separated into two fractions by reverse-phase chromatography on C-18 (Analytichem BondElut). The fraction eluted with water was subjected to HPLC on a Whatman PAC (amino–cyano) column with 0.1% trifluoroacetic acid (TFA) to give 2-oxobutanoic acid, glycine, threonine, glutamic acid, valine, *O*-methylthreonine, *N*-methylaspartic acid, and proline. Threonine and glutamic acid, which coeluted on the Whatman PAC column, were separated on a Whatman SCX (strong cation exchange) column with water; similarly, *N*-methylaspartic acid and proline coeluted on the PAC column but could be separated on the SCX column with 0.1% TFA in water.

The C-18 BondElut fraction eluted with MeOH was treated first with Ac_2O in pyridine followed by CH_2N_2 in CH_2Cl_2. Gradient HPLC of the resulting Champ-related *N*-acetyl methyl esters on silica with 1.0–50% EtOH in 1:1 hexane/CH_2Cl_2 gave a 1:1 mixture of methyl 3-acetamido-2,14-diacetoxy-4-methylpalmitates (**3**) which were isomeric at C-14.

The mixture of isomers of gross structure **3** had the following properties: $[\alpha]_D$ +17° (*c* 1.6, 1:1 MeCN/H_2O); EIMS *m/z* (relative intensity) 457 (0.1), 426 (5), 415 (3), 414 (9), 398 (6), 397 (17), 386 (15), 372 (10), 354 (14), 338 (15), 326 (96), 296 (11), 284 (15), 267 (49), 266 (100), 224 (43), 202 (25), 160 (49), 142 (50); high-resolution EIMS *m/z* 457.3014 (calcd for $C_{24}H_{43}NO_7$, −2.4 mmu error). ^1H NMR ($CDCl_3$) δ 5.50 (d, J = 10.1 Hz, NH), 5.00 (d, J = 6.3 Hz, H-2), 4.90 and 4.78 (two quintets, each of 0.5 H intensity, J = 6.3 Hz, H-14), 4.52 (ddd, J = 10.1, 6.3, 4.1 Hz, H-3), 3.72 (s, $COOCH_3$), 2.12 (s, CH_3CO), 2.04 and 2.02 (two singlets, each of 1.5 H intensity, CH_3COO on C-14), 2.00 (s, CH_3CO), 1.78 (m, H-4), 1.60 (m, 2 H on C-15), 1.25 (br m, 18 H on C-5 to C-13), 0.88 (d, J = 6.9 Hz, CH_3 on C-4), 0.86 (t, J = 7.5 Hz, H_3-16).

Acknowledgment. This research was supported by NSF Grant CHE-8800527.

Supplementary Material Available: ^{13}C and ^{15}N NMR data for 50% ^{13}C- and 90% ^{15}N-labeled **2** in methanol-d_4 and single-frequency continuous-wave ^{15}N-decoupled ^{13}C spectra of labeled **2** in methanol-d_4 (4 pages). Ordering information is given on any current masthead page.

Molecular Topology of Multiple-Disulfide Polypeptide Chains

Boryeu Mao

Contribution from Computational Chemistry, Upjohn Research Laboratories, Kalamazoo, Michigan 49001. *Received September 1, 1988*

Abstract: Molecular topology of the polypeptide chain in a stable folded protein is characterized from analyzing the graph representation of molecular covalent structure and the embedding of such a graph. A subset of topologies (i.e., embedded graphs) that do not contain knotted structures is enumerated and classified for some graphs. As a working hypothesis, it is proposed that topologies of the finite polypeptide chain in a stable, folded globular protein can be represented by this subset. Thus, for any polypeptide chain containing three or fewer disulfides, there is only *one topology* for the polypeptide. For some four-disulfide and some five-disulfide chains, their covalent structure graphs are intrinsically nonplanar (of genus 1), and in each case there are *two* enantiomorphic molecular topologies in the subset. Only one of the allowed topologies represents the stable folded tertiary structure of a protein; e.g., two mammalian-active neurotoxins from scorpions, variant 3 toxin from the North American species *Centruroides sculpturatus* Ewing and toxin II from North African scorpion *Androctonus australis* Hector, each a nonplanar four-disulfide chain, have the **D** topology. Topological stereoisomerism, i.e., the existence of multiple possibilities for the molecular topology of nonplanar polypeptide chains, indicates that the correct prediction of molecular topology must be a criterion for any scheme that predicts tertiary structure of these proteins.

Graph Representation of Polypeptide Chains

Proteins are polypeptide chains that are essentially linear structures except for disulfide bonds; each such disulfide bond links the side chains of two cysteine residues in the primary sequence. Disulfide bonds provide important structural stability in proteins, and the number of disulfide bonds found in single polypeptide chains of proteins varies from 0 to more than 12.[1] The covalent structure of a multiple-disulfide polypeptide chain is fully described by a graph in which each vertex represents the α-carbon atom of a disulfide-linked cysteine residue, and each edge represents a covalent linkage between two such cysteinyl C_α atoms.[1,2] An example of such a graph (henceforth *the covalent structure graph*) is shown in Figure 1a for a three-disulfide chain. In considering the topology of a covalent structure graph (i.e., an embedded covalent structure graph), labeling each vertex alphabetically and uniquely from the amino terminal to the carboxyl terminal essentially specifies that all edges are nonequivalent; i.e., they are edges of different colors. This is consistent with the fact that in general a partial amino acid sequence delimited by a pair of cysteine residues is different from one delimited by another pair; such labeling is also essential for the discussion of isomerism of molecular topology by topological graph theory (ref 3, and discussions in the Four-Disulfide Chains section).

Representation of a polypeptide covalent structure by a graph has been employed for studying knotting and loop penetration problems in protein structure. Crippen[2,4] estimated the probability of knotted structures in idealized polypeptide chains. Klapper and Klapper[1] described the knotting problem as a more general "loop penetration" phenomenon and also investigated the planarity/nonplanarity of different disulfide pairings in multiple-disulfide proteins. Connolly et al.[5] identified covalent and noncovalent loops in tertiary structure of proteins and studied linking and threading of such loops. Kikuchi et al.[6] devised schemes for identifying spatial arrangements of polypeptide fragments in tertiary structure of proteins. For any polypeptide chain containing three or fewer disulfide bonds (Figure 1a), the covalent structure

[1] Klapper, M. H.; Klapper, I. Z. *Biochim. Biophys. Acta* **1980**, *626*, 97–105.

[2] Crippen, G. M. *J. Theor. Biol.* **1974**, *45*, 327–338.

[3] (a) Walba, D. M. *Tetrahedron* **1985**, *41*, 3161–3212. (b) Walba, D. M. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvray, D. H., Eds.; Elsevier Science Publishers B. V.: Amsterdam, 1987; pp 23–42.

[4] Crippen, G. M. *J. Theor. Biol.* **1975**, *51*, 495–500.

[5] Connolly, M. L.; Kuntz, I. D.; Crippen, G. M. *Biopolymers* **1980**, *19*, 1167–1182.

[6] Kikuchi, T.; Nêmethy, G.; Scheraga, H. A. *J. Comput. Chem.* **1986**, *7*, 67–88.

0002-7863/89/1511-6132$01.50/0 © 1989 American Chemical Society

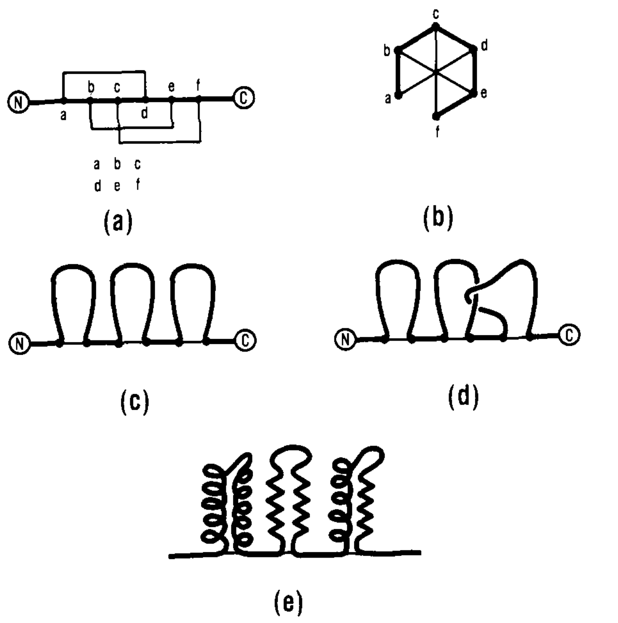

Figure 1. Disulfide linkages in a three-disulfide polypeptide chain. The covalent backbone is shown in thick lines and the disulfide linkages in thin lines. α-Carbon atoms of cysteine residues participating in disulfides are labeled alphabetically from the amino terminal, N, to the carboxyl terminal, C. (a) Disulfide pairing in the polypeptide chain. The set of three pairs (ad, be, cf) specifies the disulfide *pairing pattern* or *pairing scheme* for the six cysteine residues. (b) Graph representation of the chain in a. N- and C-terminal fragments (N-to-a and f-to-C, respectively) are omitted from the graph. (c) Planar three-disulfide chain. (d) Nonplanar embedding in which the second loop is linked with the third one. (e) Schematic drawing of secondary structures in a folded protein. Coils represent α-helices, wavy lines represent β-sheets. Packing of these secondary structure elements are also shown.

graph of the molecule (Figure 1b) is *planar*; the planarity is rigorously proved by observing that it is not a Kuratowski's graph, K(3,3) or K_5, and that it does not contain K(3,3) or K_5 subgraphs.[7] When additional disulfide bonds are incorporated into a polypeptide chain however, the number of possible cysteine pairing patterns increases combinatorially; moreover, the covalent structure graph for some chains becomes nonplanar. Figure 2a shows a nonplanar four-disulfide chain, the covalent structure graph of which is the Kuratowski's K(3,3) graph (Figure 2b).

Although the connectivity of a polypeptide chain is fully defined in its molecular structure graph, the folding of the chain in 3-dimensional space could introduce complexity in the graph embedding of the molecule. For example, the planar three-disulfide chain in Figure 1c in principle can fold into the structure shown in Figure 1d. Following the notion of Walba,[3] the two structures are *homeomorphic*, i.e., they have identical chain connectivity; they are however not *homeotopic*, i.e., they are not topologically equivalent. These two embeddings are two different *topologies* of the same graph (topological stereoisomers) that are allowed by its connectivity but cannot be made congruent with each other by continuous chain deformation in 3-dimensional space; the embedding in Figure 1d possesses *extrinsic* topological properties that the *intrinsically* planar structure in Figure 1c does not have. In this report, a subset of embeddings that do not contain such extrinsic topological complexities as linked loops or trefoil knots[3,8] or Kinoshita θ curves[9] is enumerated for planar and some nonplanar graphs. For a planar graph, this *topologically simple*[10] subset has only one topology; i.e., the embedding shown in Figure 1d and other complex topologies are excluded. For a nonplanar graph, however, more than one topology is allowed in the subset; these allowed molecular topologies, i.e., topological stereoisomers, fall into two families, **D** and **L**. A working hypothesis will be proposed that states that topologies of the finite polypeptide chain in a stable, folded globular protein are such a subset for embedding the representative covalent structure graph. Only one of the allowed topologies will describe the stable tertiary structure determined during the physical process of folding and disulfide formation.

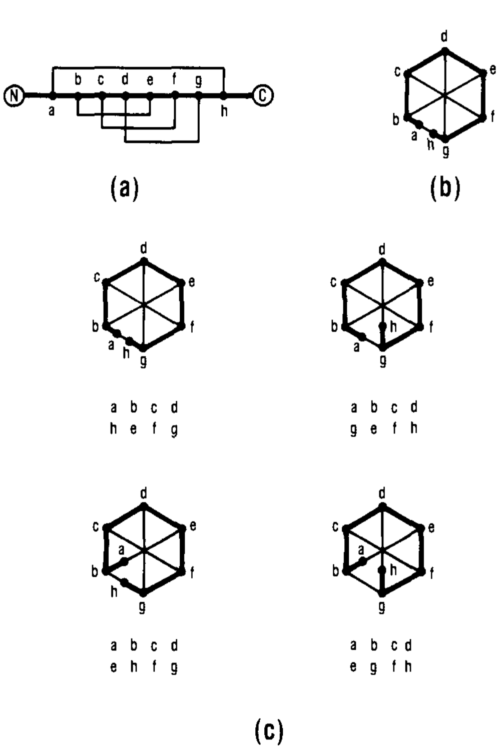

Figure 2. Nonplanar four-disulfide chain. (a) Polypeptide chain showing disulfide pairings. (b) Complete bipartite graph K(3,3), representing the chain in a. (c) Four four-disulfide pairing schemes that are nonplanar. Edge cf is observed in all four schemes.

Planar Proteins

Figure 1c,d showed an example in which an intrinsically planar graph becomes topologically complex due to linked loops. Knotted structures[10] such as linked loops and trefoil knots are *extrinsic* topological properties that are the results of graph embedding in 3-space.[3] In principle, embedding of the graph in Figure 1d could be even more complicated in that the second loop in the link can wind around the first one more than once; for even a planar graph, there are infinitely many topologies, only one of which (i.e., Figure 1c) constitutes the subset that is topologically simple.

In protein molecules, there are extensive side chain–side chain interactions between amino acid residues, and the polypeptide chains often fold into secondary structures such as α-helices and β-sheets; moreover, as shown schematically in Figure 1e, these secondary structure elements are connected by short loops and are packed together by hydrogen bonds between these secondary structure elements.[11] Thus, for a protein that is folded and of

[7] Chartrand, G.; Lesniak, L. *Graphs and Digraphs*; Wadsworth & Books: Monterey, CA, 1986.

[8] (a) Walba, D. M.; Richards, R. M.; Haltiwanger, R. C. *J. Am. Chem. Soc.* **1982**, *104*, 3219–3221. (b) Simon, J. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvray, D. H., Eds.; Elsevier Science Publishers B. V.: Amsterdam, 1987; pp 43–75. (c) Flapan, E. *Ibid.*, pp 76–81.

[9] Simon, J. *J. Comput. Chem.* **1987**, *8*, 718–726.

[10] In this article, a *knotted structure* will refer to any structure that is extrinsically complex. Thus it could be any type of linked loops, knots, or Kinoshita θ curves, or other types of extrinsic topologically complex structure. A mathematically more rigorous definition will require the identification of, among others, all the elements of topological dissymmetry.[3a] Operationally, for the relatively simple four- and five-disulfide cases discussed in this work, the procedure described in the Four-Disulfide Chains section, in conjunction of the elimination of trefoil knots (as in Figure 3c) and linked loops (as in Figure 5d and Figure 6d), determines a subset of molecular topologies that are *topologically simple*.

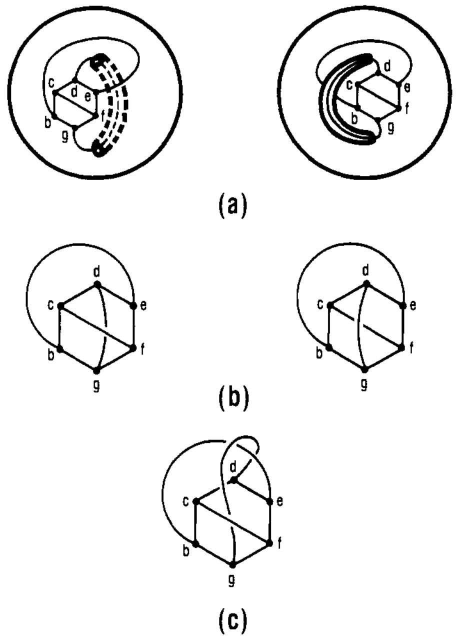

Figure 3. Embeddings of K(3,3) graph on a surface of genus 1. The *genus* of a surface is the number of handles that are attached to the surface for embedding topologically complex objects.[7] (a) Embedding of K(3,3) on the surface of genus 1, i.e., a sphere with *one* attached handle. On the left, the handle drawn in dashed lines is attached on the inside surface of the sphere; this topology is designated the **L** topology by the convention that if the thumb of a left hand points in the direction of edge cf, the fingers will follow the direction of edge dg. The topology on the right hand side by contrast is the **D** topology. (b) Simplified representations of the **L** and the **D** topology. (c) A topology with knotted structures. This is similar to the **L** topology in b except that edge dg makes an additional excursion around edge be. The loop d–c–f–e–b–g–d is a trefoil knot.

some finite length, *large open loops* from which *knotting or loop linking via disulfide formation* can occur are unlikely due to extensive interactions among amino acid residues. Since linked loops and trefoil knots in fact are not observed in proteins of known structure,[5,6] it is proposed, as a working hypotesis, that among the infinitely many embeddings of the covalent structure graph for a stable folded protein, only the subset that contains no extrinsic topological complexities will be allowed for the molecular topology. It follows then that, for each globular protein that is a stable folded polypeptide chain of finite length and that is planar, there is only one topology, that the topology does not have any knotted structures such as the link shown in Figure 1d, and that this topology is necessarily achiral.[3,8,9] A corollary is that all polypeptide chains that have an identical disulfide pairing scheme, and that are planar, are homeotopic.

Four-Disulfide Chains

Among the 105 possible four-disulfide patterns, four are represented by nonplanar graphs by virtue of their homeomorphism to Kuratowski K(3,3) graph;[1] these patterns are enumerated in Figure 2c. An interesting observation from this figure is that edge cf, an invariant among the graphs, is a necessary but not sufficient condition for the nonplanarity. It follows then that all four-disulfide polypeptide chains have graphs that can be embedded on a surface of genus 0 (planar) or genus 1 (nonplanar). Planar chains have one topology, as discussed in the previous section. A four-disulfide chain represented by any graph in Figure 2c however is *intrinsically* nonplanar and, excluding extrinsically complex ones, has two allowed topologies, shown in Figure 3a; that these two embeddings are topologically distinct (i.e., that they cannot be deformed into each other continuously in 3-space without breaking the connectivity) can be proved by following Simon's theorem on the chirality of Möbius ladders.[3b,12] The simplified representations in Figure 3b clearly show that the two topologies are in fact topological enantiomers,[3,8] i.e., topological mirror images. Alternatively, they can also be constructed by permutation of the placement of edges cf and dg onto the plane defined by the hexagon bcdefgb; in this construction, the edges are placed as straight lines without looping over other edges, and edge be is placed first and serves as a reference. This scheme is useful for extention to the discussion of five-disulfide cases. Also for extension to other multiple-disulfide proteins, the two topologies, **D** and **L**, are classified as two families, **D** and **L**, respectively, each of a single member here for the four-disulfide case.

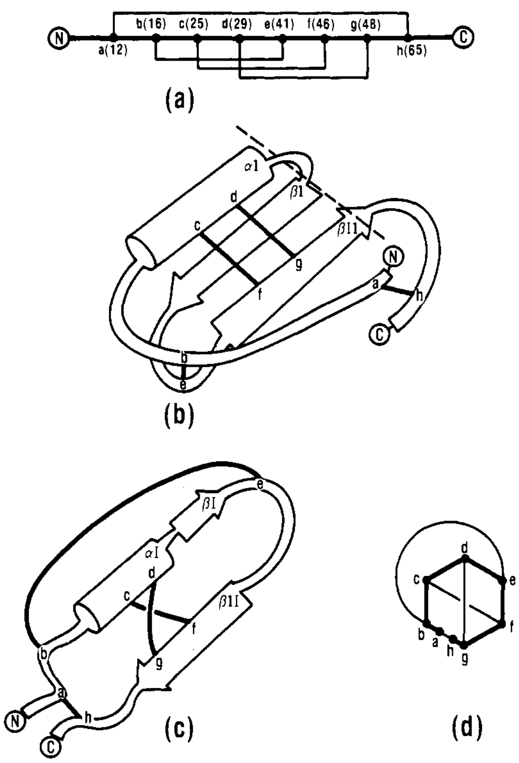

Figure 4. Scorpion neurotoxin. (a) Disulfide linkages in *Centruroides sculpturatus* Ewing variant 3 toxin. The numbers in parentheses are residue numbering of cysteines in the primary sequence. (b) Schematic drawing of the three-dimensional structure of variant 3 toxin. (c) Structure modified from b for alignment with the graph shown in d. The two β-sheets in b are rotated for 180° about the axis indicated by the dashed line, in the counterclockwise direction viewing from the lower right-hand corner. The second β-sheet is then moved longitudinally toward the helix. Vertex e is now in the upper right-hand corner. The ah-disulfide pair is also relocated, to the lower left corner. Distances between cysteinyl C_α atoms along the main peptide chain are altered such that their relative positions correspond to vertices in d. (d) Graph embedding corresponding to the structure of *C. sculpturatus* Ewing variant 3 neurotoxin. The topology is identical to the **D** topology in Figure 3b.

The two topologies shown in Figure 3a or Figure 3b for the nonplanar polypeptide chain are the only two members in the subset of the embeddings for the corresponding nonplanar *graph* that are topologically simple structures. An example of an embedding not in this subset is the more complex and topologically different structure shown in Figure 3c which has a trefoil knot. By the hypothesis discussed in the previous section and the fact that polypeptide chains are of finite dimension and the disulfide

[11] Richardson, J. S. *Adv. Protein Chem.* **1981**, *34*, 167–339.

[12] Walba[3a] had demonstrated that K(3,3) graph is equivalent to a three-rung Möbius ladder. Simon (*Topology* **1986**, *25*, 229–235) on the other hand proved that a Möbius ladder with three differentiated rungs is topologically chiral; i.e., there are topological stereoisomers for the structure. Since such a Möbius ladder cannot be superimposed with its image, changing the colors of the edges such that all edges are of different colors will maintain its chirality. Thus, the covalent structure graphs shown in Figure 3, each edge of which is unique as defined in the Introductory section, are also chiral. The two topologies can easily be verified to be mirror images, and therefore each is a distinctly different topology.

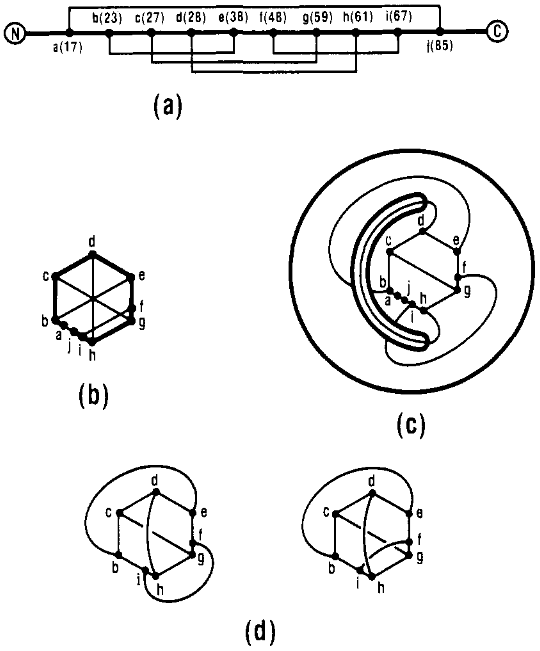

Figure 5. Porcine pancreatic colipase. (a) Disulfide linkages showing cysteine residue numbers in one of the two possible pairing schemes. In the other scheme which is planar, disulfide bonds to Cys-27 and Cys-28 are interchanged. (b) Covalent structure graph for the disulfide pattern in a. (c) Embedding of the graph in b on a surface of genus 1. (d) Two topologies in the **D** family for this polypeptide chain that are constructed from the procedure described in the Four-Disulfide Chains section. The topology on the left shows that the crossing number of the graph, i.e., the minimum number of edge crossings in a graph drawn in a plane, is 1, between edges cg and dh here. The topology on the right is ruled out for proteins since it contains linked loops c–d–h–g–c and b–e–f–i–b.

bonds are of finite length, topologies such as this are ruled out for folded proteins.

Two mammalian-directed scorpion neurotoxins share a similar primary, secondary, and tertiary structure.[13] The primary structure and the cysteine pairing scheme of one of the two neurotoxins, variant 3 toxin from the North American species *Centruroides sculpturatus* Ewing, is shown in Figure 4a. Figure 4b shows a schematic drawing of the tertiary structure of the molecule, and its topology (Figure 4d) is shown to be the **D** topology in Figure 3b.

Five-Disulfide Chains

The primary sequence of pancreatic colipase has been shown to contain ten cysteines engaging in five disulfides.[14] Two of the five disulfide pairs however cannot be determined unequivocally, and the disulfide pairing scheme is therefore unresolved; Figure 5a shows one of the two cysteine pairing schemes that is nonplanar:[15] the covalent structure graph (Figure 5b) contains K(3,3) subgraphs. The graph has a crossing number of 1 (Figure 5d) and, necessarily, must be embedded on a surface of genus 1 (Figure 5c). Although the genus of the graph in Figure 5b is 1, same as that in the nonplanar four-disulfide case, its topology has, by the procedure described in the previous section, four possibilities, two of which belong in the **D** family and are shown in Figure 5d. These two members[16] of the **D** family are the two permutations of the location for edge fi; they are topological diastereoisomeric[3,8] but not enantiomorphic since the mirror image of each belongs in the **L** family (in which edges cg and dh are interchanged from those in **D** family shown in Figure 5). On the basis of the hypothesis that protein structures do not have knotted structures, however, only one of the two topologies in Figure 5d, that on the left, is appropriate for the molecular topology of the polypeptide.

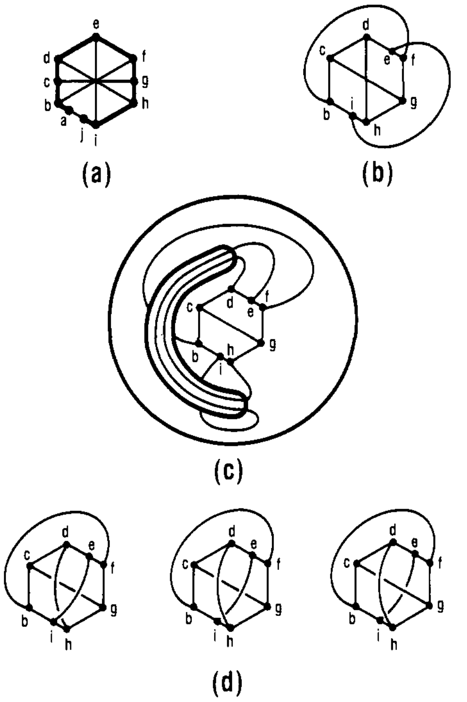

Figure 6. Second pairing scheme for a five-disulfide chain. (a) Covalent structure graph of the nonplanar five-disulfide polypeptide. Vertex i in Figure 5b is relocated to the edge following vertex b and is relabeled as vertex c here. (b) The graph has a crossing number of 2. Note that vertices are repositioned here such that **D** family and **L** family can be consistently and uniquely designated. (c) Embedding of the graph on a surface of genus 1. (d) Three topologies in the **D** family constructed by the procedure described in the Four-Disulfide Chains section. Note that these topological diastereomers are results of permuting the location of edge ei relative to edges cg and dh. Interchanging edges cg and dh of course results in topologies in the **L** family. Only the topology on the left is topologically simple and can describe polypeptides. The other two topologies contain linked loops: c–d–h–g–c and b–f–e–i–b for the topology in the center and d–e–i–h–d and b–c–g–f–b for the topology on the right.

Table I. Topological Properties of Polypeptides[a]

		nonplanar			
case no.	planar	L family	D family	genus	crossing no.
0–3SS	1				
4SS	1	1	1	1	1
5SS-1	1	1 (2)	1 (2)	1	1
5SS-2	1	1 (3)	1 (3)	1	2
5SS-3	1	1 (2)	1 (2)	1	1
5SS-4	1	1	1	1	1
5SS-5	1	1	1	1	1
5SS-6	1	1	1	1	1

[a] The first three columns are respectively the number of planar, **L** family (nonplanar), and **D** family (nonplanar) topologies. Numbers in parentheses are the numbers of topologies constructed by the procedure described in the Four-Disulfide Chains section. The last two columns are the genus and the crossing number, respectively, of the graph representing nonplanar chains in each case.

[13] (a) Almassy, R. J.; Fontecilla-Camps, J. C.; Suddath, F. L.; Bugg, C. E. *J. Mol. Biol.* **1983**, *170*, 497–527. (b) Fontecilla-Camps, J. C., Habersetzer-Rochat, C.; Rochat, H. *Proc. Natl. Acad. Sci., U.S.A.* **1988**, *85*, 7443–7447.

[14] Charles, M.; Erlanson, C.; Bianchetta, J.; Joffre, J.; Guidoni, A.; Rovery, M. *Biochim. Biophys. Acta* **1974**, *359*, 186–197.

[15] Difficulties in the sequencing experiment[14] led to the ambiguity of how Cys-59 and Cys-61 are paired to Cys-27 and Cys-28. In Figure 5a, the disulfide bond between Cys-27 and Cys-59 is equivalent to edge cf in Figure 2c, which is the invariant edge for a nonplanar four-disulfide graph. In the other possible pairing scheme, Cys-27 is disulfide-bonded to Cys-61 instead, and the graph becomes planar due to the absence of the necessary "cf edge".

[16] The left topology is a simple four-rung Möbius ladder, while the right topology is a four-rung Möbius ladder with the attachment sites of the last two rungs on one of the side pieces interchanged. These facts and several other crucial points on topological stereochemistry were kindly pointed out by the referees.

In a different nonplanar pairing scheme for a five-disulfide chain, vertex i is placed on edge bc in Figure 5b. Although this disulfide pairing pattern (Figure 6a) has an apparent crossing number of 2 (Figure 6b), it is a graph of genus 1 (Figure 6c). Moreover, by the procedure described in the Four-Disulfide Chains section, a total of six topologies can be constructed for the graph, three of which belonging in the **D** family are shown in Figure 6d. For the *molecular topology* however, only one from each family can describe polypeptide structures according to the hypothesis presented earlier in the report that excludes extrinsically complex topologies in proteins; this is also shown in Figure 6d.

In addition to the pairing scheme shown in Figure 5 (which is denoted as Case 5SS-1), and that in Figure 6 (Case 5SS-2), there are other nonplanar pairing schemes for a five-disulfide chain in which vertex i in Figure 5b or, equivalently, vertex c in Figure 6a is placed on other edges:

Case 5SS-3 where vertex c is located on edge de.
Case 5SS-4 where vertex c is located on edge ef.
Case 5SS-5 where vertex c is located on edge fh.
Case 5SS-6 where vertex c is located on edge hi.

The number of topologies for each of these cases is shown in Table I.

Molecular Topology and Protein Structure

Disulfides are important structural elements that stabilize three-dimensional structure of proteins, especially in smaller proteins.[11] In this report, molecular topology and topological stereoisomerism of multiple-disulfide polypeptide chain in stable folded proteins is characterized from graph theoretic analysis of its covalent structure.

It was shown that molecular topology is a unique property of a multiple-disulfide polypeptide chain and that this property is related to, but not implied by, the nonplanarity of the covalent structure graph of the polypeptide. For planar proteins, only one topology exists. Nonplanarity of a polypeptide chain on the other hand implies only that multiple topologies are allowed; by assuming that topologically complex structures are unlikely for folded proteins, the *number* of such allowed molecular topologies for some nonplanar chains are determined and the allowed topologies enumerated in this report. For example, a nonplanar four-disulfide chain such as variant 3 toxin from *C. sculpturatus* Ewing has two enantiomorphic topologies, one in the **D** family and one in the **L** family; for a nonplanar five-disulfide chain such as that of pancreatic colipase, there are also two possible molecular topologies, one in each of the two families. Only one of these allowed molecular topologies will represent the final, stable tertiary structure of the protein which the polypeptide chain is folded into during folding and disulfide oxidation processes. The X-ray crystallographic structures of *C. s.* Ewing variant 3 toxin and *A. australis* Hector toxin II show that they have the **D** topology. The tertiary structure of colipase and its molecular topology remain undetermined.

The classification of polypeptide molecular topologies appears to be related to the genus of its covalent structure graph. Nonplanar four-disulfide and five-disulfide cases (Table I) are shown to have covalent structure graph of genus 1 and have two possibilities for their molecular topology.

The subset of topologies without any knotted structure consists of a single topology for planar graphs; it has two families, **D** and **L**, for nonplanar graphs (Table I). The description and classification of molecular topology of polypeptide chains by this subset are based on the hypothesis that graphs representing the polypeptide chain of folded proteins do not have extrinsic topological complexity; due to extensive interactions among amino acid residues and the finite length of polypeptides, complex looping of polypeptide such as that found in linked loops or trefoil knots are not likely structures in folded proteins. Simulation of loop closures in polypeptide chains in the simplified C_α-to-C_α virtual bond representation showed that knotted structures could be observed only in a few cases and with a low probability;[2,4] if amino acid side chains and their interactions were taken into account in the simulation of loop closure however, the probability of self-avoiding loop closures would diminish for most, if not all, naturally occurring proteins. More importantly, linked loops and trefoil knots are not found in tertiary structure of proteins known to date (ref 5 and 6 and Figure 4c). It is therefore proposed, as a strong working hypothesis, that polypeptide chains of folded proteins do not have such topologically complex structures. If in fact such heretofore unknown structure were to be found, the analyses presented here represent a "lower limit" on the topological stereoisomerism in such polypeptide chains. Scorpion neurotoxins and pancreatic colipase provided examples for the discussion of the molecular topology discussed in this report. Variant 3 toxin from *C. sculpturatus* Ewing and toxin II from *A. australis* Hector however are the only examples of *known topology* for nonplanar proteins. Identification and structure determination of additional nonplanar proteins will provide more information on molecular topology of proteins. Further knowledge of protein biosynthesis and protein folding mechanism could also also provide support for the contention that knotted structures do not exist in folded protein molecules. Structural analyses of the polypeptide folding in scorpion neurotoxins on the other hand may reveal some structural basis for the existence of **D** topology vs **L** topology (in progress). This knowledge not only will be useful in the structure prediction of proteins for which a multiplicity of topologies are allowed but will also complement synthetic and theoretical work in topological stereochemistry.[3,8,9]

Finally, it should be pointed out that the analysis above has been carried out for single-chain disulfide-containing polypeptides only. It can be extended to multiple-chain proteins, synthetic polypeptide molecules, and polypeptide chains that are branched at C_α positions due to sulfide bridges.[17,18] In synthetic polypeptides where C_α atoms become fully substituted and cyclized on both *R* and *S* branches, the covalent structure graph will have vertices of order 4, and the analysis of molecular topology may require the consideration of the complete graph K_5, the other Kuratowski nonplanar graph.[7]

Acknowledgment. I thank K. C. Chou, M. A. Johnson, and G. M. Maggiora for discussions and critical reading of the manuscript and V. Nicholson (Kent State University) and P. G. Mezey (University of Saskatchewan) for discussions. Comments and suggestions by D. M. Walba (University of Colorado at Boulder) and J. Simon (University of Iowa) were important for this article and are gratefully acknowledged.

[17] Gross, E. In *Chemistry and Biology of Peptides*; Meienhofer, J., Ed.; Ann Arbor Science Publishers: Ann Arbor, MI, 1972; pp 671–678.

[18] Schnell, N.; Entian, K.-D.; Schneider, U.; Götz, F.; Zähner, H.; Kellner, R.; Jung, G. *Nature* **1988**, *333*, 276–278.