Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction

LUIS E. DONATE, STEPHEN D. RUFINO, LUC H.J. CANARD,¹ AND TOM L. BLUNDELL

The Imperial Cancer Research Fund, Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, London WC1E 7HX, United Kingdom

(RECEIVED June 12, 1996; ACCEPTED September 25, 1996)

Abstract

Loops are regions of nonrepetitive conformation connecting regular secondary structures. We identified 2,024 loops of one to eight residues in length, with acceptable main-chain bond lengths and peptide bond angles, from a database of 223 protein and protein-domain structures. Each loop is characterized by its sequence, main-chain conformation, and relative disposition of its bounding secondary structures as described by the separation between the tips of their axes and the angle between them. Loops, grouped according to their length and type of their bounding secondary structures, were superposed and clustered into 161 conformational classes, corresponding to 63% of all loops. Of these, 109 (51% of the loops) were populated by at least four nonhomologous loops or four loops sharing a low sequence identity. Another 52 classes, including 12% of the loops, were populated by at least three loops of low sequence similarity from three or fewer nonhomologous groups. Loop class suprafamilies resulting from variations in the termini of secondary structures are discussed in this article. Most previously described loop conformations were found among the classes. New classes included a 2:4 type IV hairpin, a helix-capping loop, and a loop that mediates dinucleotide-binding. The relative disposition of bounding secondary structures varies among loop classes, with some classes such as β -hairpins being very restrictive. For each class, sequence preferences as key residues were identified; those most frequently at these conserved positions than in proteins were Gly, Asp, Pro, Phe, and Cys. Most of these residues are involved in stabilizing loop conformation, often through a positive ϕ conformation or secondary structure capping. Identification of helix-capping residues and β -breakers among the highly conserved positions supported our decision to group loops according to their bounding secondary structures. Several of the identified loop classes were associated with specific functions, and all of the member loops had the same function; key residues were conserved for this purpose, as is the case for the parvalbumin-like calcium-binding loops. A significant number, but not all, of the member loops of other loop classes had the same function, as is the case for the helix-turn-helix DNA-binding loops. This article provides a systematic and coherent conformational classification of loops, covering a broad range of lengths and all four combinations of bounding secondary structure types, and supplies a useful basis for modelling of loop conformations where the bounding secondary structures are known or reliably predicted.

Keywords: loop analysis, biological functions of, classes, conformation, structure modeling and prediction; protein secondary structure

It is estimated that 30% of globular proteins comprises nonrepeating loops connecting regular elements of secondary structure. These tend to be the least well defined and the most error-prone regions of a protein in both X-ray crystallography and NMR analysis; they

are also the most difficult regions to model using knowledge-based procedures.

Loops have been analyzed and classified according to various structural properties and relationships, among them main-chain conformation, size, inter-C α distances, hydrogen bonding patterns, orientation, and type of secondary structure flanking the loop (Rose et al., 1985; Sibanda & Thornton, 1985; Leszczynski & Rose, 1986; Efimov, 1986, 1991a,b,c, 1993a,b; Edwards et al., 1987; Thornton et al., 1988; Milner-White et al., 1988; Sibanda et al., 1989; Chothia et al., 1989; Rooman et al., 1990a,b; Rice et al., 1990; Milner-White, 1990; Srinivasan et al., 1991; Sowdhamini et al., 1992; Ring et al., 1992; Yura et al., 1993; Chan et al., 1993;

Reprint requests to: Tom L. Blundell, Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom; e-mail: tom@cryst.bioc.cam.ac.uk.

¹Present address: Equipe Systèmes Moléculaires et Biologie Structurale, LMCP, Universités P6, P7 associé au CNRS (URA09), Case 115, 4 Place Jussieu, 75252 Paris Cedex 05, France.

Swindells, 1994; Nataraj et al., 1995; Sun & Blundell, 1995; Swindells et al., 1995; Wintjens et al., 1996). These analyses have shown that loops do not have completely random backbone conformations, and a limited number of favored conformations for the loops were observed. For example, Efimov (1991a,b,c; 1993a,b) described several typical short connections for $\alpha\alpha$ - and $\beta\beta$ -motif classes, as well as some $\alpha\beta$ - and $\beta\alpha$ -arches, and considered the conformation of longer loops to be a combination of the standard conformations of the short ones.

A loop region is characterized by its conformation, sequence, and the spatial relations of its bounding secondary structures, features that are interdependent. Thus, the probability of finding a particular amino acid at a position in a loop of particular conformation will depend on conformational preferences, solvent accessibility, and ionic interactions (Tramontano et al., 1989; Topham et al., 1993). For example, at loop positions requiring a positive ϕ conformation, a restricted set of residue types is observed to occur. In the case of 2:2 β -hairpins of type I', Sibanda and Thornton (1985) identified the residue preferences as Gly, Asn, and Asp for the first position in an α_L conformation and Gly for the second position in a γ_L conformation.

In general, the amino acid sequences of the loop regions do not provide a fingerprint that can be used to identify the presence of a loop of a particular conformation anywhere in a protein sequence. However, if the position and nature of the neighboring β -strands and helices are known or suspected on the basis of the known three-dimensional structure of a homologue, or from a reliable secondary structure prediction, then the particular conformation of the connecting peptide may be identified by comparison with sequence templates or patterns that characterize loop classes. These may be useful in comparative protein modelling (Greer, 1980; Blundell et al., 1988; Thornton et al., 1988; Sibanda et al., 1989; Greer et al., 1990a,b; Overington et al., 1990a,b; Topham et al., 1993) as well as in suggesting conformations of super-secondary motifs where a satisfactory structure prediction has been performed.

In this paper we describe the extraction of 2,024 loops of length between one and eight residues from a dataset of 223 protein structures. This provides a database of loop conformers that can be used for modeling and prediction. In comparative modeling the nature and the boundaries of the adjacent secondary structures, the angles between them, and the distances between the termini connected by the loop are known; our database identifies the sequence preferences and disposition in space of the bounding secondary structures of each loop conformational class. We discuss loop classes with interesting sequence preferences or spatial dispositions of bounding secondary structures, and we compare the identified loop conformational classes to those previously described, highlighting those reported for the first time. The possibility of grouping conformational classes of loops of different lengths into conformational suprafamilies is addressed. Finally, the relation of conformational class to biological function is discussed.

Results and discussion

Loop database

A loop was defined as a region of nonrepetitive secondary structure bounded by two secondary structures, while its motif was defined to include both the loop and its adjacent secondary structures. Each loop motif was described by its sequence, its main-chain conformation in terms of the five residue conformational

regions in the Ramachandran plot, and the spatial disposition of its bounding secondary structures. The spatial disposition of the bounding secondary structures in a loop motif was defined by the separation and the angle between vectors (Chou et al., 1984; Srinivasan et al., 1991; Sowdhamini et al., 1992) representing the local axis of the proximal termini (i.e., that bordering the central loop) of the secondary structures. An angle of zero indicates a linear arrangement of secondary structure vectors while an angle of 180 degrees denotes a chain reversal.

From a database of 223 protein and protein domain structures of 2.7 Å or better resolution, checked with PROCHECK (Morris et al., 1992), 2,757 loops were identified. These ranged in length from 0 to 75 residues in the case of a loop in beef liver catalase (7 cat), although the latter included three 3_{10} -helices of only three residues (Martin et al., 1995). Lengths of 0 were expected for loops linking β -strands to helices, helices to β -strands, and helices to helices, because contiguous 3_{10} -, α -, or π -helices were not merged. Of the 2,757 loops, 1,065 (39%) linked two β -strands (EE), 598 (22%) linked a β -strand to a helix (EH), 567 (21%) linked a helix to a β -strand (HE), and 527 (19%) linked two helices (HH). These percentages are very similar to the 37% of EE loops, 24% of EH loops and HE loops, and 15% of HH loops that would be predicted if there were no sequential preferences.

The length distribution of EH loops peaks at one while that of HE loops peaks at four (Fig. 1). This difference is also reflected in the average length of the loops, with values of 5.4 for EH loops and 6.5 for HE loops. The length distributions of both EE loops and HH loops have two sharp peaks. In the case of EE loops, the first peak is at length two and the second at length five, while in the case of HH loops the first peak is at length one and the second at length five. Of the 27 HH loops of length zero, 13 link a 3_{10} -helix to an α -helix and 14 an α -helix to a 3_{10} -helix. To see whether small β -bulges represent a large fraction of short loops, the intervector angle distributions of EE loops of length one to six were calculated at 20-degree intervals (Fig. 2A). EE loops of length one, which do not have the conformational freedom to be involved in sharp chain turns, do not have a major peak at around 160 degrees; instead they have one at 40–80 degrees, possibly reflecting the observation that EE loops of length one are major contributors to orthogonal $\beta\beta$ -motifs (Sowdhamini et al., 1992). Although some of the EE loops of length one might correspond to β -bulges, they will be involved in large distortions of at least five residues in their paired β -strand because the procedure used for identifying secondary structures includes in β -strands only those β -bulges involving at the most one residue in one β -strand and four in the other (Kabsch & Sander, 1983). The distribution of loops of length two linking two β -strands shows no major peaks at small angle values while that of loops of length three shows a reduced peak at 160 degrees and a second major peak at small angle values around 40–80 degrees. With the exception of loops of length zero, the intervector angle distributions of HH loops of six or fewer residues show preferred values around 80 and 180 degrees (Fig. 2B). Only the intervector distribution of HH loops of length zero show major density at small angle values with a peak at 20–80 degrees.

Except where specifically stated, loops of length zero are not considered any further in this study because its focus is on loop main-chain conformations. Loops longer than eight were also ignored because their small numbers and the vast conformational space available to them do not allow significant identification of populated loop conformational classes. Of the 2,757 loops, those with lengths of one to eight number 2,058 and of these 2,024 have

Fig. 1. Length distribution of loops segregated according to the type of their bounding secondary structures.

main-chain bond lengths and peptide bond angles within acceptable ranges (Engl & Huber, 1991). These loops represent 73% of the 2,757 initial loops, with percentages ranging from 69% for EH loops to 77% for HE loops.

Clustering of loops and identification of loop conformational classes

The 2,024 loops of length one to eight were grouped into a total of 32 groups according to their length and to the type of their bounding secondary structures (Table 1). Secondary structure type is considered because capping patterns and other interactions are specific to α -helices and β -strands. The loops within each group were pairwise superposed, and populated loop main-chain conformation classes were then identified (Table 2) by clustering the loops according to their pairwise structural similarity (Felsenstein, 1985) (Fig. 3A and B). The variability of the main-chain conformation, intervector separation, and angle of their member loops in the identified clusters increases with loop length. Although the wide intervector relation distributions of some clusters clearly reflect spatial plasticity, in some cases a few isolated outlying members uncharacteristic of their clusters were removed and clusters with clearly bimodal distributions were split.

A loop conformational class is defined (Table 2) as a cluster containing at least four nonhomologous loops or loops differing in sequence by at least one quarter of their residues. In addition to the "populated" classes, "small" classes are clusters containing at least three loops, differing in sequence by at least one quarter of their residues, from at the most three nonhomologous groups. Of the 2,024 motifs studied, 63% cluster into a total of 161 loop classes with 51% in 109 "populated" classes and an additional 12% in 52 "small" classes. The name of a loop class is derived from the type of its bounding secondary structures and the five-state main-chain conformation (a, b, e, g, and t) of its member loops. Alternate conformations are indicated by a backslash; for example, Haab/tH describes a loop class of length three, conformation aab or aat, linking two helices. The "small" classes are indicated by a final "s" in the name (e.g., EaeaaHs).

Some loop classes were further subdivided on the basis of their sequences. This was achieved by constructing sequence templates describing the probabilities of each of the 20 amino acids at each loop position for each loop class (Rufino et al., 1996). The amino acid probabilities of a loop class were determined by applying conformationally constrained environmental substitution tables (Topham et al., 1993) to the sequences of its member loops. The sequence template of each loop class was then tested against each of its member loops so as to identify loops whose sequences scored poorly. Most such cases were confined to short loop classes of length one, with one class of length two, and were due to opposing trends within the loop class sequence template. For example, the HaH class was found to contain two sets of loops, the first with a hydrophobic residue such as Val, Ile, Leu, Ala, or Phe and the second with a Gly or a charged or polar residue such as Asp, Thr, or Tyr. In such cases the loop classes were split, with the HaH class being separated into the Ha-VILAF-H and Ha-GDTY-H classes. When the sequence pattern involved more than one position, a simple numbering scheme was used to distinguish them with, for example, Eaa-2-E describing one of several loop classes of length two, conformation aa, linking two β -strands. In all, 10% of clustered loops belonged to classes split on the basis of conflicting sequence patterns.

Distribution of loop conformational classes

A smaller proportion of the longer loops cluster into classes compared with shorter loops, with values decreasing from 87% for loops of length one to 29% for loops of length eight, as expected from the fact that longer loops have a lower density in conformational space. A total of 54 (33%) classes are EE classes, 42 (26%) HE classes, 33 (20%) EH classes, and 33 (20%) HH classes, reflecting the overall distribution of loops. Wintjens et al. (1996) report a similar distribution, with 23 $\beta\beta$ families, 15 $\alpha\beta$ families, 15 $\beta\alpha$ families, and 10 $\alpha\alpha$ families. The number of classes decreases slightly from 23 at length one to 21 at length two, rises to 28 at length four and then decreases to seven at length eight, with six of these being EE classes and one an HH class (Fig. 4). For

Fig. 2. Intervector distribution for (A) EE loops and (B) HH loops grouped according to their length.

short loops the number of loop classes is limited by the number of available conformations, while for long loops it is limited by the loop density in conformational space. Overall, 65% of EE loops cluster into classes, as do 61% of EH loops, 64% of HE loops, and 62% of HH loops. These percentages are much greater than the 15% of EH loops and HE loops classified into structural families by Edwards et al. (1987), a fact probably explained by the small number of loops, 129, available to those authors at the time.

Distribution of intervector relations

For each loop class the sequence preferences and average values and standard deviations of member loop intervector relations were recorded. While some classes have very restricted intervector separation and angle, others show greater variation. For example, the EaaagE class, which corresponds to the type I 4:4 β -hairpins iden-

tified by Sibanda and Thornton (1985; Sibanda et al., 1989), has standard deviations for its distribution of intervector separations and angles of 0.32 Å and 13.4 degrees, respectively, while the Et/a/g/baabE class has standard deviations of 1.58 Å and 28.6 degrees, respectively.

Grouping loop classes into suprafamilies

Because assignments of helical and β -strand termini, and therefore of loop boundaries and length, depend on definition and identification of hydrogen bonds, loop length cannot be viewed as an absolute criterion. It is, therefore, expected that some longer loops will contain the conformation of the smaller ones, flanked by residues in the same conformation as the bounding secondary structures. For example, loops of conformation HbaE, Habae, HbabE, HaabaE, and so on, might be included in an H-ba-E suprafamily

Table 1. Distribution of loops according to their length and the type of their surrounding secondary structures

	No. of loops	No. of loops in classes	% of loops in classes	No. of classes
EE1	27	20	74	3
EE2	164	122	74	7
EE3	104	75	72	9
EE4	118	73	62	9
EE5	148	104	70	9
EE6	72	33	46	5
EE7	69	33	48	5
EE8	81	39	48	6
EH1	85	74	87	7
EH2	67	46	69	4
EH3	73	36	50	5
EH4	55	35	64	6
EH5	55	37	67	7
EH6	35	13	37	3
EH7	25	6	24	1
EH8	20	0	0	0
HE1	20	17	85	5
HE2	43	34	79	5
HE3	81	67	83	8
HE4	80	61	76	7
HE5	65	28	43	5
HE6	77	41	53	8
HE7	48	23	48	4
HE8	23	0	0	0
HH1	76	69	91	8
HH2	39	21	54	4
HH3	52	38	73	4
HH4	58	49	84	6
HH5	61	36	59	5
HH6	41	18	44	3
HH7	38	6	16	2
HH8	24	4	17	1

Note: The number and percentage of loops that cluster into a conformational class, as well as the number of classes, is indicated.

containing the basic conformational unit [ba]. Such groupings could be due to poor definition of the bounding secondary structures in a motif, while others could reflect a genuine "softness" characteristic of certain loops that allows them to bridge suboptimal geometric dispositions. The loops in each of the four major types, namely EE-, EH-, HE-, and HH loops, were examined and numerous examples of loops sharing such conformational units were found.

Prior to grouping into suprafamilies, loop motifs were first classified as hairpin-like, V-shaped, or L-shaped, according to the angle between their secondary structures, and EE loops were classified as hairpins when their β -strands were hydrogen bonded. Only loops with similar geometric arrangements were considered for grouping into a suprafamily, thus β -hairpins with incompatible hydrogen bonding ladders were classified in different suprafamilies.

Suprafamilies with at least two conformational classes and containing nine or more members are summarized in Table 3. Suprafamilies are given a subscript of two letters denoting hairpin (HP), V-shaped (VS), or L-shaped (LS) loops. There were 18 suprafamilies that met the two former criteria, accounting for 59

loop conformational classes and 600 loops. Each suprafamily was assessed by first extending all its member loops to the length of its longest loop and then pairwise superposing and clustering these in the context of all loops of this maximum length. Both EE-loop and HH-loop suprafamilies were found to cluster reasonably tightly while some of the EH-loop and HE-loop suprafamilies formed split clusters or clusters that included unrelated loops. Overall, suprafamilies with member loops of very diverse length and those with long loops did not cluster tightly, while the best clustering was obtained for the β -hairpin suprafamilies.

EE loops were the most numerous and thus provided the greatest number (eight) of suprafamilies, of which six are β -hairpin suprafamilies. In contrast, EH loops provided only one suprafamily. Within the suprafamilies, the degree of relaxation at the ends of secondary structures was observed to vary; it is at its minimum for the β -hairpin suprafamilies, which usually contain representatives of only two different lengths. Thus 88% of all the [gg] hairpin loops are classified as E-[gg]_{HP}-E with very few loops in the EbgbE conformation, and the great majority of loops in the E-[aa]_{HP}-E (Table 3) suprafamily have the EbaagbE conformation, which corresponds to the 3:5 turns of Sibanda and Thornton (1985), providing an example of highly favored β -strand relaxation. However, the E-[aa]_{HP}-E suprafamily significantly deviates from this trend because it includes representatives of all four possible β hairpins, with the basic unit [aa] at the center of the loop. On the other hand, relaxation of either the helix or the strand or both in the H-[g]_{HP}-E suprafamily is very common.

In most cases, a similar amino acid pattern was found in the basic conformational units of loops of all lengths belonging to the same suprafamily. For example, in the E-[ab]_{HP}-H suprafamily, Asp, Asn, and Ser are common at the second position, and in the H-[gbb]_{HP}-H and H-[gbb]_{LS}-H suprafamilies, Gly is highly preferred at the first position, hydrophobic residues occupy always the second, and Ser, Pro, and Asp and Asn are common at the third. However, in the E-[aaag]_{HP}-E suprafamily, three quarters of the member loops including all of EaaagE conformation show a strong preference for Asn (followed by His and then Gly) at the fourth position in the basic unit, while others including all of EbbaagbE conformation prefer Gly in the fourth position of the [aaag] unit. In view of such observations we have maintained the individual loop conformational classes for prediction and homology modeling purposes.

Sequence frequencies

The observed residue frequencies at each loop position allow specific sequence preferences to be identified. For example, the EggE class (Table 4a), which corresponds to the type I' 2:2 β -hairpins, has a marked preference for Asn, Asp, and Gly at the first position, with frequencies of 0.26, 0.24, and 0.18, respectively, while Gly is strongly preferred at the second position, with a frequency of 0.9 followed distantly by Asp with a frequency of 0.06. These preferences are similar to the sequence patterns identified by Sibanda and Thornton (1985; Sibanda et al., 1989). In the case of the EaaE class (Table 4b), which corresponds to the type II' 2:2 β -hairpins, residue preferences are Gly and Asp at the first position, with frequencies of 0.8 and 0.12, respectively, and Ser, Asp, Asn, and Thr at the second position, with frequencies of 0.3, 0.2, 0.16, and 0.12, respectively, which are again similar to those found by Sibanda and Thornton (1985; Sibanda et al., 1989).

Table 2. The identified loop conformational classes

Class	Length	Member	Weight	Separation		Angle		Reference
				Average	Deviation	Average	Deviation	
EaE	1	10	3	5.72	0.73	78	26	$\beta\alpha\beta \beta\gamma\beta$; Efimov, 1993
EbE	1	7	6	6.95	0.91	43	21	
EtEs	1	3	2	5.44	0.18	53	17	
EbaE	2	8	5	9.00	0.97	91	30	
EeaE	2	35	25	5.28	0.37	153	14	$\epsilon\gamma_R$, Sibanda & Thornton, 1985
Eaa-1-E	2	8	7	5.47	0.29	147	12	
Eaa-2-E	2	8	7	5.31	0.44	155	10	$\alpha_R\gamma_R$; Sibanda & Thornton, 1985
EgbEs	2	4	3	9.47	0.26	84	34	
EbbE	2	5	5	9.46	0.27	93	37	
EggE	2	56	39	5.32	0.41	147	14	$\alpha_L\gamma_L$; Sibanda & Thornton, 1985
EbaeEs	3	12	2	11.94	0.60	43	23	
Etae/bEs	3	4	2	12.04	0.18	19	5	
EaagE	3	6	5	5.93	0.77	126	18	$\beta\alpha_R\alpha_R\alpha_L\beta$; Srinivasan et al., 1991
EaaaE	3	8	5	5.60	1.10	145	25	$\beta\alpha\alpha\alpha\beta$; Efimov, 1993
Ebga/a/bE	3	9	7	6.18	1.29	148	11	$\beta\gamma_L\alpha_R \beta\alpha_L\alpha_R$; Nataraj et al., 1995
EggaE	3	8	6	4.75	0.49	150	10	$\alpha_L\alpha_L\alpha_R$; Nataraj et al., 1995
EabgEs	3	7	3	8.35	0.46	67	2	
EbbbE	3	11	8	12.47	0.39	51	20	
Ea/g/tbbE	3	11	8	11.38	1.58	93	28	
Eabbb/a/eE	4	5	5	12.26	0.72	72	15	
Et/bgggE	4	8	6	5.18	0.38	151	10	
EagabE	4	5	3	4.91	0.26	162	6	
EbbgbEs	4	4	3	8.00	0.98	136	29	$\beta\beta\alpha_L\beta$; Efimov, 1993
EaaagE	4	28	18	5.28	0.32	156	13	$\alpha_R\alpha_R\gamma_R\alpha_L$; Sibanda & Thornton, 1985
Et/a/g/baabE	4	4	4	12.06	1.26	65	28	
EbaabE	4	9	9	4.94	0.36	161	9	$\beta\gamma_R\gamma_R\beta$; Efimov, 1991
EbabbEs	4	5	3	12.86	0.96	134	22	
EeaagEs	4	5	2	4.94	0.18	148	6	
EbaagbE	5	49	35	4.96	0.46	149	13	$\beta\alpha_R\gamma_R\gamma_L\beta$; Sibanda & Thornton, 1985
EbbgggEs	5	14	2	6.82	1.80	156	4	
EabaabE	5	14	6	5.06	0.35	149	11	
Ebba/bb/ebEs	5	8	3	13.11	0.31	158	4	
EbbgggEs	5	3	3	5.09	0.27	154	0	
EagabEs	5	5	3	4.87	0.13	153	6	
EaababEs	5	4	1	10.37	0.16	144	2	
EbbgbtEs	5	3	1	5.06	0.13	166	1	
Eb/taaab/aE	5	4	4	11.67	0.73	61	26	
EbaagabE	6	10	5	4.64	0.65	145	24	
Eaabb/ab/tbE	6	5	4	5.26	0.70	148	9	
EbaaagbE	6	9	8	4.78	0.22	155	12	$\beta\alpha_R\alpha_R\alpha_R\alpha_L\beta$; Srinivasan et al., 1991
EbaaagbEs	6	4	1	4.55	0.03	162	2	
Eb/taababEs	6	5	1	10.14	0.15	146	1	
Eb/tbb/a/eg/t/a/ebbbE	7	14	10	5.20	0.33	156	10	
Ebb/ag/tbbaaEs	7	5	2	8.68	0.27	152	2	
Eg/bbbaaEs	7	4	2	16.95	3.94	124	9	
Ea/tb/abbbba/bEs	7	5	2	16.83	1.26	128	4	
Ebaaa/ta/bgb/aE	7	5	4	5.86	1.36	143	12	
EbbaagbbaE	8	9	6	5.10	0.47	147	21	
EbbbgataaEs	8	6	1	14.45	0.16	92	5	
Eab/tbbaaEs	8	5	2	12.91	0.30	149	1	
EbgbaaEs	8	9	3	13.97	0.76	148	2	
Egbb/taabbEs	8	4	1	13.53	0.29	128	2	
Ebb/ea/eb/abbE	8	6	4	5.05	0.46	150	24	
Eb/pH	1	5	5	7.02	0.52	90	16	
EbH	1	5	4	7.94	1.58	111	25	
Eb-HTAL-H	1	14	12	7.01	0.93	93	25	
Eb-MKEW-H	1	7	5	6.96	0.48	106	30	
Eb-GSDP-H	1	32	19	7.07	0.71	108	34	
EaH	1	5	4	5.63	0.13	67	17	
EtH	1	4	4	6.86	0.32	59	9	

(continued)

Table 2. Continued

Class	Length	Member	Weight	Separation		Angle		Reference
				Average	Deviation	Average	Deviation	
Ebb/eH	2	10	3	9.19	0.23	104	6	$\beta\alpha_2$; Rooman et al., 1990
EabH	2	20	11	8.24	0.59	147	13	
Ebb/tH	2	14	11	10.10	0.61	90	50	
EbgHs	2	3	2	7.58	0.14	124	29	
Eaaa/tH	3	9	8	7.65	0.99	129	24	$\beta\alpha_3$; Edwards et al., 1987
Eba/ebH	3	8	7	10.71	0.80	142	28	
EbaaH	3	5	5	7.43	0.50	85	23	
Eet/abH	3	9	4	9.75	0.29	168	6	
EbbbH	3	7	4	13.41	0.79	83	23	$\beta\alpha\beta\beta$; Efimov, 1993
Eb/aaab/aH	4	13	12	10.01	1.99	102	43	
Ea/babb/aH	4	4	4	10.68	0.39	116	24	
Eb/t/ebbb/eH	4	5	5	14.03	0.35	105	50	
Ebba/tb/eH	4	6	5	14.44	0.98	122	9	$\beta_m\beta\beta\alpha_L\beta\beta\alpha_n$; Efimov, 1993
EaeaaHs	4	3	2	11.26	0.11	132	38	
EaggbHs	4	4	2	11.15	0.34	155	5	
Eb/c/a/gg/b/tbabH	5	5	5	13.95	0.53	142	8	
Eb/ebbgb/tHs	5	5	3	12.35	0.79	153	15	$\alpha\alpha_L\beta$; Srinivasan et al., 1991
Ea/baab/tbH	5	7	6	12.12	1.40	143	24	
Eb/aaaab/e/tH	5	4	4	9.63	0.97	160	12	
Eb/ab/tga/bg/t/bHs	5	4	3	9.15	0.35	133	15	
Ea/t/g/bbb/t/g/abbH	5	7	6	14.16	1.67	117	27	$\alpha\alpha_L\beta$; Srinivasan et al., 1991
Eb/abaaaH	5	6	5	9.42	1.43	114	16	
Ea/b/taaba/tb/a/tH	6	7	6	10.88	1.76	115	39	
EbbbgagHs	6	3	1	8.45	0.77	144	4	
EbbaabbHs	6	3	3	14.54	0.62	131	12	$\alpha\alpha_L\beta$; Srinivasan et al., 1991
Ea/b/gbb/abaaaH	7	6	4	14.86	2.02	116	52	
Hb-QLRS-E	1	5	4	6.33	0.87	56	12	
Hb-G-Es	1	3	2	7.47	0.39	125	27	
HeEs	1	3	2	7.72	1.06	127	30	$\alpha_m\gamma\alpha_L\beta\beta_n$; Efimov, 1993
Hg-G-E	1	4	4	6.65	1.23	128	12	
Hg-SR-Es	1	2	2	6.74	0.08	149	10	
HaaE	2	8	5	6.38	0.60	55	15	
HgbE	2	10	10	9.12	0.65	148	13	$\alpha\beta\alpha_L\beta$; Srinivasan et al., 1991
HbbEs	2	5	2	9.39	0.43	109	30	
HbaE	2	5	5	7.85	0.33	138	9	
HagE	2	6	5	5.95	0.94	121	17	
Haga/tE	3	5	5	8.11	1.41	102	21	$\alpha\beta\alpha_L\beta$; Srinivasan et al., 1991
HagbE	3	28	20	7.99	1.16	143	16	
HbgbE	3	5	4	8.74	0.99	108	15	
Haab/t/aE	3	5	5	8.53	0.75	70	41	
HbabEs	3	3	3	10.63	1.13	117	34	$\beta\alpha\beta\beta$; Efimov, 1993
Hg/abaE	3	13	11	10.41	1.36	139	22	
Hg/abbE	3	6	5	10.89	0.65	159	4	
Hbb/tbE	3	5	5	12.20	0.48	92	25	
Ht/eaab/aE	4	10	8	9.87	1.50	150	15	$\alpha\beta\alpha_L\beta$; Srinivasan et al., 1991
Hb/a/gabb/a/eE	4	9	7	11.37	1.42	133	27	
HagbbE	4	8	7	9.96	1.23	157	3	
HagbaE	4	20	15	9.93	0.73	161	9	
Ha/gagb/aE	4	6	6	9.11	3.38	154	15	$\alpha_m\gamma\alpha_L\beta\alpha\beta\beta\beta_n$; Efimov, 1993
Hgbt/bbE	4	5	4	14.38	1.07	159	7	
HagagEs	4	3	2	10.22	1.00	139	27	
Hb/abgbE	5	4	4	9.65	3.01	142	30	
Hagb/tb/ab/t/gE	5	9	7	11.12	1.10	149	23	$\alpha_m\beta\beta\alpha_L\beta\beta_n$; Efimov, 1993
Ha/taaba/bE	5	7	7	10.51	0.97	147	16	
HagabaEs	5	4	3	9.85	1.41	161	9	
Hb/aaagaE	5	4	4	8.33	0.13	129	5	
HagababEs	6	4	3	11.46	0.38	131	2	$\alpha_m\gamma_L\alpha\beta\beta_p\alpha\alpha\gamma\beta_n$; Efimov, 1993
HgbbaaaEs	6	5	3	11.87	1.45	131	18	
HtaagagE	6	11	6	8.77	0.27	47	11	
Haabb/ag/ab/aE	6	6	4	8.89	0.80	141	25	
Hag/agabb/a/eE	6	6	5	11.10	1.90	164	11	

(continued)

Table 2. Continued

Class	Length	Member	Weight	Separation		Angle		Reference
				Average	Deviation	Average	Deviation	
Hg/agb/taabEs	6	5	2	8.39	1.25	58	6	
Ht/aaaabaEs	6	3	3	15.11	0.43	155	11	
Hab/taabaEs	6	3	3	9.23	0.96	129	25	
Ht/baagagbE	7	10	4	8.66	0.17	49	4	
Ht/aaabab/atEs	7	4	2	12.05	1.86	158	16	
Hagab/tabb/a/tEs	7	5	2	13.64	2.25	152	5	
Hagab/ag/b/aa/bb/aE	7	4	4	13.27	1.51	142	20	
HeH	1	5	4	7.80	0.45	120	6	$\alpha_m\epsilon\alpha_n$; Efimov, 1993
HbHs	1	3	2	5.95	0.92	40	19	$\alpha_m\beta\gamma\alpha_n$; Efimov, 1993
Hb-SQYT-H	1	6	5	6.17	0.38	100	18	$\alpha_m\beta\alpha_n$; Efimov, 1993
Hb-DWN-H	1	10	5	6.19	0.87	84	11	$\alpha_m\beta\alpha_n$; Efimov, 1993
Ht-FYT-H	1	9	3	5.61	0.73	120	14	
Ht-DHN-H	1	13	7	6.22	0.25	128	11	
Ha-GDTY-H	1	12	8	4.53	0.62	47	31	
Ha-VILAF-H	1	11	8	5.09	0.83	79	29	
Hat/aH	2	4	4	7.16	1.54	85	19	
Ha/gbH	2	6	4	8.47	1.11	108	30	$\alpha_m\gamma\alpha_L\beta\alpha_n$; Efimov, 1993
HbbH	2	7	7	9.45	0.45	145	15	$\alpha_m\gamma\beta\beta_P\alpha_n$; Efimov, 1993
HagH	2	4	4	5.91	0.47	55	24	$\alpha_m\gamma_L\alpha_n$; Efimov, 1993
Hg/abbH	3	10	8	11.78	1.04	101	12	$\alpha_L\beta\beta$; Thornton et al., 1988
Haab/tH	3	15	13	8.96	0.99	111	43	$\alpha_m\gamma\beta/\delta\alpha_n$; Efimov, 1993
Haga/tH	3	10	10	7.89	0.96	125	37	$\alpha G\alpha$; Wintjens et al., 1996
HtaaHs	3	3	3	9.88	1.14	85	36	$\alpha_m\delta\alpha/\gamma\alpha_n$; Efimov, 1993
Ha/gbaa/bH	4	6	4	9.21	2.68	123	34	
Ha/g/babbH	4	10	9	12.01	0.74	130	32	$\alpha BABB\alpha$; Wintjens et al., 1996
HbabaHs	4	4	1	7.84	0.35	110	1	
HagbbH	4	12	8	11.75	0.94	123	30	$\alpha_m\gamma\alpha_L\beta_P\beta_P\alpha_n$; Efimov, 1993
Ha/c/b/taat/a/bH	4	11	7	8.83	1.55	104	58	$\beta\alpha\gamma\beta$; Efimov, 1993
Hg/a/bbba/b/tH	4	6	5	13.12	2.18	129	7	$\alpha BBBBB\alpha$; Wintjens et al., 1996
Ht/a/caab/tH	5	8	6	12.90	1.10	110	27	
HgbbbbbH	5	6	4	16.68	0.62	121	15	$\alpha\alpha_L\beta\beta\beta\beta\alpha$; Srinivasan et al., 1991
HaagbbH	5	14	12	10.31	0.99	115	23	
He/baat/bbH	5	5	4	12.90	2.18	128	6	$\beta\alpha_R\alpha_R\beta\beta$; Thornton et al., 1988
Hb/aab/tbbHs	5	3	3	11.58	1.18	88	23	
Hagbb/a/tb/tb/aH	6	7	5	14.90	1.86	134	26	
Ht/g/aaaat/bbHs	6	7	3	15.31	0.55	119	16	
HaagbbaHs	6	4	1	12.42	0.47	136	3	
Hagbbaaa/bHs	7	3	3	14.44	1.77	123	5	
HgbbbbbbHs	7	3	1	21.69	0.16	97	0	
HaabgabbHs	8	4	1	12.93	0.17	130	3	

Note: Includes length, number of member loops, sum of member loops' weight, and distribution of intervector separation and angle in terms of average values and standard deviations.

Sequence entropy

The sequence entropy of a distribution is a measure of its information content. Thus, a class loop position for which all amino acid types are observed with equivalent frequencies among member loops is permissive with a normalized entropy of one. On the other hand, a position with only one observed amino acid type has an entropy of zero. Such residues have been called key residues. The normalized entropy of a loop position is a function of the number of nonhomologous member loops (Fig. 5). Loop positions with unexpectedly restrictive amino acid preferences were then identified as those whose normalized entropies were less than one half the average value for their respective group. A cut-off in

number of standard deviations could not be used because the entropies of loop classes with few independent members do not follow a normal distribution.

Of the 638 loop positions, 130 have normalized entropies below the selected cut-off. A high proportion, 99 (76%), of loop positions with low cut-offs are from "small" classes, although the cut-off was defined as a function of class weighted size. The residues with the highest frequency at these low entropy loop positions are Gly with 30%, Asp with 15%, Pro with 8%, Ser and Thr with 6%, and Ala, Leu, and Phe with 5% (Table 5). Comparing these frequencies to those among all loop positions, Cys is found to be 3.1 times as frequent, Gly 2.1 times, Phe 1.5 times, Asp 1.4 times, and Pro 1.3 times. All other residues are less frequent in loop positions

Fig. 3. **A:** Clustering of EH loops of length three based on their pairwise structural superposition. The 37 loops not belonging to clusters have been removed to improve clarity. **B:** Superposition of some loops of length four from the HagbBE class [generated using the program MOLSCRIPT (Kraulis, 1991)]. All main-chain atoms of the loop region are drawn, while the bounding secondary structures are represented by their C α traces.

of low entropy compared with all loop positions. When the residue frequencies of low entropy loop positions are compared with those among the 40,594 residues from the 223 proteins studied, Cys is found to be 4.0 times as frequent, Gly 3.3 times, Asp 2.4 times, Pro 2.0 times, and Phe 1.3 times. It is interesting that the residues that are more frequent in low entropy loop positions relative to all loop positions are the same as those relative to all residues. Gly is the major contributor to 41 low entropy loop positions, Asp to 23, Pro to 12, Phe to 6, and Cys to 4. These low entropy positions are involved in function, in defining secondary structure termini, or in defining loop conformations.

Functionally important residues are frequent among conserved Cys and Asp positions and include the rubredoxin iron-binding Cys residues (EbbaagbEs, positions 1 and 4), ferredoxin iron-cluster binding Cys residues (Hb/aaagaE, position 1), a loop-contributing Asp to the catalytic triad of bacterial serine proteinases (EbbgbtEs, position 1), pepsin-like aspartic proteinases catalytic Asp (Eb/taababEs, position 1), calmodulin, and parvalbumin-like calcium-binding Asp residues (Ht/baagagbE, positions 1 and 3, and HtaagagE, positions 1 and 3). Some conserved positions, although not themselves directly involved in the function, do belong to functionally important loops; for example, the catalytic Asp-containing loop of the bacterial serine proteinases (EbbgbtEs, position 2) includes a largely conserved *cis*-Pro.

In order to investigate whether low entropy loop positions are involved in helix capping (Richardson & Richardson, 1988; Dasgupta & Bell, 1993), we extended helices to include adjacent residues in α conformation (a) and define the helical N cap as the

residue preceding the helix and the helical C cap as the one following the helix. Five of the 11 Gly-rich positions and four of seven Asp-rich positions are N caps of helices, with the aspartates forming side-chain hydrogen bonds to main-chain NH of residues at $i + 2$, $i + 3$, or both (Richardson & Richardson, 1988; Dasgupta & Bell, 1993). One of the three Pro-rich positions is one residue after a helix N cap. Both Phe-rich loop classes with C-terminal helices have their conserved Phe residues, which are one position prior to the N cap, interacting with hydrophobic residues four positions after the N cap, a pattern known as a hydrophobic-staple (Muñoz et al., 1995). Of the 18 Gly-rich loop classes with N-terminal helices, 13 have their conserved Gly position at the helical C cap, 11 of which are in positive $\phi(g)$ or extended positive $\phi(e)$ conformations (Nagarajaram et al., 1993). Two of the five classes with conserved Pro residues, bound at their N terminus by a helix, have conserved Pro residues one position after the C cap.

Colloc'h and Cohen (1991) have identified preferences for Asp, Lys, Thr, Gly, and Glu at the residue position preceding a β -strand (N-terminal β -breakers), and for Gly, Asp, Pro, Ala, Lys, and Ser at the residue position following a β -strand (C-terminal β -breakers). With β -breakers loosely defined as the non- β positions bounding a β -strand, 17 of the 30 Gly-rich loop classes with C-terminal β -strands have their conserved Gly position as the N-terminal β -breaker, and 10 of the 23 loop classes with N-terminal β -strands have their conserved Gly position as the C-terminal β -breaker. In the case of Pro-rich positions, 5 of 7 loop classes with N-terminal β -strands have their conserved Pro position at the C-terminal β -breaker, while only 1 of 10 loop classes with

Fig. 4. Number of classes as a function of loop length.

C-terminal β -strands has its conserved Pro position at the N-terminal β -breaker, a preference identified by Colloc'h and Cohen (1991). Of the 16 Asp-rich loops with C-terminal β -strands, 9 have their conserved Pro position at the N-terminal β -breaker, and 3 of 8 loop classes with N-terminal β -strands have their conserved Asp position at the C-terminal β -breaker.

The conserved Phe-rich loop positions all have buried side chains. While in five of these the Phe-rich position is within two residues of a secondary structure terminus, pinning down the loop end, in the remaining case the Phe-rich position is at the center of an immunoglobulin heavy chain variable domain hypervariable H1 loop (Egbb/taabbEs, position 4), which it stabilizes by burial of its side chain, as first identified by Tramontano et al. (1989).

Comparison with previously reported loop conformations

To assess whether the initial set of loops is complete and the derived classification comprehensive, the conformational classes were compared with previously described loop conformations. This comparison is rendered somewhat subjective by the different data sets used in the various studies as well as by the variety of secondary structure assignments. Loop conformational classes (Table 2) were found that coincide with type I, I', and II' 2:2 β -hairpins, as well as type I 4:4 and type I 3:5 β -hairpins, described by Sibanda and Thornton (1985). The two $\alpha\beta$ loop classes, of lengths one and two, and the $\beta\alpha$ loop class, of length three, identified by Edwards et al. (1987) from $\beta\alpha\beta$ units with hydrogen bonded β -strands, have corresponding conformational classes, although of slightly different lengths. The present classification also includes the ten $\alpha\alpha$ -turn families described by Wintjens et al. (1996), of which six had been previously reported (Thornton et al., 1988; Efimov, 1991a,b,c; 1993a,b).

Although comparison with the work of Rooman et al. (1990a) is difficult because the fragments they compared were not specifically bound by secondary structures, the present analysis includes

one of the three new fragment conformations reported. Of the 46 loop conformational classes of one to eight residues in length, identified by Efimov (1986; 1991a,b,c; 1993a,b) among the four types of super-secondary motifs, only the $\beta_m\epsilon\beta_n$, $\beta_m\alpha\gamma\epsilon\beta_n$, $\alpha_m\gamma\alpha_L\beta\alpha\beta\beta_m$, and $\alpha_m\gamma\alpha\beta\beta_P\alpha\gamma\beta_n$ classes are not found in the present work. The 15 loop conformational classes identified by Srinivasan et al. (1991) among loops of length one to five are all found in the current classification.

Of the α -turns identified by Nataraj et al. (1995) all but one were found among the loop conformational classes; their minor g4 group corresponded to the two isolated EbbgE loops that do not constitute a conformational class in our analysis.

New loop classes

Of the 161 loop conformational classes in our database, to our knowledge, we report here 111 classes for the first time. However, 46 of these are members of suprafamilies or related to previously identified loop conformations. Of a total of 53 classes in the EE loops, 26 are unique new ones, as are 16 of the 33 EH classes, 16 of the 42 HE classes, and 7 of the 33 HH classes. We discuss now a selection of new classes that can neither be assimilated to reported ones by allowing relaxation of the bounding secondary structures nor belong to poorly populated long loops.

EE-loop classes

EagabE class, which comprises five nonhomologous loops from three different protein families (thermolysin, porin, and carbonic anhydrase), is a reasonably compact class, with low standard deviations for both the intervector separations and angles. Four of the examples are β -hairpins and the other one is a hairpin-like structure. Gly is the only residue allowed at the second position in the loop, while in the first there is a very strong preference for Asn. Those members that are hairpins belong to a novel 2:4 type IV hairpin class in which the side chain of the residue at the first

Table 3. Suprafamilies of loop conformational classes

A. EE					
A1. Hairpin structures					
b[gl]b _{HP} [23]	[gg] _{HP} [65]	b[gg] _{HP} [9]	[aa] _{HP} [32]	[aag] _{HP} [58]	[aaag] _{HP} [43]
^a EbgbbE [2]	^b EggE [62]	^a EbggE [6]	^b EaaE [18]	^b EaagE [6]	^b EaaagE [30]
^b EbbgbbE [15]	^a EbggbE [3]	^b EbbggE [3]	^b EbaabE [12]	^b EbaagE [52]	^b EbaagbE [13]
^b EbbgbbE [6]			^c EbbaabE [1]		
			^c EbbbaabE [1]		
A2. Non-hairpin structures					
[a] _{VS} [41]	[g] _{LS} [13]				
^b EaE [13]	^c EgE [1]				
^b EbaE [17]	^c EbgE [1]				
^c EabE [3]	^b EgbE [4]				
^a EabbE [8]	^b EgbE [4]				
	^c EgbbbE [1]				
	^c EbbbbbE [2]				
B. EH		C. HE			
[ab] _{HP-I} [38]	[g] _{HP-I} [79]	[ba] _{HP-I} [19]	[gba] _{HP-I} [37]		
^b EabH [21]	^b HgE [6]	^b HbaE [5]	^a HgbaE [11]		
^a EabH [11]	^b HagbE [30]	^a HbabE [2]	^a HagbabE [4]		
^a EbbabH [3]	^c HaaagbbE [1]	^a HaaabaE [4]	^b HagbaE [20]		
^a EbabaH [1]	^b HagE [6]	^a HaaaabaE [1]	^c HgbabbE [1]		
^a EbbbabH [1]	^b HgbE [12]	^c HababbE [1]	^c HagbabE [1]		
^c EbabaaaH [1]	^a HgbbE [6]				
	^a HaagbE [3]				
	^b HagbbE [8]				
	^a HgbbbE [2]				
	^c HaaagbE [1]				
	^a HagbbbE [2]				
	^c HaaagbbE [1]				
	^c HagbbbbE [1]				
D. HH					
[b] _{LS} [41]	[g] _{VS} [9]	[t] _{LS} [36]	[bb] _{HP-I} [12]	[gbb] _{HP-I} [21]	[gbb] _{VS} [24]
^b HbH [21]	^c HgH [2]	^b HrH [22]	^b HbbH [8]	^b HagbbH [13]	^a HgbbH [8]
^c HabH [3]	^a HagH [4]	^a HarH [2]	^c HbbaH [1]	^a HgbbbaH [2]	^b HaagbbH [14]
^a HbaH [3]	^a HagaH [3]	^c HraH [1]	^a HaabbH [3]	^b HaagbbaH [4]	^c HaaagbbaH [2]
^a Haab [7]	^c HaagH [1]	^a HaarH [3]		^a HagbbaaaH [2]	
^a HabaaH [3]		^b HraaH [3]			
^c HbaaaH [1]		^a HaaaH [2]			
		^a HaaaH [2]			
		^a HtaaaH [1]			
		^c HaaaaH [2]			

Note: Number of loops populating the suprafamily, conformational class, or isolated observation given within brackets. HP = hairpin structure; HP-I = hairpin-like structure; VS = V-shaped structure; LS = L-shaped structure.

^aDenotes a loop conformation from a class in which alternative conformations in some of the positions have been allowed (e.g., the conformation EbbabbE belongs to the loop conformational class Ebb/a/bb/ebEs).

^bDenotes a loop conformational class.

^cDenotes a loop conformation observed in our database, but not forming a class.

position of the loop (always Asn) is hydrogen-bonded to the main chain of the residue at the forth position in the loop. An example is presented in Figure 6A.

EbaaabE class has 14 members from six different proteins: mammalian Ser proteinases, histocompatibility antigens, Ser proteinase inhibitors, interleukin, glutathione reductase, and bacteriochlorophyll protein a. All but one of the examples are hairpin-like structures. There is no strong preference for any particular residue at any position of the loop.

EbaagabE class comprises eight loops from five different proteins: histocompatibility antigen, ferredoxin, immunoglobulin (variable region), glucoamylase, and glucose permease. Six members are hairpins and the other two are hairpin-like structures. Asn/Asp is highly preferred for the first position, Glu for the second and Gly for the fourth. There is no clear pattern for the other positions in the loop. The side chains of the charged/polar residues in the two first positions are exposed to the solvent and do not interact with other residues.

Table 4. Observed residue percentage frequencies for the EggE (A) and EeaE (B) classes

Pos.	Conf.	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	1/2SS
A.																						
1	g	5.6	0.0	26.1	5.1	0.0	18.0	2.6	0.0	1.3	2.6	0.0	23.7	0.0	5.1	0.0	7.5	0.0	0.0	0.0	2.6	0.0
2	g	0.0	0.0	0.0	1.3	0.0	89.7	0.0	0.0	0.0	0.0	0.0	6.4	0.0	0.0	0.0	2.6	0.0	0.0	0.0	0.0	0.0
B.																						
1	e	0.0	0.0	12.0	4.0	0.0	80.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0
2	a	10.0	0.0	16.0	4.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	30.0	12.0	0.0	0.0	0.0	0.0

Note: Half cystines are indicated as 1/2SS.

EH loop classes

EtH class contains four loops from three different proteins: glucose permease, aspartate amino transferase, and thymidilate synthetase. All are V-shaped motifs and have Asp as the residue in the loop. The side chain hydrogen bonds with the main chain of the third residue in the following helix, capping it as shown in Figure 6B.

Eet/abH class comprises nine loops from three different proteins: glutathione reductase, lactate dehydrogenase, and alcohol dehydrogenase. All members of this class are hairpin-like structures with a narrow distribution of both the intervector separation and angle values. This is the first loop in the $\beta 1$ - αA - $\beta 2$ motif in dinucleotide-binding proteins. Gly is the only residue observed at the first and third positions of the loop, while there is a very strong preference for hydrophobic residues at the second position. This hydrophobic residue fixes the conformation of the whole loop (which by featuring Gly at both the first and third positions could have a good deal of conformational freedom) by maintaining (Fig. 6C) hydrophobic contacts with three other residues. These are the third residue in the following helix (which is usually Leu), the aliphatic part of the side chain of an Asp residue at a position

some 20 residues further down the sequence (which hydrogen bonds to the ribose of NAD/FAD), and a residue that is usually aromatic and on another element of secondary structure.

HE-loop classes

HbgbE class has five loops from four different proteins: isocitrate dehydrogenase, amino binding proteins, β -lactamase, and dehydrofolate reductase; four are L-shaped motifs. Gly is highly preferred in the second position of the loop and Lys in the third position, which is exposed to the solvent.

Haga/tE class has five independent loops from four different proteins: adenylate kinase, phosphofructose, amino acid transferase, and ferredoxin reductase. There is a striking variation in both the intervector separation and angle values, and examples of hairpin-like, V-shaped, and L-shaped motifs are found among the members of this class. Gly is strongly preferred in the first position of the loop.

Hgbt/bbE class comprises four nonhomologous loops (five in total) from four different proteins: rubisco, CD4, glutathione reductase family, and xylose isomerase. All are hairpin-like structures with a low standard deviation for both the intervector angle

Fig. 5. Entropy distribution of loop class positions as a function of the sum of the member loop weights. A curve approximating average entropy (values of which were estimated as described in Materials and methods) was built using average values only, derived from at least ten loop positions.

Table 5. Frequencies and frequency ratios of the 21 amino-acid type in low entropy loop positions, loop positions and whole proteins

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET
Low entropy	4.87	2.69	14.9	1.34	4.84	30.3	0.44	1.41	2.96	4.61	0.00
All loop positions	7.06	0.86	10.49	5.09	3.18	14.27	2.27	2.38	6.44	5.32	0.86
All proteins	8.96	0.75	6.30	5.63	3.81	8.93	2.12	4.97	6.34	7.70	1.87
Low entropy/all loops	0.69	3.13	1.42	0.26	1.52	2.12	0.19	0.59	0.46	0.87	0.00
Low entropy /all	0.54	3.95	2.37	0.24	1.27	3.39	0.21	0.28	0.47	0.60	0.00

	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR	1/2 SS
Low entropy	3.81	8.33	0.30	0.85	6.39	5.80	3.88	0.02	2.17	0.00
All loop positions	5.86	6.39	3.26	3.04	9.23	5.86	4.26	0.78	2.44	0.55
All proteins	4.28	4.28	3.54	3.86	7.07	5.95	7.25	1.45	3.68	1.27
Low entropy/all loops	0.65	1.30	0.09	0.28	0.69	0.99	0.91	0.03	0.89	0.00
Low entropy /all	0.89	1.95	0.08	0.22	0.90	0.97	0.54	0.01	0.59	0.00

Note: Half cystines are indicated as 1/2SS.

Fig. 6. MOLSCRIPT drawings of representatives of the (A) EagabE, (B) EtH, (C) Eet/abH, and (D) Ha/gbaa/bH conformational classes. (Figure continues on facing page.)

values and the separation of the termini. Gly is always found at the first position in the loop.

HH-loop classes

Ha/gbaa/bH class contains six nonhomologous loops from four different proteins: globins, lysozyme, phosphoribosyl-anthranilate isomerase, and thymidylate synthase. The standard deviation of the intervector separation and angle values of the members of this class is high, reflecting the existence of both hairpin-like structures and L-shaped motifs (three examples of each). Charged/polar residues are preferred at the second and fourth positions in the loop, where the sidechains of these residues are usually exposed to the solvent but sometimes involved in capping the preceding or succeeding helices. An example of the latter is shown in Figure 6D.

Hagbb/a/tb/tb/aH class comprises seven members from four different proteins: calcium-binding protein, adenylate kinase, thymidylate synthase, and cytochrome c. Two types of geometric arrangements are observed in the members of this class, hairpin-like and L-shaped motifs. Gly is the only residue allowed at the second position of the loop and hydrophobic residues, which are buried in a hydrophobic environment, are very much preferred at the fifth and sixth.

Loop classes and biological function

Loop regions often form binding sites for ligands or contribute to enzyme active sites. For example, the GTP-binding site in the GTP-binding proteins is made up by five of the six loops connecting strands of the central β -sheet with the α -helices in the α/β structure; and the active site residues in mammalian Ser proteinases are situated on loops provided by the two antiparallel β -barrel domains. We now ask whether loops playing similar functional roles belong to the same conformational class, and if so, whether they are the only members populating the class. Functionally specific loop classes could provide valuable clues in the study of proteins of unknown function.

The loop in the helix-turn-helix motif in DNA-binding proteins

The DNA-binding helix-turn-helix (HTH) motif (Pabo & Sauer, 1984; Brennan & Matthews, 1989; Steiz, 1990; Harrison & Aggarwal, 1990) interacts with DNA by insertion of its C-terminal helix or recognition helix into the DNA major groove (Harrison, 1991), while its loop, three residues in length, is essential for maintaining the 120-degree angle between the helices. The HTH

Fig. 6. Continued.

motif has been identified in prokaryotic repressors and activators, as well as in the eukaryotic homeobox domain. HTH-like motifs have been reported in cytochrome *c* peroxidases and ribosomal L7/L12 protein (Richardson & Richardson, 1988).

The five HTH motifs, from the five HTH DNA-binding proteins in the initial database, belong to the two related loop classes Hg/abbH and HagbbH. Four of the eight HgbbH loops are DNA-binding HTH motifs from 1LRD, 1R69, 2CRO, and 2WRP, and four do not bind DNA, including one from L7/L12 ribosomal protein. The HagbbH group contains the last DNA-binding HTH motifs from 3GAP, but also four HTH-like motifs from 2CRO and 1R69 (Yura et al., 1993) and eight loops from other families that do not bind DNA. Because neither of the two HTH containing loop conformational classes is restricted to DNA-binding loops, identifying or predicting a loop as belonging to either of them does not necessarily indicate DNA-binding activity. As suggested by Wintjens et al. (1996), who also found a mixed α GBB α class, the HTH motif may require a very precise shape achieved by the interactions with the third helix in the bundle, a feature specific to the DNA-binding proteins.

The loop in the calcium-binding motifs

The helix-loop-helix calcium-binding motifs of parvalbumin, calmodulin, troponin C, and other proteins that regulate cellular activities through the binding of calcium, contain a 12 residue long loop that binds calcium by means of Asp and Glu carboxyl side chains and main-chain carbonyls (Kretsinger, 1980). In fact, a short intervening β -strand is assigned by the Kabsch and Sander (1983) procedure, implying a calcium-binding HE loop of six or seven residues in length. All the calcium-binding loops, from eight proteins, in the initial database belong to the HtaagagE and Hb/taagagE loop conformational classes, which contain no other loops. These two classes have a specific main-chain conformation as well as specific side chains that make them unique to calcium binding. However, at present all the members in these two classes belong to proteins of similar folds.

Catalytic loops in pepsin-like aspartic proteinases

The catalytic Asp aspartic residues of aspartic proteinases are located on loops of two Greek ψ -like motifs, each contributed by a domain related by a pseudo dyad axis in the pepsin-like enzymes and a perfect dyad between subunits in the retroviral proteinases. The regions around the catalytic Asp residues have characteristic sequences: (hydrophobic, generally Phe)-Asp 32-Thr-Gly-Ser in the N-terminal domain and (hydrophobic)-Asp 215-Thr-Gly-Ser/Thr (pepsin numbering). Our protein database contained five pepsin-like aspartic proteinases: rhizopuspepsin (2APR), penicillopepsin (3APP), endothiapepsin (4APE), chymosin (4CMS), and pepsin (5PEP). The Asp 215 loop of all five aspartic proteinases belongs to class Eb/taababEs, where Asp 215 is the first residue in the loop, and the Asp 35 loop, of all the enzymes but 2APR, in the related Eaa-babEs class, where Asp 35 takes up the last position of the first strand. The Asp 35 loop of 2APR belongs to the closely related conformation EaababE, which appears only once in our database. Both Eaa-babE and Et/baababE are unique to pepsin-like aspartic proteinases.

Others

We have observed several other examples of loop conformations uniquely associated with a biological function. These include the

iron-binding loops of the rubredoxins, which all belong to the EbbaagbEs class. (Interestingly, the EbbaagbE class includes nine loops from seven different proteins, none of which are known to bind iron, and lack the Cys residues needed to coordinate Fe atoms.) The loops of iron-sulfur clusters in the ferredoxins belong to class HbaagaE; the Asp 102-bearing loops of bacterial Ser proteinases belong to the EbbgbEs class; and the loops of the β 1 α β 2 motif in the dinucleotide-binding proteins belong to the Eet/abH class. Thus, longer loops with a specific conformation are more likely to be indicative of function.

Application on protein structure modelling

An approach to loop selection in knowledge-based modelling procedures has been developed based on the identification of preferred loop conformational classes. This approach is the subject of a paper (Rufino et al., 1996) to be published elsewhere. In summary, during the modeling process, the information available for the prediction of the conformation of a loop comprises its sequence and the spatial disposition of its bounding secondary structures as defined by the structurally conserved regions of the model. To predict the conformation of a loop, its motif is compared with the templates of each identified loop class of matching length and compatible bounding secondary structure types, and each comparison is evaluated in terms of both sequence match and bounding secondary structure vectors fit. This approach differs from most other loop modeling procedures in that it predicts the class of a loop rather than identifying a specific loop. Classes contain much more information about sequence preferences and geometric plasticity than do isolated loops.

Materials and methods

Protein structure database

We selected 223 protein and protein domain structures from the Brookhaven Protein Data Bank (Bernstein et al., 1977), including 182 structures from a database of 66 protein families (Overington et al., 1990a), structurally aligned using the MNYFIT (Sutcliffe et al., 1987) and COMPARE (Šali & Blundell, 1990) programs, and 41 unique structures (Sowdhamini & Blundell, 1995). All of the structures were of resolution 2.7 Å or better and had at least 65% of their residues with $\phi\psi$ within the highly populated regions of the $\phi\psi$ plot as defined by PROCHECK (Morris et al., 1992).

Identification of secondary structure elements

Secondary structure elements were defined as continuous fragments of identical secondary structure (Kabsch & Sander, 1983). 3_{10} -helices containing only three residues, which often play a role in polypeptide chain reversal or redirection, were not considered as secondary structure elements. Although 3_{10} -, α -, and π -helices were grouped in a general helical class, contiguous 3_{10} -, α -, or π -helices were not merged.

Identification of loop motifs

A loop motif was defined to include both the region of nonrepetitive conformation or loop and the pair of bounding secondary structures. To avoid erroneous classification of β -hairpin motifs due to hydrogen bonding with β -strands outside the motif, the proximal termini (i.e., those bordering the central loop) of β -hairpin β -strands were determined using only the hydrogen bonding patterns within the β -hairpin itself.

Associated vectors

For each loop motif two vectors were determined, representing the directions and positions of the proximal termini of its two secondary structures. Vectors associated with β -strands, 3_{10} -, α -, and π -helices were calculated using all of their residues up to a maximum of 6, 7, 8, or 9 residues, respectively, proximal to the loop. Where possible, an extra distal residue was used for determining the vector associated with a β -strand of only two residues. Vectors associated with secondary structures of four or more residues were defined so as to minimize the sum of the squares of the distances from the vector to the $C\alpha$ atoms of the corresponding secondary structure (Chou et al., 1984; Srinivasan et al., 1991; Sowdhamini et al., 1992). In the case of a β -strand of three residues or a β -strand of two residues extended to three, a vector was fitted to the three $C\alpha$ atoms with each of the two terminal $C\alpha$ atoms, contributing one-half the weight of the central $C\alpha$ atom. The vector associated with a β -strand containing only two residues was defined as the virtual bond linking its two $C\alpha$ atoms.

For each loop motif, the intervector separation and angle were calculated. The separation between the two vectors of a motif was defined as the distance between the projections on the vectors of the proximal $C\alpha$ atoms of their respective secondary structures. The angle between two vectors was calculated as the angle between their directions so that an angle had a value of zero if the vectors pointed in the same direction.

Main-chain conformation

The main-chain conformation of the residues in the loop region of a motif was originally determined and described using a reduced notation of seven residue conformational classes (Topham et al., 1993): α -helical (a), antiparallel β -sheet (b), other β -sheet (p), transition between α -helical and β -sheet (t), left handed helix (l), positive ϕ (g), and positive ϕ in an extended conformation (e). Later, these seven classes were reduced to five by the merging of p and b and the g and l conformational classes into the larger classes b and g, respectively.

Superposition of loops

Loops were superposed (Kearsley, 1989) using their carbonyl oxygen and carbon atoms, their $C\alpha$ atoms, and their amide nitrogen and theoretical hydrogen atoms. So as not to overemphasize the fitting of the main-chain atoms versus the $\phi\psi$ angles, the carbonyl oxygen and the amide nitrogen were given weights of 1.5, while all other atoms were given weights of 1.

Clustering of loops

All loops of compatible length and bounding secondary structure type were superposed (Kearsley, 1989) pairwise and a distance measure was derived from the RMS deviation (RMSD) of individual pairwise superpositions:

$$D = \ln(1 + \text{RMSD}),$$

with values ranging from zero for identical loops to approximately $\ln(\text{RMSD})$ for very different loops.

Loops were then clustered using the hierarchical clustering program KITSCH (Felsenstein, 1985), which selects the dendrogram that minimizes the differences with calculated distances (Fitch & Margoliash, 1967).

Description of loop conformational classes

A loop class, identified from clustering, was described by the main-chain conformations, sequences, and intervector separation and angle of all the loop members of the class. Sequence information was described in terms of observed residue frequencies at each loop position and intervector relations in terms of average value and standard deviation. Twenty-one residue types were considered, the 20 amino acids encoded by the genetic code with the addition of the half-cystine, an oxidized Cys residue involved in a disulfide bridge. The contribution of a loop motif to the description of a loop class was weighted by the inverse of the number of its homologous member loops. The main-chain conformation of a loop class was reflected in the class's name by indicating for each loop position the observed main-chain conformations. Alternate main-chain conformations were separated by a backslash.

Entropy of amino acid distribution at conformational class loop positions

The information content of the amino acid distribution at a loop position was approximately estimated by its entropy, H , defined as:

$$H = - \sum_{i=1}^n f_i * \ln(f_i),$$

where f_i is the observed frequency of the amino acid of type i . The entropy of a permissive loop position with equivalent frequencies for all amino acid types will be:

$$H = - \sum_{i=1}^n \frac{1}{n} * \ln\left(\frac{1}{n}\right) = -\ln\left(\frac{1}{n}\right) = \ln(n).$$

Considering the 23 amino acid types, including half-cystine, the entropy of a loop position with equal amino acid frequencies is:

$$\ln(23) = 3.14.$$

Because

$$0 * \ln(0) = 0 \quad \text{and} \quad 1 * \ln(1) = 0,$$

the entropy of a loop position accepting a unique amino acid type is zero. The normalized entropy was defined as

$$E = H/\ln(23),$$

with a range of zero to one. We chose not to use actual residue frequencies in the determination of entropies, so that the conserved positions would be equally scored whatever the conserved residue and interesting conservations at frequently occurring residues would be not missed.

Acknowledgments

We would like to thank Dr. N. Srinivasan for his vector determination program SCHELAX, Drs. N. Srinivasan and N.H. Yennawar for their superposition program SUPER, and Dr. Zhan-Yang Zhu for his program DRAWTREE. We thank all colleagues at the ICRF Unit of Structural Molecular Biology and Birkbeck College for their kind help. L.H.J.C., who was fulfilling his National Service duties, thanks the staff at the French Embassy in London for their support. L.E.D. was a recipient of a Human Capital and Mobility fellowship from the European Union and of a P.F.P.I. grant from the Spanish Ministerio de Educación y Ciencia. S.D.R. was a

recipient of an Imperial Cancer Research Fund bursary, and L.H.J.C. was funded by Dorian.

References

- Bernstein FC, Koetzle TF, William GJB, Meyer EF, Brice MD, Rodgers JR, 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.
- Blundell T, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T, Overington J, Singh OA, Sibanda BL, Sutcliffe M. 1988. Knowledge-based protein modelling and design. *Eur J Biochem* 172:513–520.
- Brennan RG, Matthews BW. 1989. The helix-turn-helix DNA-binding motif. *J Biol Chem* 264:1903–1906.
- Chan EAW, Hutchinson EG, Harris D, Thornton JM. 1993. Identification, classification and analysis of β -bulges in proteins. *Protein Science* 2:1574–1590.
- Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlam EA, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ. 1989. Conformations of immunoglobulin hypervariable regions. *Nature* 342:877–883.
- Chou K-C, Nemethy G, Scheraga HA. 1984. Energetic approach to the packing of α -helices. 2. General treatment of nonequivalent and nonregular helices. *J Am Chem Soc* 106:3161–3170.
- Collopy H N, Cohen FE. 1991. β -Breakers: An aperiodic secondary structure. *J Mol Biol* 221:603–613.
- Dasgupta S, Bell JA. 1993. Design of helix ends. *Int J Peptide Res* 41:499–511.
- Edwards M, Sternberg M, Thornton JM. 1987. Structural and sequence patterns in the loops of $\beta\alpha\beta$ units. *Protein Engng* 1:173–181.
- Efimov AV. 1986. Standard conformations of polypeptide chain in irregular regions of proteins. *Mol Biol (Moscow, USSR)* 20:250–260.
- Efimov AV. 1991a. Structure of α - α -hairpins with short connections. *Protein Engng* 4:245–250.
- Efimov AV. 1991b. Structure of coiled β - β -hairpins and β - β -corners. *FEBS Letters* 284:288–292.
- Efimov AV. 1991c. Long and medium-sized irregular regions in proteins as combinations of small standard structures. In: Ramakrishnan C, Balaram P, eds., *Molecular conformation and biological interactions*. Bangalore: Indian Academy of Science. pp 19–29.
- Efimov AV. 1993a. Patterns of loop regions in proteins. *Curr Opin Struc Biol* 3:379–384.
- Efimov AV. 1993b. Standard structures in proteins. *Prog Biophys Mol Biol* 60:201–239.
- Engl RA, Huber R. 1991. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst A* 47:392–400.
- Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.
- Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155:279–284.
- Greer J. 1980. Model for haptoglobin heavy chain based upon structural homology. *Proc Natl Acad Sci USA* 77:3393–3397.
- Greer J. 1990a. Comparative modelling methods: Application to the family of the mammalian Ser proteinases. *Prot Struct Funct and Genes* 7:317–334.
- Greer J. 1990b. Comparative modelling of proteins in the design of novel renin inhibitors. *Biophysical J* 57:p.A.207.
- Harrison SC. 1991. A structural taxonomy of DNA-binding domains. *Nature* 353:715–719.
- Harrison SC, Aggarwal H. 1990. DNA-recognition by proteins with the helix-turn-helix motif. *Ann Rev Biochem* 59:933–969.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
- Kearsley SK. 1989. On the orthogonal transformation used for structural comparisons. *Acta Cryst A* 45:208–210.
- Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Applied Cryst* 24:946–950.
- Kretsinger RH. 1980. Structure and evolution of calcium-modulated proteins. *CRC Crit Rev Biochem* 8:119–174.
- Leszczynski JF, Rose GD. 1986. Loops in globular proteins: A novel category of secondary structure. *Science* 234:849–855.
- Martin ACR, Toda K, Stirk HJ, Thornton JM. 1995. Long loops in proteins. *Protein Engng* 8:1093–1101.
- Milner-White EJ. 1990. Situations of γ -turns in proteins. Their relation to α -helices, β -sheets and ligand binding sites. *J Mol Biol* 216:385–397.
- Milner-White EJ, Ross BM, Ismail R, Belhadj-Mastefa K, Poet R. 1988. One type of γ -turn, rather than the other, gives rise to chain reversal in proteins. *J Mol Biol* 216:385–397.
- Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein structure coordinates. *Proteins* 12:345–364.
- Muñoz V, Blanco FJ, Serrano L. 1995. The hydrophobic-staple motif and a role for loop-residues in α -helix stability and protein folding. *Nature Struc Biol* 2:380–385.
- Nagarajaram HA, Sowdhamini R, Ramakrishnan C, Balaram P. 1993. Termination of the right handed helices in proteins by residues in left handed helical conformations. *FEBS Letters* 321:79–83.
- Nataraj DV, Srinivasan N, Sowdhamini R, Ramakrishnan C. 1995. α -Turns in protein structures. *Curr Sci* 5:434–447.
- Overington J, Johnson MS, Sali A, Blundell TL. 1990a. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc R Soc Lond B* 241:132–145.
- Overington J, Johnson MS, Topham C, McLeod A, Sali A, Zhu Z-Y, Sibanda BL, Blundell TL. 1990b. Applications of environment specific amino acid substitution tables to identification of key residues in protein tertiary structure. *Curr Sci* 59:867–874.
- Pabo CO, Sauer RT. 1984. Protein-DNA recognition. *Ann Rev Biochem* 53:293–321.
- Rice PA, Goldman A, Steitz TA. 1990. A helix-turn-strand structural motif common in α - β proteins. *Proteins* 8:334–340.
- Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of α -helices. *Science* 240:1648–1652.
- Ring CS, Kneller DG, Langridge R, Cohen FE. 1992. Taxonomy and conformational analysis of loops in proteins. *J Mol Biol* 224:685–699.
- Rooman MJ, Rodriguez J, Wodak SJ. 1990a. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol* 213:327–336.
- Rooman MJ, Rodriguez J, Wodak SJ. 1990b. Relations between protein sequence and structure and their significance. *J Mol Biol* 213:337–350.
- Rose GD, Gierasch LM, Smith JA. 1985. Turns in peptides and proteins. *Advan in Prot Chem* 37:1–109.
- Rufino SD, Donate LE, Canard LHJ, Blundell TL. 1996. Predicting the conformational class of short and medium size loops connecting regular secondary structures: Application to comparative modelling. *J Mol Biol*. In press.
- Sali A, Blundell TL. 1990. Definition of general topological equivalence in protein structures. *J Mol Biol* 212:403–428.
- Sibanda BL, Blundell TL, Thornton JM. 1989. Conformation of β -hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol* 206:759–777.
- Sibanda BL, Thornton JM. 1985. β -Hairpin families in globular proteins. *Nature* 316:170–174.
- Sowdhamini R, Blundell TL. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci* 4:506–520.
- Sowdhamini R, Srinivasan N, Ramakrishnan C, Balaram P. 1992. Orthogonal $\beta\beta$ motifs in proteins. *J Mol Biol* 223:845–851.
- Srinivasan N, Sowdhamini R, Ramakrishnan C, Balaram P. 1991. Analysis of short loops connecting secondary structural elements in proteins. In: Ramakrishnan C, Balaram P, eds., *Molecular conformation and biological interactions*. Bangalore: Indian Academy of Science. pp. 59–73.
- Steitz TA. 1990. Structural studies of protein-nucleic acid interaction: The sources of sequence-specific binding. *Q Rev Biophys* 23:205–280.
- Sun Z, Blundell TL. 1995. The pattern of common supersecondary structure (motifs) in protein database. Proceedings of the 28 Annual Hawaii International Conference on System Sciences.
- Sutcliffe MJ, Haneef I, Carney D, Blundell TL. 1987. Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engng* 1:377–384.
- Swindells MB. 1994. Loopy similarities. *Struc Biol* 1:421–422.
- Swindells MB, MacArthur MW, Thornton JM. 1995. Intrinsic ϕ , ψ propensities of amino acids, derived from the coil regions of known structures. *Nature Struc Biol* 2:596–603.
- Thornton JM, Sibanda BL, Edwards MS, Barlow DJ. 1988. Analysis, design and modification of loop regions in proteins. *Bioessays* 8:63–69.
- Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL. 1993. Identification of key residues in structurally variable regions of proteins using conformationally-constrained environmental substitution tables: Applications to loop fragment ranking in modelling of protein structure. *J Mol Biol* 229:194–220.
- Tramontano A, Chothia C, Lesk AM. 1989. Structural determinant of the conformations of medium-sized loops in proteins. *Proteins* 6:382–394.
- Wintjens RT, Rooman MJ, Wodak SJ. 1996. Automatic classification and analysis of α -turn motifs in proteins. *J Mol Biol* 255:235–253.
- Yura K, Tomoda S, Go M. 1993. Repeat of a helix-turn-helix module in DNA-binding proteins. *Protein Engng* 6:621–628.