Loops in Globular Proteins: A Novel Category of Secondary Structure

JACQUELYN F. LESZCZYNSKI AND GEORGE D. ROSE*

The protein loop, a novel category of nonregular secondary structure, is a segment of contiguous polypeptide chain that traces a "loop-shaped" path in three-dimensional space; the main chain of an idealized loop resembles a Greek omega (Ω). A systematic study was made of 67 proteins of known structure revealing 270 omega loops. Although such loops are typically regarded as "random coil," they are, in fact, highly compact substructures and may also be independent folding units. Loops are almost invariably situated at the protein surface where they are poised to assume important roles in molecular function and biological recognition. They are often observed to be modules of evolutionary exchange and are also natural candidates for bioengineering studies.

THE SECONDARY STRUCTURE OF PROTEINS FALLS INTO three classes: α-helices, β-sheet, and reverse turns (*1–4*). Helices and sheet are termed "regular" structures because their residues have repeating main-chain torsion angles, and their backbone N-H and C=O groups are arranged in a periodic pattern of hydrogen bonding (*1*). In contrast, turns are "nonregular" structures with nonrepeating backbone torsion angles and, at most, one internal N-H . . . O=C hydrogen bond (*1–4*). Remaining residues, by subtraction, are often classified as "random coil," although, as Richardson has pointed out, they are neither random nor coil (*1*).

In this article we examine another category of nonregular secondary structure—the loop. A loop may be described as a continuous chain segment that adopts a "loop-shaped" conformation in three-dimensional space, with a small distance between its segment termini. The main-chain trace of an idealized loop resembles a Greek omega (Ω). Backbone torsion angles for such a structure are nonrepeating, and there are few, if any, backbone hydrogen bonds. A simple loop subsumes no proper subsets that are also loops, while a compound loop contains at least one smaller embedded loop. Only simple loops are considered in the following discussion.

Loops have been discussed in relation to specific structures, such as the conspicuous loops in superoxide dismutase (*5*) and in immunoglobulin domains (*6*), the autolysis loop in serine protease zymogens (*7*), and the calcium-binding loops in parvalbumin (*8*). However, there has been no systematic study of these structures. Kuntz alluded to larger loops in his definitive paper on peptide chain turns (*2*), and the topic is mentioned briefly in a recent review (*3*). Looped-out regions are also evident in schematic representations of protein structure such as those of Richardson (*1*) or Lesk and Hardman (*9*). While such examples are clearly "looplike," their description has been only qualitative.

In our study, loops are defined explicitly. Stringent defining criteria are chosen deliberately to exclude those structurally ambiguous examples containing substantial amounts of regular secondary structure. The definition was implemented in the form of a computer algorithm and used to identify all loops in 67 proteins of known structure. The set of identified loops was then characterized with respect to residue composition, size and shape, compactness, accessibility to solvent, and role in protein taxonomy.

Our survey reveals an abundant population of loops, on the order of four per protein molecule. Almost always, they are situated at the molecular surface; often, they are implicated in molecular function. Most of these loops are highly compact, globular structures, with low x-ray temperature factors and a packing efficiency that rivals that of β-sheet. The observed compactness is a consequence of loop side-chain atoms that pack tightly within the loop core. In view of such characteristics, the description of these chain segments as "random coil" warrants revision.

Loops are choice candidates for protein bioengineering studies. The catalog of loops presented here should be useful for the design of such experiments, as well as in the further study of nonregular protein secondary structure.

Identification of loops from x-ray coordinates. A loop is a continuous segment of polypeptide chain that is defined in terms of its (i) segment length, (ii) absence of regular secondary structure, and (iii) distance between segment termini. These criteria are now specified in detail.

The segment length must be between 6 and 16 residues. The lower length limit serves to eliminate reverse turns. Superficially, it might seem that a turn is merely a small loop, but an important characteristic distinguishes the two. Turns, which range from three to five residues in length, have backbone groups that pack together closely, forcing side chains to project outward (*3*). This stereochemical restriction is relaxed in larger segments where side-chain atoms can pack within the loop's own core. The upper length limit imposes a practical threshold that eliminates most of the compound loops.

A loop may contain no regular secondary structure. This criterion excludes adjacent strands of antiparallel β-sheet as well as structurally ambiguous cases. Secondary structure assignments for the residues were taken from the Kabsch and Sander (K&S) dictionary of protein secondary structure (*10*). However, two minor exceptions to the K&S classification were adopted: two-residue strands and single turns of helix (four or five residues) are not counted as regular secondary structure. Although they are ignored in most classification schemes, K&S includes β-strands that are just two residues in length. Strand lengths are distributed in a statistically well-behaved fashion, with the exception of these two-residue strands. Four- and

The authors are members of the Department of Biological Chemistry, Milton S. Hershey Medical Center, Pennsylvania State University, Hershey, PA 17033.

*To whom correspondence should be addressed.

Table 1. Summary of 270 omega loops in 67 x-ray elucidated proteins.*

PROT	FIRST	NUM	SEQUENCE
1ABP	93	7	V N...K P
1ABP	142	7	A N...T A
1ABP	203	6	G M...S T
1ABP	236	13	A V...G F
1ABP	289	6	I T...N F
1ABP	299	6	E K...L G
2ACT	8	6	R S...A V
2ACT	58	7	R T...R G
2ACT	89	15	Y P...L Q
2ACT	139	6	A A...A F
2ACT	141	16	G D...P C
2ACT	182	11	N S...E G
2ACT	198	8	R N...G T
2ACT	203	7	A G...I A
4ADH	14	8	L W...P F
4ADH	100	13	C R...C L
4ADH	115	8	D L...G T
4ADH	122	7	T M...T S
4ADH	282	6	C Q...Y G
2ADK	133	10	G E...D N
1ALP	217	8	N V...N N
2APP	41	15	F S...S V
2APP	129	8	N T...S Q
2APP	139	11	F D...Q P
2APP	184	9	V D...W S
1APR	8	10	T D...Y Y
1APR	18	14	G Q...N L
1APR	43	16	G S...D K
1APR	61	9	P S...K A
1APR	76	8	I G...S A
1APR	90	14	D T...G P
1APR	129	10	D T...S S
1APR	189	9	I D...W A
1APR	203	9	A T...L G
1APR	216	11	A I...L I
1APR	227	6	L P...A A
1APR	233	16	V G...L G
1APR	243	8	Q D...G F
1APR	261	13	S I...E I
1APR	280	8	A E...C T
1APR	291	9	G A...A I
1AZU	9	7	G N...Q F
1AZU	35	12	H P...G H
1AZU	67	6	G L...D Y
1AZU	73	11	L K...A H
1AZU	84	9	T K...E K
1AZU	112	7	C T...H S
2B5C	32	16	L T...L R
1BP2	23	8	N N...C G
†1BP2	25	15	Y G...V D
1BP2	56	11	K K...V D
2BP2	23	8	N N...C G
†2BP2	25	15	Y G...V D
2BP2	61	8	C K...N P
156B	16	10	V I...K A
156B	47	12	T P...P M
351C	16	11	H A...P A
351C	51	12	G S...M P
155C	21	8	I Q...T D
155C	47	8	A S...K Y
155C	83	13	K P...G A
155C	128	6	J J...J J
2C2C	18	16	H T...L F
2C2C	30	14	P N...H K
2C2C	41	16	A H...M K
2C2C	74	16	P K...K S
2CAB	6	7	G Y...N G
2CAB	17	8	S K...A N
2CAB	78	10	V L...D S
2CAB	98	7	G S...H G
2CAB	108	7	T V...K Y
2CAB	128	13	Y S...D G
2CAB	197	8	S L...L Y
2CAB	230	11	L S...V P
1CAC	5	12	W G...H W
1CAC	17	7	H K...I A
1CAC	98	6	G S...G E
1CAC	108	7	T V...K Y
1CAC	128	13	Y G...D G
1CAC	166	7	S I...G K
1CAC	197	8	S L...L L
1CAC	232	8	D G...E E
2CHA	70	9	E F...S E
2CHA	94	9	Y N...N D
2CHA	114	6	F S...V S
2CHA	217	8	S S...S T
3CNA	13	9	P N...P S
3CNA	97	8	T G...T N
3CNA	116	8	K S...Q T
3CNA	147	9	T T...L E
3CNA	160	6	S S...S P
3CNA	199	11	I K...D G
†3CNA	222	14	P S...P D
3CNA	229	9	L L...A N
3CPA	128	14	K T...G V
3CPA	142	15	D A...G A
3CPA	156	11	A S...Y H
3CPA	205	9	P Y...S I
3CPA	231	7	K S...T S
3CPA	244	7	I T...Q A
3CPA	272	14	R D...S Q
1CPV	18	6	C K...D S
1CPV	64	14	K L...A L
1CRN	33	12	I I...D Y
1CTX	1	15	I R...C P
1CTX	26	10	C D...G K
1CYC	18	15	H T...N L
1CYC	30	14	P N...Q A
1CYC	40	15	T G...K S
1CYC	70	15	N P...A G
3CYT	18	15	H T...N L
3CYT	34	10	G L...Q A
3CYT	40	15	T G...K S
3CYT	70	15	N P...A G
1ECD	33	10	S I...F A
1ECD	41	9	F A...S I
1EST	69	12	G E...T E
1EST	94	11	W N...Y D
1EST	112	7	V T...Y V
1EST	142	10	G L...L A
1EST	165	14	Y A...T V
1EST	216	11	V S...R K
3FAB	24L	6	G S...N I
3FAB	122L	11	P S...K A
3FAB	168L	6	K Q...N K
3FAB	182L	6	L T...Q W
3FAB	72H	6	N T...N Q
3FAB	99H	7	L I...I D
3FAB	132H	9	S K...T A
1FDX	12	12	G A...I I
1FDX	30	12	I D...S C
1FDX	39	12	G S...A P
3FXN	54	8	S A...V L
2GCH	70	9	E F...S E
2GCH	94	8	Y N...N N
2GCH	112	7	A S...T V
2GCH	165	12	N T...K I
2GCH	217	8	S S...S T
1GPD	47	6	D S...G V
1GPD	76	7	E M...N I
1GPD	121	9	P S...F V
1GPD	128	10	F V...K Y
1GPD	183	16	K T...R G
2GRS	83	7	A V...P S
2GRS	139	9	T I...K Y
2GRS	162	11	T P...A S
2GRS	239	7	E N...E V
2GRS	256	6	K T...G L
2GRS	268	7	A V...L P
2GRS	300	8	L N...Q T
2GRS	315	6	V D...Q N
2GRS	331	7	D V...A L
2GRS	404	12	T P...K T
2GRS	465	8	A I...S E
1HIP	20	7	N Q...K S
1HIP	28	14	R V...E Q
1HIP	43	7	C A...F M
†1HIP	44	16	A D...D E
4LDH	173	16	R Y...G V
4LDH	192	9	I G...V P
4LDH	203	16	W S...L G
†4LDH	212	14	L H...D W
4LDH	219	8	N K...W K
4LDH	239	8	V I...Y T
4LDH	275	11	V K...N V
1LDX	70	9	S L...K I
1LDX	79	8	V G...S L
1LDX	102	7	Q Q...S R
1LDX	193	8	G R...G V
1LDX	207	7	N N...L Q
1LDX	211	6	N L...G M
1LDX	218	7	W E...E G
1LDX	236	1	A Y...Y E
1LDX	276	8	K E...K E
†1LH1	41	13	K D...E V
1LH1	47	8	L K...V P
1LHB	46	14	P A...L T
1LHB	55	10	F K...E L
7LYZ	18	8	D N...S L
7LYZ	36	7	S N...Q A
7LYZ	44	9	N R...T D
7LYZ	60	16	S R...N L
1LZM	134	6	A K...W Y
1MBN	40	8	L E...F K
1MBS	37	14	P E...L K
1MBS	49	6	L K...D D
1MBS	78	7	K K...E A
2MHB	40A	9	K T...D L
2MHB	39B	16	Q R...A V
2MHB	47B	11	D L...G N
1NXB	6	8	Q H...Q T
2PAB	49A	6	T S...G E
8PAP	8	6	R Q...A V
8PAP	60	8	S Y...G Y
8PAP	86	15	Y P...E K
8PAP	138	16	G K...P C
8PAP	175	11	N S...N G
8PAP	191	8	R G...Y G
8PAP	198	6	G V...L Y
1PCY	6	8	G A...L A
1PCY	41	16	F D...I S
1PCY	63	6	L N...G E
1PCY	84	9	C S...G M
3PGM	11	15	S E...D V
3PGM	98	12	A Q...K F
3PGM	109	12	F N...P P
3PGM	123	8	I D...F S
3PGM	132	14	K G...V L
3PGM	209	16	L V...S Y
2PTN	69	12	G E...N E
2PTN	94	9	Y N...N D
2PTN	112	7	A S...R V
2PTN	142	11	G N...Y P
2PTN	184A	8	G Y...K D
2PTN	217	8	S G...N K
1REI	91	6	Y Q...P Y
1RHD	34	10	S W...E A
1RHD	43	15	A R...S F
1RHD	60	14	I E...V M
1RHD	85	6	G S...I S
1RHD	99	7	N G...G S
1RHD	185	7	G R...T Q
1RHD	193	7	E P...G L
1RHD	216	8	L T...E K
1RHD	284	10	P E...K G
1RNS	36	6	T K...C K
1RNS	87	10	T G...C A
2RXN	18	11	G X...G T
2RXN	38	8	V C...V G
1SBT	17	6	H S...Y T
1SBT	37	8	S S...K V
1SBT	74	13	A L...A P
1SBT	96	6	L G...G S
1SBT	157	8	G S...S T
1SBT	181	7	D S...R A
1SBT	257	10	L G...K G
2SGA	16	16	I A...S L
2SGA	93	7	S F...D Y
2SGA	218	7	G N...G G
3SGB	16	16	I S...S L
3SGB	48	8	V R...Y Y
3SGB	66	9	W A...T V
3SGB	93	7	S F...D Y
3SGB	118	7	T V...D I
3SGB	167	14	A T...G M
3SGB	190	12	V C...P L
3SGB	199	9	L Y...I G
3SGB	235	6	L V...G V
2SNS	43	10	E T...V E
2SNS	114	6	V Y...N N
2SNS	136	6	K L...W S
2SOD	50	9	D N...S A
2SOD	67	12	K K...R H
2SOD	103	7	S L...Y S
2SOD	122	16	D D...G N
†2SOD	132	6	S T...G N
2SSI	19	7	G V...T A
1TIM	67A	13	Y K...I S
1TIM	169A	6	A I...G K
3TLN	24	8	Y S...L Q
3TLN	32	7	D N...D G
3TLN	44	10	A K...G S
3TLN	55	16	W A...P A
3TLN	91	7	L S...N N
3TLN	125	6	G D...T F
3TLN	188	16	I G...L R
3TLN	204	10	S M...G D
3TLN	214	6	P D...S K
3TLN	221	13	Y T...I N
3TLN	248	8	G T...S V

five-residue helices are classified as type III reverse turns. Again, the distribution of helical segment lengths is statistically well behaved, except for these single-turn helices.

The distance between segment termini, that is, the end-to-end distance, is measured as the distance from the first α-carbon to the last α-carbon in the segment. The end-to-end distance must be less than 10 angstroms and may not exceed two-thirds the maximum distance between any two α-carbons within the segment under consideration. This criterion selects as loops those segments with termini that "neck in" like an omega (Ω). The set of discovered loops is not overly sensitive to the coefficient of two-thirds. In practice, the end-to-end distance varies between 3.7 and 10.0 Å.

Loops identified as described above are frequently members of small families. Such families arise whenever a range of segments, all of similar length, satisfies the definition. For example, if residues i through j comprise a loop, then it often happens that residues i through $j + 1$ also comprise a loop. In our study, each family is represented by its most compact member. To choose these individual representatives, we evaluated the compactness of each segment in every family, and the most compact loops were selected.

The coefficient of compactness of Zehfus and Rose (*11*) was used to assess compactness. This coefficient, Z, is a sensitive single-value figure of merit that identifies those segments with the smallest solvent-accessible surface area for their volume. Explicitly,

$$Z = \frac{\text{accessible surface area of segment}}{\text{accessible surface area of sphere of equal volume}} \quad (1)$$

Solvent-accessible surface areas and volumes were calculated by the methods of Lee and Richards (*12*) and Pavlov and Federov (*13*), respectively.

Z is a dimensionless ratio and should show no dependence on unit size. However, the configurational freedom of very small segments is restricted in comparison to larger ones, and is thus biased toward more compact arrangements. To adjust for this apparent size dependency, a compensating correction term was applied to all Z values used in our study. This term,

$$0.488 \times e^{(-.068 \times N_r)} + 0.970$$

(where N_r = number of residues in the segment), is chosen to yield a standard normalized value when multiplied by Z values of the most compact units of all sizes (*11*).

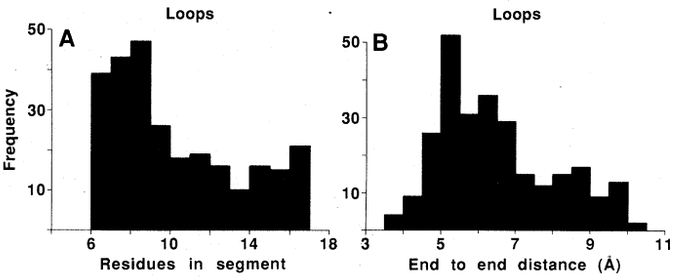

Fig. 1. Histograms showing the distribution of loop sizes for the 270 loops in Table 1 by number of residues (A) and by end-to-end distance (B). The mean values are: (A) 9.8 and (B) 6.4.

To identify loops, all continuous segments in 67 proteins from the Brookhaven database (*14*) were screened. (The Brookhaven database, a U.S. government supported resource, maintains atomic coordinates of x-ray elucidated proteins.) Only proteins from the K&S dictionary (*10*) were used in order to ensure self-consistency in the secondary structure assignments. Those continuous segments 6 to 16 residues in length that satisfied the end-to-end distance criterion were retained, if devoid of regular secondary structure. Coefficients of compactness were then calculated for all survivors and the most compact segment was chosen to represent each family cluster.

Characterization of loops. Compact loops are common structures in proteins. In the 67 proteins included in our survey, 270 loops were found, an average of more than four per molecule. The distribution of loop sizes is shown in Fig. 1 and the full set of loops is listed in Table 1.

Of those examined, only six proteins are without loops entirely: glucagon, insulin, mellitin, ovomucoid, avian pancreatic polypeptide, and pancreatic trypsin inhibitor. These proteins are all less than 60 amino acid residues in length and are among the nine smallest proteins in the database. Three, glucagon, mellitin, and pancreatic polypeptide, are also nonglobular, as determined by their axial ratios (as is discussed below).

Structures perceived intuitively as loops may not satisfy our stringent definition; the calcium-binding loops of parvalbumin (*8*) are an example. When the defining criteria are relaxed slightly to avoid elimination of loops containing three-residue β-strands, 22 additional loops are identified (Table 2), including several of these common examples.

Figures 2 and 3 illustrate typical loops from cytochrome c (*15*) and thermolysin (*16*). In each case, the loop main chain surrounds an internal cavity that is packed with side-chain atoms from loop residues. This kind of arrangement results in a highly compact chain fold for the segment. Occasionally, a metal ligand is also included, as shown in the thermolysin loop. In proteins with multiple loops, the individual loops occur frequently in spatial clusters, as they do in superoxide dismutase (Fig. 4).

Some loops with irregular tails satisfy our defining criteria both with and without their tail segment; in these cases the most compact representative is chosen, as previously stated. Occasionally, however, the larger version of such a loop, although more compact, is excluded because it exceeds the 16-residue upper limit. For example, the superoxide dismutase loop 67–78 is part of a larger, more compact loop 61–80. Despite this restriction, the upper size limit rarely eliminates a loop entirely; when the threshold is extended to 30 residues, only one de novo loop is found (phosphoglycerate mutase 191–211). As was mentioned above, the 16-residue cutoff serves to eliminate compound loops. If we use this upper bound,

*Key. PROT, Brookhaven name of protein (*14*). FIRST, residue number of loop NH_2-terminus; the Brookhaven numbering system, which need not correspond to the numbering in the Kabsch and Sander dictionary (*10*), was used. NUM, number of residues in loop. SEQUENCE, first, second, penultimate, and ultimate residues in loop. Single letter abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu: F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr; X, Asx; J, unknown. †Indicates a compound loop. Proteins used (and their parenthesized Brookhaven file names) are: l-arabinose-binding protein (1ABP), actinidin (2ACT), alcohol dehydrogenase (4ADH), adenylate kinase (2ADK), alphalytic protease (1ALP), penicillopepsin (2APP), rhizopuspepsin (1APR), azurin (1AZU), cytochrome b5 (2B5C), phospholipase A2 (1BP2), prophospholipase A2 (2BP2), cytochrome b562 (156B), cytochrome c551 (351C), cytochrome c550 (155C), cytochrome c2 (2C2C), carbonic anhydrase B (2CAB), carbonic anhydrase C (1CAC), alpha chymotrypsin (2CHA), concanavalin A (3CNA), carboxypeptidase (3CPA), calcium-binding parvalbumin (1CPV), crambin (1CRN), alpha cobratoxin (1CTX), ferrocytochrome c (1CYC), cytochrome c (3CYT), erythrocruorin (1ECD), tosylelastase (1EST), lambda immunoglobulin Fab NEW (3FAB), *Peptococcus* ferredoxin (1FDX), flavodoxin (3FXN), gamma chymotrypsin (2GCH), glucagon (1GCN), glyceraldehyde-3-phosphate dehydrogenase (1GPD), glutathione reductase (2GRS), high potential iron protein (1HIP), insulin (1INS), apolactate dehydrogenase (4LDH), lactate dehydrogenase isoenzyme (1LDX), acetate-met-leghemoglobin (1LH1), lamprey methemoglobin cyanide V (1LHB), egg lysozyme (7LYZ), T4 lysozyme (1LZM), whale metmyoglobin (1MBN), seal metmyoglobin (1MBS), horse aquomethemoglobin (2MHB), mellitin (1MLT), neurotoxin B (1NXB), ovomucoid third domain (1OVO), prealbumin (2PAB), papain (8PAP), plastocyanin (1PCY), phosphoglycerate mutase (3PGM), avian pancreatic polypeptide (1PPT), trypsin (2PTN), pancreatic trypsin inhibitor (4PTI), Bence-Jones immunoglobulin REI (1REI), rhodanese (1RHD), ribonuclease S (1RNS), rubredoxin (2RXN), subtilisin BPN (1SBT), *Streptomyces* proteinase A (2SGA), *Streptomyces* proteinase B (3SGB), staphylococcal nuclease (2SNS), Cu, Zn superoxide dismutase (2SOD), *Streptomyces* subtilisin inhibitor (2SSI), triose phosphate isomerase (1TIM), and thermolysin (3TLN).

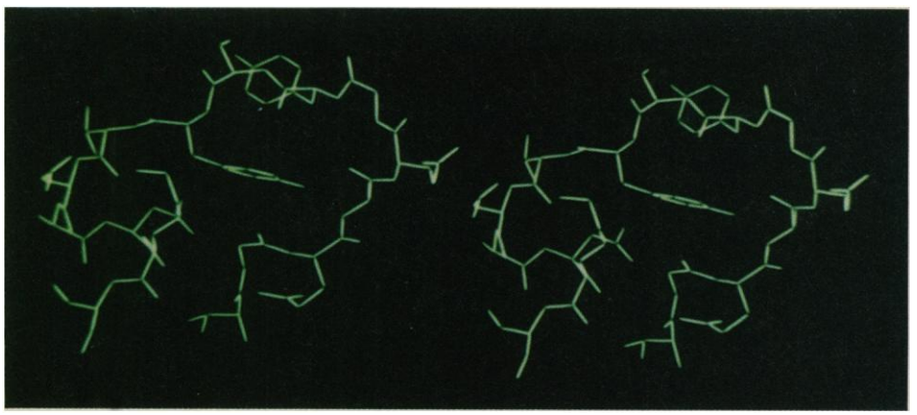

Fig. 2. Stereoview of a typical loop from cytochrome c (residues 40–54).

only seven compound loops fail to be excluded; these loops are indicated by a dagger in Table 1.

Loops may contain one or more reverse turns; these facilitate the main-chain direction changes needed to bring segment termini together. If loop curvature is sufficiently gradual, the chain direction can be reversed without resorting to an explicit turn, but this is unusual. All of the loops contain at least one turn or bend residue, as defined by Kabsch and Sander (*10*), but not every loop contains a complete turn or bend. Because they include reverse turns, loops do not constitute a pure structural category. Nonetheless, a loop and a turn are distinct moieties. While both result in changes in the overall direction of the polypeptide chain, a loop cannot be viewed merely as an "overgrown" turn. A turn, unlike a loop, has backbone groups that pack together closely, forcing side chains to project outward (*3*). For steric reasons, a segment of main chain cannot circumscribe an interior cavity of polyatomic dimensions until it exceeds a length of five residues.

The residue composition of loops was assessed by calculating the normalized frequency of occurrence, f, for each residue type, X, such that

$$f = \frac{X_L X_T}{N_L N_T} \quad (2)$$

where X_L is the number of residues of type X in loops, X_T is the total number of residues of type X, N_L is the total number of residues in loops, and N_T is the total number of residues in the database. A value of $f = 1$ implies that X is distributed randomly in loops. Values greater than unity imply that X is found preferentially in loops; conversely, f values less than unity imply a less than average frequency of occurrence of X in loops.

Examination of f values for all residues in loops reveals that residues present most often in reverse turns (*1–4, 17*) are also found most often in loops (Gly, Pro, Asp, Asn, and Ser) with the notable addition of Tyr (Table 3). All but Tyr have short side chains, and all but Pro are polar; Pro favors turns for steric reasons (*3*). Hydrophobic residues are strongly disfavored in loops; these include Val, Met, Ile, Leu, and Ala with aliphatic side chains, and His, Trp, and Phe with aromatic side chains.

To quantify the accessibility of a loop, we calculated the solvent-accessible surface area (*12*) in each of three successive states: in the standard state (*18*), as an isolated secondary structure, and within the protein (Fig. 5). Approximately half of the area of regular secondary structure is lost upon the formation of the isolated secondary structure, and the remaining half is lost when that secondary structure is buried within the protein (*19*). In our sample of 67 proteins, the percentage of the area lost when the chain folds into an isolated loop (34 percent) is comparable to the area loss upon formation of an isolated helix (35 percent). However, the subsequent loss when the loop is incorporated into the protein (47 percent) is less than that of the helix (60 percent). These statistics indicate that loops tend to be somewhat more accessible to solvent than helices.

These fractional accessibilities reveal that loops are almost invariably situated at the molecular surface (Figs. 4 and 5). It should be noted that the definition of a loop does not require that it be at the surface. Moreover, the data on solvent accessibility are consistent with the residue composition; loops are found at the protein surface and contain a preponderance of hydrophilic residues.

Loops are as compact as the proteins that contain them (Fig. 6). The coefficient of compactness, Z, is used to assess compactness, as described above. The Z values of loop segments range between 1.43 and 1.86 with a mean (± standard deviation) of 1.61 ± 0.07. In comparison, Z values for the 67 proteins range between 1.36 and 1.93 (with a single outlier at 2.09); the mean of this distribution is 1.67 ± 0.13.

It is conceivable that the apparent compactness of loops is biased by the use of the most compact segment to represent a loop family. As a control, the largest member of each family was chosen instead and used as the representative member; the coefficient of compactness was then calculated for these largest representatives. The distribution for these largest representatives is similar to that for compact representatives, ranging from 1.48 to 1.86 with a mean (± standard deviation) of 1.64 ± 0.07. This control demonstrates that loops are inherently compact structures.

As would be expected from their observed compactness, loops are not flat, but globular. This visual impression is confirmed when we calculate the principal moments of inertia for all loops, helices,

Table 2. Additional loops found in 67 x-ray elucidated proteins with defining criteria relaxed to allow three-residue strands of β sheet. The key is the same as that for Table 1.

PROT	FIRST	NUM	SEQUENCE	PROT	FIRST	NUM	SEQUENCE
4ADH	130	8	F T...P I	1CPV	51	12	D Q...D E
4ADH	158	16	A K...I G	1CPV	89	9	G D...K I
1ALP	190	12	A C...S W	1GPD	279	15	V S...F D
1ALP	200	8	I T...A Q	2GRS	370	7	V V...P P
2APP	212	11	G I...L L	1HIP	65	15	L F...A S
2APP	290	8	N S...L I	4LDH	289	14	L P...I V
1APR	31	12	L N...W V	3PGM	166	11	I A...M I
1APR	161	7	A A...S D	2RXN	5	8	T C...Y I
1APR	169	11	D F...N K	2SGA	119	7	Y L...S Y
1APR	310	10	V V...I R	2SGB	138	9	R R...T H
155C	36	14	P N...S E	1OVO	23	9	V C...T Y

strands of sheet, and protein monomers within the set of 67 proteins. Ratios of the largest to the smallest eigenvalues were formed. The axial ratios for loops resemble those for whole proteins (Fig. 7), while ratios for helices and strands are more rodlike.

The free energy change upon closing a protein segment into a loop consists of an unfavorable entropic contribution and a compensating enthalpic contribution. The enthalpy needed to counterbalance loop-closing entropy may be due to either intrasegment or extrasegment interactions, or both. At one extreme, a loop might be stabilized by interactions within its own core. At the other extreme, the rest of the protein might provide a stable framework that pinches together the termini of an intervening segment, forcing that segment to "loop out." Greater structural autonomy would be expected in the former case.

The number of noncovalent contacts between loops can be used to provide a rough estimate of loop enthalpy. Using united-atom radii (*12*), we plotted the number of noncovalent contacts as a function of the number of atoms in the loop, and obtained a line described by the equation (± standard error):

$$\text{Number of contacts} = 3.0\ (\pm 0.03) \times \text{number of atoms} - 11.3\ (\pm 2.7) \quad (3)$$

The equation can be interpreted to mean that loop enthalpy is essentially a linear function of loop length. (The negative intercept is expected upon extrapolation to zero length because a threshold of several atoms would be required to establish any contacts.) If we assume a binding energy of −0.03 kcal/mol per contact, the average loop enthalpy is on the order of −0.6 kcal/mol per residue.

Table 3. Residue frequencies in loops, normalized with the use of equation 2.

Gly	1.35	Glu	1.09	Trp	0.85
Pro	1.28	Thr	1.07	His	0.83
Tyr	1.28	Lys	1.02	Ala	0.77
Asp	1.22	½ Cys	0.94	Leu	0.76
Asn	1.22	Gln	0.93	Ile	0.68
Ser	1.20	Arg	0.91	Met	0.67
Cys	1.16	Phe	0.90	Val	0.64

The entropy of loop closure scales linearly with the logarithm of segment length (*20*), but a confident numerical estimate requires theory that takes into account heterogeneous loops of approximately one statistical segment in length. While such an estimate is beyond the scope of this article, it is evident that the loop-closing entropy for these compact loops of 6 to 16 residues is offset, at least in part, by extensive favorable contacts within the loop.

Protein secondary structure has often been codified into a small number of states on the basis of backbone dihedral angles and hydrogen-bonding patterns. The usual categories include helix, sheet, reverse turn, and random coil. Identification of these categories is not always straightforward, and a given segment may be classified differently by different investigators. Not surprisingly, estimates of the relative abundance within these categories vary

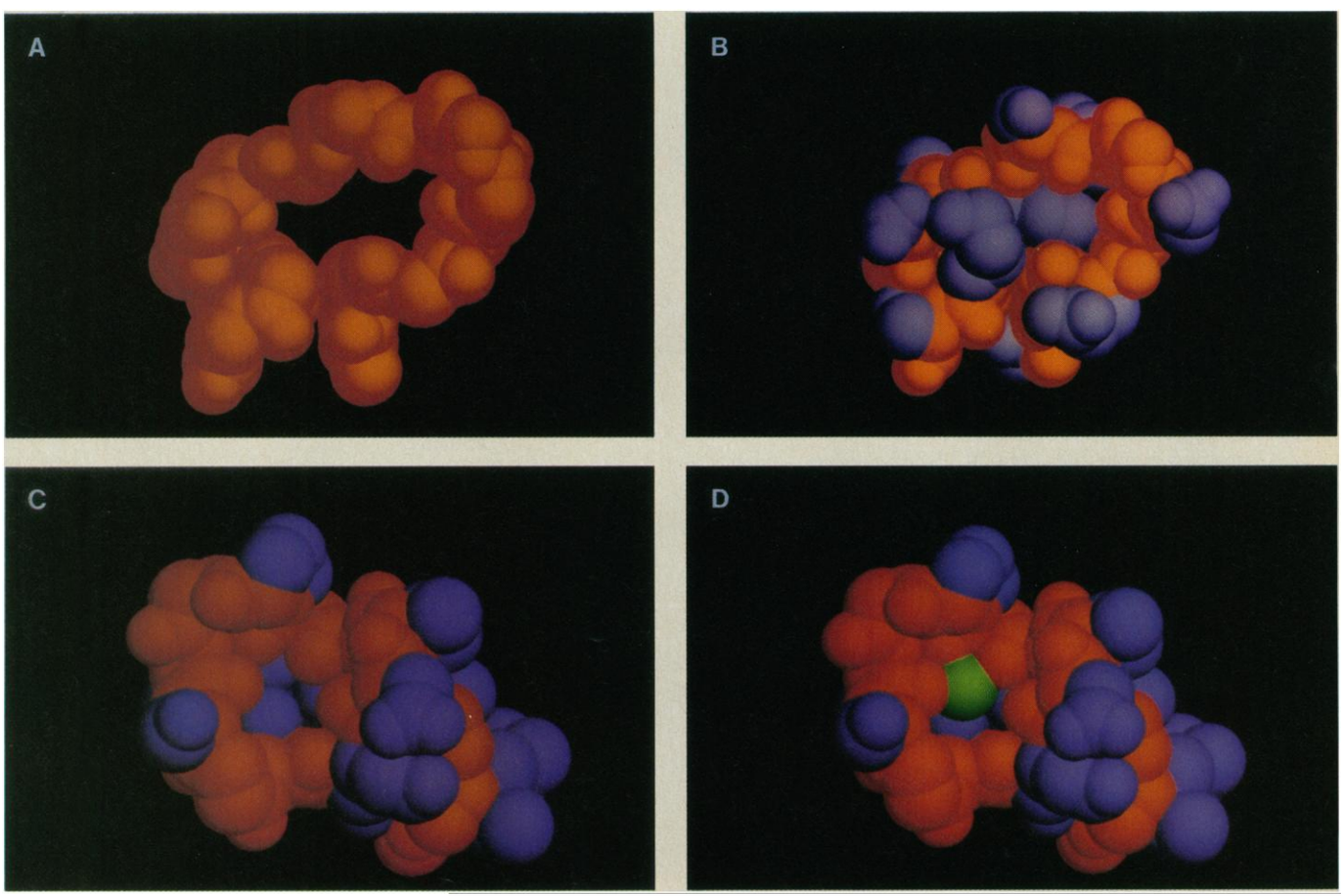

Fig. 3. Typical loops from cytochrome c (residues 40–54) and thermolysin (residues 188–203). (A and B) Space-filling representation of the cytochrome c loop, with the same orientation as Fig. 2. Backbone atoms are shown in red and side-chain atoms in blue. The loop main chain forms an internal cavity that is filled by side-chain groups from loop residues. (C and D) Space-filling representation of the thermolysin loop with backbone atoms shown in red and side-chain atoms in blue; (C) without the metal ligand, and (D) with the metal ligand, shown in green.

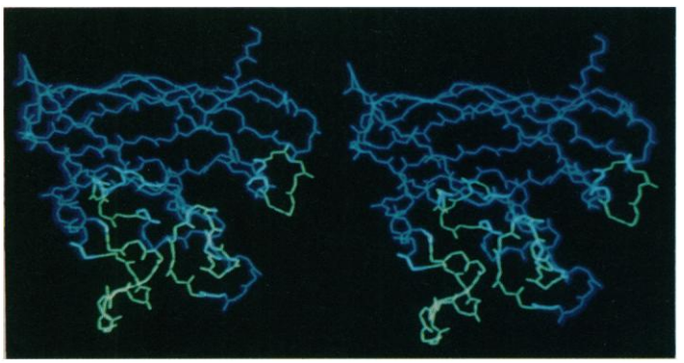

Fig. 4. Stereoview showing clustering of loops in superoxide dismutase.

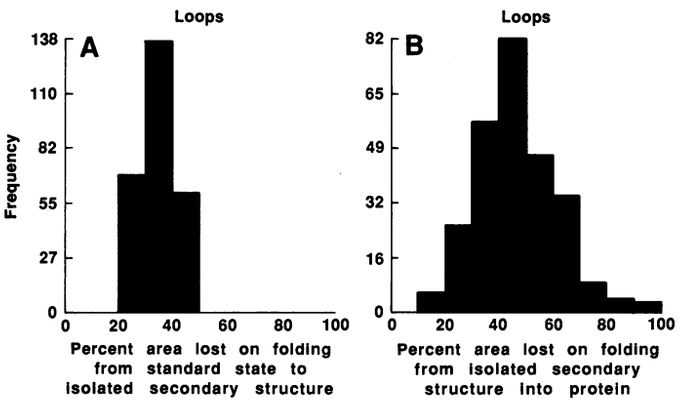

Fig. 5. Histograms showing distribution of the percentage of surface area lost for loops on folding from (A) the standard state (*18*) to the isolated secondary structure and (B) the isolated secondary structure to the native protein. The mean values (± standard deviation) are: (A) 34.4 percent (± 6.5) and (B) 46.7 percent (± 14.6).

somewhat, particularly in the case of turns (*3*). However, a consensus estimate finds that regular secondary structures—helices and sheet—make up slightly less than half of proteins, on average. Chou and Fasman (*17*) allocate another third of all residues to turns, although other estimates are closer to a quarter (*10, 21*).

Two comprehensive studies that codify residues into discrete states based on objective criteria can be found (*10, 22*). We used the assignment of Kabsch and Sander (*10*) to eliminate regular secondary structure prior to loop identification, as discussed above.

The 67 proteins in our study contain 11,885 residues: 26 percent in helix, 19 percent in sheet, 26 percent in turns, and 21 percent in loops. On the basis of the K&S assignments, calculation of the percentage of residues in helix and sheet is straightforward, but assignment of residues to turns or loops is confounded because loops contain reverse turns. In that our statistics count such residues among loops, the percentage of residues in reverse turns must be reduced accordingly. When subdivided, 11 percent of all residues are found in turns within loops and 15 percent in turns external to loops. A further correction, although slight, should be made for those two-residue strands and single-turn helices that are counted both in loops and as regular secondary structure; in combination, these two minor categories contain less than 1 percent of all residues. Subject to these adjustments, helices, sheet, reverse turns, and loops account for approximately 80 percent of all residues in the 67 proteins of our study. When our conservative criteria for defining loops are relaxed only slightly, more than 90 percent of all residues are included in the accounting.

Loops as a definitive category for structure analysis. The principal question raised by our findings is whether loops comprise a distinct class of secondary structure. Of course, the classical definition of secondary structure as hydrogen-bonded backbone structure (*23*) automatically excludes loops. In practice, however, secondary structure has come to be synonymous with the conformation of continuous segments of the polypeptide chain (*24*).

The fact that loops exist in a range of conformations would seem to argue against their classification as a discrete category. Yet, the situation is not entirely dissimilar to that of reverse turns which can range between three and four residues and adopt multiple conformations (*1–4*). This range of variability increases exponentially with segment length, and, in loops, it is extremely large.

The conspicuous compactness of most loops is the factor that most convincingly underwrites their classification as a discrete entity. They are autonomously well-folded structures because their observed compactness does not depend on interactions with the rest of the protein. Indeed, were loops amorphous, their location at the molecular surface would render them ready targets for indiscriminate proteolysis (*25*), leading to rapid protein turnover, but there is no evidence for this.

Because loops contain reverse turns, they should perhaps be viewed as structural composites, akin to supersecondary structure (*1, 26*). In any event, loops can be identified objectively in x-ray elucidated proteins, and they occur with a frequency comparable to that of the β-sheet. We propose that protein segments which satisfy our defining criteria be called omega (Ω) loops.

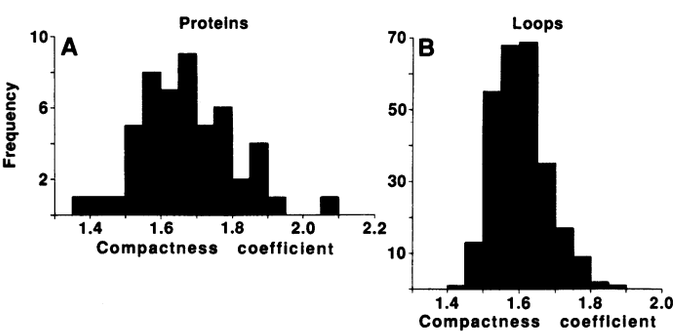

Fig. 6. Histograms showing distribution of the compactness coefficient (*Z*) for (A) the proteins used in this study and (B) the 270 loops from Table 1. The mean values (± standard deviation) are: (A) 1.67 (± 0.13) and (B) 1.61 (± 0.07).

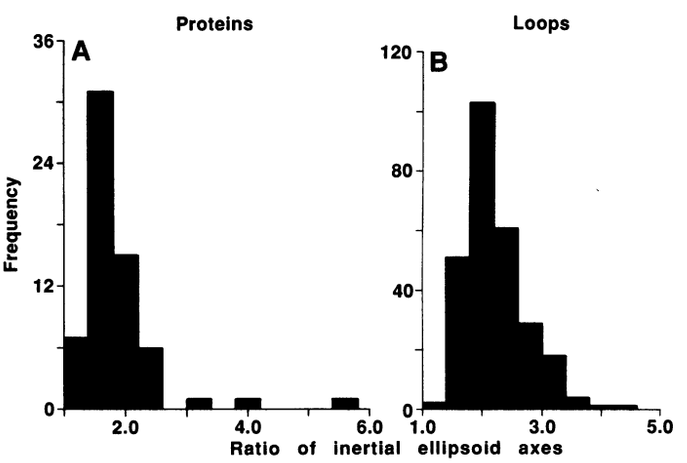

Fig. 7. Histograms showing the distribution of axial ratios for (A) the proteins used in this study and (B) all loops from Table 1. The mean values (± standard deviation) are: (A) 1.77 (± 0.66) and (B) 2.17 (± 0.51). The three outliers in (A) are glucagon, mellitin, and avian pancreatic polypeptide, three of the six proteins without loops.

The term loop has also been used to describe a chain segment that is cross-linked by a disulfide bond. Although one can envision an Ω loop with a disulfide bond between its ends, none are observed among the loops in Table 1. There are interloop disulfide bonds, but these are not situated between loop termini. In addition, the cystines in these omega loops form loop-loop and loop-protein disulfide bridges.

Some loop residues have been implicated in antibody binding. Lysozyme loops 18 to 25, 44 to 52, and 60 to 75 contain antigenic residues 19, 21, 45 to 48 (*27*), and 64 to 80 (*28*). Since the entire protein surface is thought to be potentially antigenic (*29*), loop involvement in antigenic sites may be a consequence of the fact that loops are on the protein surface.

It is unclear whether an isolated loop segment will fold independently in solution (*30*), but the proposition is testable. The questions are analogous to those raised by Bierzynski *et al.* (*31*) and Kim and Baldwin (*31*) in their studies of the independent stability of the C-peptide helix from ribonuclease. The isolated segment could be monitored for nativelike interactions by nuclear magnetic resonance. Alternatively, the conformation of loops containing suitable ligands might be probed with metal ions (*32*).

Isolated loop segments are also attractive peptides for use in model studies. They are highly solvated, both within the protein and alone in solution; and they are readily cyclized by addition of cysteines at their termini. An analogous use of cyclized peptides to model reverse turns has been successfully exploited by Gierasch and co-workers (*3–4*).

Omega loops are appealing candidates for bioengineering studies. From the data in Table 1, experiments could be designed to test the hypothesis that loops function as integral units and have the potential for modular exchange between proteins. Thus, a loop might be "swapped" or excised entirely, and the consequences for protein stability and enzymatic activity can be assessed.

Evolution appears to have established precedents for loop swap experiments. Several protein families in our database such as the cytochromes c and the serine proteases contain loops at homologous locations. Some of those homologous loops are structurally similar, while others have conserved end points, but differing overall structures.

Macromolecular recognition is a hallmark of biological systems. Recognition sites for glycosylation, phosphorylation, supramolecular assembly, and transport all reside on the protein surface. It is plausible that omega loops assume a central role in such processes.

REFERENCES AND NOTES

1. J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981).
2. C. M. Venkatachalam, *Biopolymers* **6**, 1425 (1968); I. D. Kuntz, *J. Am. Chem. Soc.* **94**, 8568 (1972); P. N. Lewis, F. A. Momany, H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **68**, 2293 (1971).
3. G. D. Rose, L. M. Gierasch, J. A. Smith, *Adv. Protein Chem.* **37**, 1 (1985).
4. J. A. Smith and L. G. Pease, *CRC Crit. Rev. Biochem.* **8**, 315 (1980).
5. J. A. Tainer, E. D. Getzoff, K. M. Beem, J. S. Richardson, D. C. Richardson, *J. Mol. Biol.* **160**, 181 (1982).
6. D. R. Davies, E. A. Padlan, D. M. Segal, *Annu. Rev. Biochem.* **44**, 639 (1975); J. S. Richardson, D. C. Richardson, K. A. Thomas, E. W. Silverton, D. R. Davies, *J. Mol. Biol.* **102**, 221 (1976).
7. A. A. Kossiakoff, J. L. Chambers, L. M. Kay, R. M. Stroud, *Biochemistry* **16**, 654 (1977).
8. R. H. Kretsinger and C. E. Nockolds, *J. Biol. Chem.* **248**, 3313 (1973).
9. A. M. Lesk and K. D. Hardman, *Science* **216**, 539 (1982).
10. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
11. M. H. Zehfus and G. D. Rose, *Biochemistry* **25**, 5759 (1986).
12. B. K. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971). Probe radius that we used was 1.4 Å. The atomic radii, in Å, are: tetrahedral C, 2.0; trigonal carbon, 1.7; carbonyl O, 1.4; hydroxyl O, 1.6; carboxyl O, 1.5; tetrahedral N, 2.0; trigonal N, 1.7; divalent S, 1.85; and sulfhydryl S, 2.0.
13. M. Y. Pavlov and B. A. Federov, *Biopolymers* **22**, 1507 (1983).
14. F. C. Bernstein *et al.*, *J. Mol. Biol.* **112**, 535 (1977).
15. T. Takano and R. E. Dickerson, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6371 (1980).
16. M. A. Holmes and B. W. Matthews, *J. Mol. Biol.* **160**, 623 (1982).
17. P. Y. Chou and G. D. Fasman, *ibid.* **115**, 135 (1977); *Annu. Rev. Biochem.* **47**, 251 (1978); M. Levitt, *Biochemistry* **17**, 4277 (1978).
18. The standard state surface area of a residue, X, is the average area of that residue in a representative ensemble of Gly-X-Gly tripeptides [see table 1 in G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, M. H. Zehfus, *Science* **229**, 834 (1985)]. The standard state surface area of a segment is taken as the sum of its residue standard states.
19. C. Chothia, *J. Mol. Biol.* **105**, 1 (1976), F. M. Richards, *Carlsberg Res. Commun.* **44**, 47 (1979).
20. P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969).
21. G. D. Rose and J. P. Seltzer, *J. Mol. Biol.* **113**, 153 (1977); G. D. Rose and D. B. Wetlaufer, *Nature (London)* **268**, 769 (1977).
22. M. Levitt and J. Greer, *J. Mol. Biol.* **114**, 181 (1977).
23. K. Linderstrom-Lang, *Lane Memorial Lectures—Proteins and Enzymes* (Stanford Univ. Press, Stanford, CA, 1952), vol. 6, p. 1.
24. A. Lehninger, *Principles of Biochemistry* (Worth, New York, 1982).
25. A. L. Goldberg and A. C. St. John, *Annu. Rev. Biochem.* **45**, 747 (1976).
26. M. Levitt and C. Chothia, *Nature (London)* **261**, 552 (1976).
27. S. J. Smith-Gill *et al.*, *J. Immunol.* **128**, 314 (1982); S. J. Smith-Gill, T. B. Lovoie, C. R. Mainhart, *ibid.* **133**, 384 (1984).
28. R. Arnon, *Immunochemistry of Enzymes and Their Antibodies*, M. R. J. Salton, Ed. (Wiley, New York, 1977).
29. D. C. Benjamin *et al.*, *Annu. Rev. Immunol.* **2**, 67 (1984); D. W. Fanning, J. A. Smith, G. D. Rose, *Biopolymers* **25**, 863 (1986); D. J. Barlow, M. S. Edwards, J. M. Thornton, *Nature (London)* **322**, 747 (1986).
30. D. B. Wetlaufer, *Adv. Protein Chem.* **34**, 61 (1981).
31. P. S. Kim and R. L. Baldwin, *Annu. Rev. Biochem.* **51**, 459 (1982); A. Bierzynski, P. S. Kim, R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2470 (1982).
32. W. D. Horrocks, Jr., *Progr. Inorg. Chem.* **31**, 1 (1984).

33. We thank Micheal Zehfus for suggestions, Lyndon Hibbard and William Young for discussion during preliminary stages of this work, and an anonymous reviewer for valuable comments. Supported by NIH grants GM29458 and AGO6084.

6 May 1986; accepted 25 September 1986