D.R., Eds.), pp. 93–108. Academic Press, New York.

Kikuchi, T., Némethy, G., & Scheraga, H. (1986). Spatial geometric arrangements of disulfide-crosslinked loops in proteins. *J. Comput. Chem. 7*, 67–88.

Kikuchi, T., Némethy, G., & Scheraga, H. (1989). Spatial geometric arrangements of disulfide-crosslinked loops in non-planar proteins. *J. Comput. Chem. 10*, 287–294.

Klapper, M. & Klapper, I. (1980). The 'knotting' problem in proteins. *Biochim. Biophys. Acta 626*, 97–105.

Le Nguyen, D., Heitz, A., Chiche, L., Castro, B., Boigegrain, R., Favel, A., & Coletti-Previero, M. (1990). Molecular recognition between serine proteases and new bioactive microproteins with a knotted structure. *Biochimie 72*, 431–435.

Mao, B. (1989). Molecular topology of multiple-disulfide polypeptide chains. *J. Am. Chem. Soc. 111*, 6132–6136.

Mao, B., Chou, K., & Maggiora, G. (1990). Topological analysis of hydrogen bonding in protein structure. *Eur. J. Biochem. 188*, 361–365.

Meirovitch, H. & Scheraga, H. (1981a). Introduction of short-range restrictions in a protein folding algorithm involving a long-range geometrical restriction and short-, medium-, and long-range interactions. *Proc. Natl. Acad. Sci. USA 78*, 6584–6587.

Meirovitch, H. & Scheraga, H. (1981b). An approach to the multiple-minimum problem in protein folding, involving a long-range geometrical restriction and short-, medium-, and long-range interactions. *Macromolecules 14*, 1250–1259.

Richardson, J. (1985). Describing patterns of protein tertiary structure. *Methods Enzymol. 115*, 341–358.

Roberts, F. (1984). *Applied Combinatorics*. Prentice-Hall, Englewood Cliffs, New Jersey.

Rolfsen, D. (1976). *Knots and Links*. Publish or Perish Press, Berkeley, California.

Scheraga, H., Konishi, Y., & Ooi, T. (1984). Multiple pathways for regenerating ribonuclease A. *Adv. Biophys. 18*, 21–41.

Sela, M. & Lifson, S. (1959). On the reformation of disulfide bridges in proteins. *Biochim. Biophys. Acta 36*, 471–478.

Thornton, J.M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol. 151*, 261–287.

Walba, D. (1985). Topological stereochemistry. *Tetrahedron 41*, 3161–3212.

Weissman, J. & Kim, P. (1991). Reexamination of the folding of BPTI. *Science 253*, 1386–1392.