LOOP PENETRATION IN PROTEINS

A Not So Knotty Problem

Michael H. Klapper and Issac Z. Klapper, *Department of Chemistry, The Ohio State University, Columbus, Ohio 43210 U.S.A.*

Protein chain entanglement has generally been discussed in terms of knots (understood by most biochemists as a structure which, when grasped at the two free ends, becomes a true knot in the mathematical sense). We propose to replace this conceptual model with that of loop penetration: one segment of the protein encircled by another. Of these two structural concepts the latter is the more general, since all knots contain at least one penetrated loop, while all structures with penetrated loops are not necessarily knots.

Under certain conditions loop penetration can be detected in a protein, the three-dimensional structure of which is unknown. To begin, a protein is represented by a linear graph. As previously suggested by Crippen (1), the cysteine α-carbons are the vertices of the graph, and an edge is drawn between two vertices only if there is, between the two corresponding cysteine residues, a covalent pathway through either the protein backbone or a disulfide bond. This particular graph representation is an adjacency relation for cysteine residues in proteins. It is trivial when no disulfide bonds are present; thus, only proteins with disulfide bonds are worth considering.

The determination of loop penetration is based upon the conjecture that, if a protein's graph representation is nonplanar, then its three-dimensional structure contains a penetrated loop. A graph is planar when it can be drawn on a plane in such a way that no one edge crosses any other edge; otherwise, the graph is nonplanar. Determination of planarity is based on a restatement of the more general theorem due to Kuratowsky (discussed in most texts on graph theory; e.g., reference 2): the graph representation of a protein is planar if and only if it does not contain the nonplanar $K_{3,3}$ as a subgraph. ($K_{3,3}$ is a graph with six vertices partitioned into two sets of three each. A vertex is joined to each of the three not within its own set; hence all vertices are trivalent.) The conjecture that nonplanarity implies loop penetration is based upon visual inspection of a three-dimensional model for the $K_{3,3}$ graph. This model contains a penetrated loop in spite of all structural manipulations, excluding cutting.

When the disulfide pairing is known, a protein's graph representation is constructed easily. Planarity can be determined manually, but a mechanical procedure is easier when many graphs must be investigated; the computer algorithm we used is similar to that proposed by Mei and Gibbs (3). To reduce sample bias, proteins were chosen so that a superfamily (4) was represented only once, unless more than one protein included in different families had different pairing patterns. The resultant sample library contains 67 proteins, but only 26 have more than three disulfide bonds, a necessary condition for nonplanarity, since the $K_{3,3}$ graph contains six trivalent vertices. Of these 26, only 2 —colipase (5) and Androctonus neurotoxin II (6) —have nonplanar graph representations, and thus contain a penetrated loop. How many nonplanar graphs would be expected were all disulfide pairing patterns equally probable?

The total number of distinct pairing patterns for proteins with n cysteine residues is $\Pi^{n}{}_{i-1} (2i - 1)$, assuming n is even, and all cysteine residues are oxidized. Of the 105 different graphs associated with all possible pairings of four disulfide bonds, 4 are nonplanar; for five disulfide bonds, 130 are nonplanar out of a total of 945. The number of distinct graphs

TABLE I
STATISTICS OF NONPLANAR PROTEIN GRAPH REPRESENTATIONS

Disulfide bonds	Total number of distinct graphs	Fraction of nonplanar graphs*	Number of proteins‡	Nonplanar graphs (observed)	Nonplanar graphs (expected)§
1	1	0.	15	0	0
2	3	0.	10	0	0
3	15	0.	16	0	0
4	105	0.0381	11	1	0–1 (0.42)
5	945	0.1376	6	1	0–1 (0.83)
6	10,395	0.30 (0.010)	2	0	0–1 (0.60)
7	135,135	0.45 (0.012)	3	0	1–2 (1.35)
≥12	—	—	4	0	—

*In the case of 6 and 7 disulfide bonds, estimates were made by randomly generating pairing patterns and calculating the fraction of these that were nonplanar. The standard deviation (in parentheses) of the estimate was calculated as $[f(1 - f)/N]^{1/2}$, where N is the total number of structures generated, and f is the fraction found to be nonplanar.

‡Protein sequences were obtained in large part from reference 4.

§The number of nonplanar graphs expected on the basis of the number of proteins in the sample was calculated as the product of entries in columns 3 and 4 and is given in parentheses.

possible for proteins with more than five disulfide bonds are sufficiently large, so that each one was not checked individually for nonplanarity. Instead, a Monte Carlo procedure was used to estimate the fraction of nonplanar graphs. As might have been expected, the probability of finding a nonplanar graph increases with the number of disulfide bonds (Table I).

Were disulfide pairing patterns equiprobable, and were the protein library truly random, then we would have expected to find approximately three proteins with seven or fewer disulfide bonds, and an associated nonplanar graph, which compares well with the number found. However, because of the small sample size, a firm conclusion cannot be drawn. The sample library contains four proteins with more than eleven disulfide bonds, each of which yields a planar graph. These four have multidomain structures, with each domain being a"small protein"covalently linked to its neighbors. Therefore, these four were not included in the statistical argument.

We conclude that loop penetration (as detected by nonplanar graph representations), while infrequent, may not be rare. This contradicts the apparent consensus that protein chain entanglement does not occur. We should also emphasize that a count of nonplanar structures yields only a minimal estimate of loop penetration. Proteins with no disulfide bonds (e.g., carbonic anhydrase and subtilisin), or with planar graph representations may nonetheless contain a penetrated loop, since nonplanarity is a sufficient, but not necessary, condition. We suggest that loop penetration may be fairly common, although its importance as a structural feature in proteins remains to be established.

Received for publication 14 November 1979.

REFERENCES

1. Crippen, G. M. 1974. Topology of globular proteins. *J. Theor. Biol.* **45**:327–338.
2. Marshal, C. W. 1971. Applied Graph Theory. John Wiley & Sons, Inc., New York.
3. Mei, P.–S., and N. E. Gibbs. 1970. A planarity algorithm based on the Kuratowsky theorem. *AFIPS Conf. Proc.* **36**:91–93.

4. Dayhoff, M. O. 1978. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, D.C. 5:9–22 (suppl. 3).
5. Kopeyan, C., G. Martinez, S. Lissitzky, F. Miranda, and H. Rochat. 1974. Disulfide bonds of toxin II of the scorpion Androctonus Australis *Eur. J. Biochem.* 47:483–489.
6. Erlanson, C., J. A. Barrowman, and B. Borgstrom. 1977. Chemical modifications of pancreatic colipase. *Biochim. Biophys. Acta.* 489:150–162.

STRUCTURE OF HAPTOGLOBIN HEAVY CHAIN AND OTHER SERINE PROTEASE HOMOLOGS BY COMPARATIVE MODEL BUILDING

Jonathan Greer, *Department of Biological Sciences, Columbia University, New York, New York 10027 U.S.A.*

Proteins often occur in families whose structure is closely similar, even though the proteins may come from widely different sources and have quite distinct functions. It would be useful to be able to construct the three-dimensional structure of these proteins from the known structure of one or more of them without having to solve the structure of each protein *ab initio.* We have been using comparative model building to derive the structure of an unusual protein of the trypsin-like serine protease family (1). We have recently extended this comparison to include other serine protease homologs for which a primary structure is available.

To generate structures for the different members of the serine protease family, it is necessary to extract the common structural features of the molecule. Fortunately, three independently determined protein structures are available: chymotrypsin (2), trypsin (3, 4), and elastase (5). These three structures were compared in detail and the structurally conserved regions in all three, mainly the β-sheet and the α-helix, were identified. The variable portions occur in the loops on the surface of the molecule. By using these structures, the primary sequences of these three proteins were aligned. From this alignment, it is clear that sequence homology between the proteins occurs mainly in the structurally conserved regions of the molecule, while the variable portions show very little sequence homology.

The protein that has been built is haptoglobin (Hp), a serum protein that forms a highly specific and exceedingly strong complex with the blood protein hemoglobin. The in vivo function of Hp is to permit the recycling of red blood cell free hemoglobin iron and to prevent loss of heme iron in the urine and related damage to the kidney tubules eventually causing renal failure (6). Kurosky et al. (7, 8) have shown that the sequence of the heavy chain of Hp (HpH) is clearly homologous to the mammalian serine proteases, although the protein exhibits no protease activity.

The first step in modeling HpH into the known serine protease structure is to align the sequence to those of the known structures so that homology is maximized in the structurally conserved regions. Strong sequence homology was found for every structurally conserved region. No additions or deletions were found in these regions; all such occurred in the external loops where deviations are also found between the three known proteins. The resulting alignment shows that HpH must be very closely homologous to the proteases in structure as well as in sequence.

Coordinates were generated for HpH using the known homologus structures. Side chains