Linkers of secondary structures in proteins

V. GEETHA AND P.J. MUNSON

Analytical Biostatistics Section, Laboratory of Structural Biology, Division of Computer Research & Technology,
Bethesda, Maryland 20892-5626

(RECEIVED March 12, 1997; ACCEPTED August 8, 1997)

Abstract

Linkers that connect repeating secondary structures fall into conformational classes based on distance and main-chain torsion clustering. A data set of 300 unique protein chains with low pairwise sequence identity was clustered into only a few groups representing the preferred motifs. The linkers of two to eight residues for the nonredundant data set are designated H-L_n-H, H-L_n-E, E-L_n-H, E-L_n-E, where *n* is the length, H stands for α -helices, and E for β -strands. Most of the clusters identified here corroborate earlier findings. However, 19 new clusters are identified in this paper, with many of them having seven and eight residue linkers. In our first analysis, the secondary structures flanking the linkers are both interacting and noninteracting and there is no precise angle of orientation between them. A second analysis was performed on a set of proteins with restricted orientations for the flanking elements, namely, mainly α class of proteins with orthogonal architecture. Two definite clusters are identified, one corresponding to linkers of orthogonal helices and the other to linkers of antiparallel helices. Loops forming binding sites or involved in catalytic activity are important determinants of the function of proteins. Although the structural conservation of the residues around the catalytic triad of serine proteases has been studied widely, there has not been a systematic analysis of the conformation of the loops that contain them. Residues of the catalytic triad reside in the linkers of β -strands, with varying lengths of more than eight residues. Here, we analyze the structural conservation of such linkers by superposition, and observe a conserved structural feature of the linkers incorporating each of the three residues of the catalytic triad.

Keywords: catalytic triad of serine proteases; clustering; conformational analysis; linkers, orthogonal mainly α proteins

Linkers are loops connecting secondary structural elements in proteins. Previous studies have demonstrated the regularity of certain loops and the possible structural motifs they adopt, contradicting the notion of “randomness” in the loop regions (Efimov, 1986, 1991, 1993). Such structural motifs leading to supersecondary structural features were described by Thornton and others (Sibanda & Thornton, 1985; Milner-White & Poet, 1986; Thornton et al., 1988; Rice et al., 1990; Rooman et al., 1990; Sibanda & Thornton, 1991; Ring et al., 1992; Scheerlinck et al., 1992; Tramontano & Lesk, 1992; Wintjens et al., 1996). Loops are important for protein structure and function, as understood from the EF-hands of calcium-binding proteins, the helix-turn-helix motif of repressor proteins, and the antigen binding sites of immunoglobins. Linkers between two functional domains in proteins have long been implicated in determining the orientation of these domains (Wong et al., 1986; Govindaraj & Poulos, 1995). The catalytic and cellulose-binding domains in endoglucanase A from *Cellulomonas fimi* are connected by a 23-residue linker, which plays a crucial role in maintaining the separation and relative orientation of the two domains necessary for normal function (Wong et al., 1986). A 16-residue

linker connecting the reductase and heme domains in Cytochrome P₄₅₀ is responsible for precise orientation of these two domains required for efficient electron transfer (Govindaraj & Poulos, 1995).

Understanding the role of linkers in connecting the secondary structures is a problem that has intrigued many others. Earlier workers have discussed the existence of structural motifs in loop regions (Sibanda & Thornton, 1985; Thornton et al., 1988; Rooman et al., 1990; Efimov, 1991, 1993; Wintjens et al., 1996). Folding units made up of adjacent secondary structural elements along the chain have been found to associate rapidly to form compact structures (Efimov, 1993). With the growing number of protein structures now available, it is fruitful to reinvestigate the classification of linker regions in proteins. Our preliminary results on the short linkers (Geetha & Munson, 1997) showed the existence of structural motifs and corroborated earlier findings. We now present the results of the clustering of the entire database of protein linkers. A similar classification of loops has been published recently (Donate et al., 1996), but in comparison, proteins analyzed here are selected to have a better resolution. Description of linker motifs clearly depends upon the method of classification of the secondary structural elements and their termini. A recent taxonomy of loops (Kwasigroch et al., 1996) has considered the end-to-end distances and an axial descriptor using virtual positions to identify families of clusters. That method combines linkers of same length irrespective of

Reprint requests to: V. Geetha, ABS/LSB, Bldg. 12A/Rm. 2041, DCRT, NIH, Bethesda, Maryland 20892-5626; e-mail: geetha@helix.nih.gov.

the flanking elements. We feel that the nature of the flanking elements is important for understanding the linker regions. The present study of linkers of secondary structures seeks an understanding of the nonregular regions through a systematic classification scheme. Like other methods, ours also depends heavily upon the secondary structural assignment method. Here, instead of a consensus assignment, we chose the STRIDE program (Frishman & Argos, 1995) for secondary structural assignment of the protein chain. The authors of STRIDE have presented convincing evidence for its relatively superior performance, compared to DSSP (Kabsch & Sander, 1983), which assigns an excess of short helices outside the core secondary structure. Our analysis of linkers should pave the way for improved loop modeling procedures and even aid in their prediction and design.

Results and discussion

The nonredundant data set used here is comprised of high-resolution proteins with $\leq 45\%$ pairwise sequence identity (Materials and methods, Data set I). In earlier studies of loops and turns, pairwise sequence identity of as much as 50% was allowed (Sibanda & Thornton, 1985; Cohen et al., 1993; Chelvanayagam et al., 1994; Mattos et al., 1994; Kwasigroch et al., 1996). Various linkers of size ranging from two to eight residues (L_n , $n = 2-8$), connecting different secondary structures, are clustered here (see the Materials and methods). There are about 3,133 total linkers (Fig. 1), among which the E- L_n -E linkers comprise a significant proportion. The ϕ, ψ conformation of all the linker residues irrespective of the

flanking elements has a relatively large population of beta-parallel (Fig. 2, region P) and left-handed alpha-helix (Fig. 2, region L). This distribution was found to contribute to the irregularity of the linkers (Kwasigroch et al., 1996). Antiparallel sheets (Fig. 2, region B) are present at much lower density compared to their occurrence within the secondary structural elements of proteins.

There are four types of linkers of length two, depending on the adjacent secondary structures (Table 1). Two-residue linkers between helices, H-L2-H, have two distinct clusters. The first has PP conformation, with the first residue (denoted L_1) mostly hydrophobic and the second residue (denoted L_2) one of Ser/Thr/Pro. This grouping of these three residues is based on an earlier observation that Ser and Thr could mimic Pro at certain positions (Wilmot & Thornton, 1988). The second cluster shows preference for Gly/Asn at L_1 and Ser/Thr at L_2 . A single major cluster for H-L2-E linkers shows preference for LP conformation at the L_1 and L_2 positions, respectively. L_1 is Gly/Asn and L_2 is hydrophobic, with a g^+ side-chain conformation for both L_1 and L_2 . The E-L2-H linker with two distinct clusters has BP conformation for one and AB for the other. We find that the E-L2-E linkers form four distinct clusters in the α -carbon distance space corresponding to the hydrogen bonded turn types I, I', II, and II'. It is interesting to note the presence of a distinct cluster (PL) for turn type II. The β -strands that straddle linkers in this cluster might either be hydrogen bonded to form a well-defined β -hairpin, or might not interact. We have made no attempt to delineate the interacting and noninteracting strands and hence it is hard to conclude if the cluster corresponding to turn type II is an allowed conformation for β -hairpins, which would contradict an earlier report (Thornton et al., 1988).

Fig. 1. Distribution of linker size for linkers connecting helices (H) and strands (E), from data set I (see Materials and methods).

Fig. 2. ϕ, ψ Torsional angle distribution for all the linkers irrespective of the flanking elements, represented as contours. The ϕ, ψ regions identified here are right-handed helix (α_R or A), left-handed helix (α_L or L), beta-extended (β_E or B), beta-parallel (β_P or P), bridge region between α -helix and β -sheet region (Y, named here for convenience), and epsilon (ϵ or E), occupied mostly by Gly (Efimov, 1993). Note the relatively large population of β_P (P), found to contribute to the irregularity of the linkers.

Three-residue linkers of type H-L3-H and E-L3-E were found previously to form distinct structural motifs (Wintjens et al., 1996). Here, the H-L3-H linkers clustered into four different structural motifs. The first cluster with an LBB backbone conformation prefers Gly at the first position of the linker and hydrophobic residues at the second. The second cluster with ABB main-chain conformation is distinct from the first. Our first two clusters seem to subdivide a single cluster (L/ABB) reported earlier (Donate et al., 1996). The second cluster does not prefer Gly at the first position, but is mostly hydrophobic at the second position with the side chain mostly in g^+ conformation and the third residue preferring a g^- side-chain conformation. The H-L3-E linkers fall largely into a single cluster with the structural motif LPA, having position L_1 mostly Gly and L_2 hydrophobic. The E-L3-E linker has motif AAL with residue L_3 mostly Gly (Fig. 3). The side chain of position L_2 prefers g^+ conformation. Most of the clusters identified here correspond to various loop structural motifs (Sibanda & Thornton, 1985; Thornton et al., 1988; Rooman et al., 1990; Efimov, 1991, 1993; Srinivasan et al., 1991). We present the entire results of our classification of linkers in order to corroborate earlier results (Donate et al., 1996) and to validate our clustering technique.

Linkers of length greater than three residues

Many linkers comprised of four residues connecting two helices, H-L4-H, fall into a cluster with a defined structural motif. In this cluster, the first two residues have nonspecific main-chain torsions and the last two residues have an extended backbone conformation with L_3 mostly occupied by hydrophobic residues. For the H-L4-E linker, one of the clusters shows a very strong preference for Pro at the L_2 position, which constrains the ϕ angle to be negative,

whereas L_3 prefers Asp. There is no sequence preference at any of the four positions of the linker region in E-L4-H. The cluster of E-L4-H linker is similar to one of the clusters of the H-L4-E linker with the difference that there is a strong preference of Pro at the second position of the linker residue in H-L4-E. The E-L4-E linker has four clusters (Table 1). One of them, with a AAL/PP, has not yet been identified and appears to be a new finding of the present analysis. Our observations so far are quite consistent with the similar identification of clusters in most of the cases observed, even though some additional clusters have been identified for the short linkers (Donate et al., 1996).

The five-residue linkers between helices, H-L5-H, fall into three major clusters; one of them, with conformation LBBBB, has a strong preference for Gly at the first position (Fig. 4), which might act as a helix breaker. L_3 and L_4 show preference for hydrophobic residues, whereas L_5 prefers Thr/Pro. The second cluster has a folded turn conformation with i to $i + 3$ hydrogen bond in the linker region and the third cluster has a 3_{10} -helical conformation in the linker region. The second and third clusters show no residue preference. The extended linker conformation allows the central aromatic hydrophobic residues to point inward into the gap between the flanking helices. The second cluster shows no residue preference.

The linker H-L5-E shows two clusters. L_2 is found to prefer Pro in the case of the first cluster and is mostly hydrophobic for the second. The cluster with BBB at L_2 , L_3 , L_4 is found to be a new cluster not identified earlier. The E-L5-H linker shows a specific preference for B/PAA conformation for L_2 , L_3 , L_4 , respectively. The amino acid residues are not specific in any of the positions of this linker. This cluster is also a novel finding of the present analysis. The E-L5-E linker has a single large cluster with BAALB as the preferred motif. The third residue shows preference for Asp and the fourth is mostly Gly. It is quite interesting to note that the five-residue linker with the first and the fifth residues in extended conformation is quite similar to the E-L3-E linker. If the first and last residues of E-L5-E linker are part of the flanking strands, then this linker motif would be similar to E-L3-E. As we stated earlier, the linker lengths strongly depend on the structural assignment, especially at the termini. Other authors combined such loop motifs into "supra families" (Donate et al., 1996).

The hexapeptide linking two helices H-L6-H is found to contain two major clusters, which have not been described previously (Donate et al., 1996; Kwasigroch et al., 1996). In the first cluster, the fourth residue of the linker has a preference for Gly and the last residue has a preference for Asp and Ser/Thr/Pro. Novel motifs, one each for H-L6-E, E-L6-H, and two new clusters for E-L6-E linker, have also emerged from the present study.

The seven-residue linker between helices (Table 1) is found to have a cluster that agrees substantially with the earlier observation. The H-L7-E and E-L7-H linkers each are found to have a major cluster representing a novel motif. There are no specific sequence preferences in any of the positions of these linker residues.

Seven novel clusters are found for the eight-residue linkers with the number of examples between four and nine for each of those clusters. Our results (Table 1) thus extend the loop motifs reported previously (Donate et al., 1996). In summary, we have found a smaller number of more densely populated clusters, compared to earlier work. Also, we have found 19 new clusters through the present analysis.

The proteins in our data set have as much as 45% pairwise sequence identity, raising the possibility that some conformations

Table 1. Linkers and their main-chain conformations

Linker	Cluster number	Number of examples	L ₁ ^a	L ₂	L ₃	L ₄	L ₅	L ₆	L ₇	L ₈	Comments ^b
H-L2-H	1	21	P	P							<i>bb</i>
	2	28	ns	B							<i>a/gb</i>
H-L2-E	1	84	L	P							<i>gb</i>
E-L2-H	1	27	B	P							<i>bb</i>
	2	37	A	B							<i>ab</i>
E-L2-E	1	66	A	A							<i>aa</i>
	2	110	L	L							<i>gg</i>
	3	20	P	L							New
	4	68	E	A							<i>ea</i>
H-L3-H	1	23	L	B/P	P						<i>g/abb</i>
	2	6	A	B	B						Combined in 1
	3	6	B	B/P	B						<i>bbb</i>
	4	16	A	A	Y						<i>aab/t</i>
H-L3-E	1	40	L	P	A						<i>gba</i>
E-L3-H	1	38	ns	P	P						<i>bbb</i>
	2	29	ns	A	B						<i>ba/eb</i> or <i>et/ab</i>
E-L3-E	1	37	A	A	L						<i>aag</i>
	11		A	A	A						<i>aaa</i>
H-L4-H	1	27	ns	ns	B	B					<i>agbb</i> or <i>a/g/babb</i>
H-L4-E	1	34	ns	A	A	ns					<i>t/eaab/a</i>
	2	20	ns	A	B	ns					<i>b/a/gabb/a/e</i>
E-L4-H	1	26	ns	A	A	ns					<i>b/aaab/a</i>
E-L4-E	1	17	A	A	L/P	P					New
	2	15	P	A/L	ns	P					<i>t/hggg</i>
	3	69	A	A	A	L					<i>aaag</i>
	4	28	B/P	A	A	ns					<i>baab</i>
H-L5-H	1	16	L	B	B	B	B				<i>gbbbbb</i>
	2	19	Y	A/P	A/L	B	B				Ambiguous
	3	6	ns	ns	A	A	A/B				Ambiguous
H-L5-E	1	26	ns	A	A	ns	ns				Ambiguous
	2	24	ns	B	B	B	ns				New
E-L5-H	1	21	ns	B/P	A	A	ns				New
E-L5-E	1	61	B	A	A	L	B				<i>baagb</i>
H-L6-H	1	12	L/B	B	B	A	A	A			New
	2	18	A/B	A	A	A/L	B	B/L			New
H-L6-E	1	11	L	B	B	B	B	B/A			New
E-L6-H	1	9	A/B	A	A	A	A/B	B/A			New
E-L6-E	1	24	B/A	A	A	A/B	L/B	B/P			<i>baaagb</i>
	2	18	P/A	A	B	P/E	ns	PA			New
	3	7	P/Y	A/P	L	B	P/Y	P/A			New
H-L7-H	1	25	L/B	A/B	B/L	B	B	ns	B		<i>gbbbbbb</i>
H-L7-E	1	19	L/B	ns	B	B	B/A	ns	B/A		New
E-L7-H	1	10	A/B	A/B	A	A/L	Y	B/Y	B/P		New
E-L7-E	1	8	ns	A	A	B/P	A/P	ns	ns		Ambiguous
	2	11	ns	ns	P	L/A	A/B	ns	P		Ambiguous
H-L8-H	1	7	L	P/B	ns	B/P	B	B/P	B/P	ns	New
	2	4	B	A	B/E	B/P	B/A	B	P/A	B/P	New
	3	6	Y	A	A	A/B	A/P	A/B	A/B	P/B	New
	4	6	ns	B/A	L/B	ns	A	A/B	B	B	New
H-L8-E	1	6	A/B	A/B	A/B	A	ns	A/Y	ns	A/E	New
E-L8-H	1	9	ns	A	A/P	A/P	A/P	A/B	P/A	ns	New
E-L8-E	1	4	A	A	A/B	A/B	A	A	ns	ns	New

^aMain-chain conformation (ϕ/ψ) of linker residues; P = β_P , A = α_R , L = α_L , B = β_E , Y = bridge region between helix and extended conformations in the ϕ,ψ map, ns = nonspecific, if more than two states are observed.

^bNotations from Donate et al. (1996), where *a* = α_R , *b* = $\beta_E + \beta_P$, *g* = $\alpha_L + \gamma_L$, *t* = transition between α -helical and β -sheet, *e* = positive ϕ in an extended conformation.

Fig. 3. Sequence logo of the linker E-L3-E with a motif AAL, preferring Gly at L₃ position.

within a cluster might be simply homologous copies. We examined the pairwise sequence identity of those proteins belonging to novel clusters, especially of linkers of length seven to eight residues, and found that most (>80%) of the members of each cluster are definitively not homologous (<25% pairwise sequence identity) and only a negligible number of them have pairwise sequence identity between 25 and 34%. Alignment of proteins within clusters revealed the following. The single cluster of the H-L7-E linkers had a maximum pairwise sequence identity (MPSID) of 43%, but a subset of 16 of them had MPSID < 25%. The single cluster of E-L7-H linker had MPSID of 38%, with a diverse subset of eight of them having MPSID < 25%. The first cluster of H-L8-H had MPSID of 31%, with a set of four sequences below the 25% threshold; the second cluster had MPSID of 29%, with three of

them below the 25% threshold; the third cluster had MPSID of 30%, with four of them below the threshold; and the fourth cluster had MPSID of 33%, with four of them below the threshold of 25%. The single cluster of E-L8-H linker had MPSID of 35%, with five of them below the threshold and the single cluster of E-L8-E linker had MPSID of 30%, with eight of them below the threshold. Even when whole protein sequences are forced into alignment, the relevant linker regions do not align themselves, but rather appear in distant regions of the protein chain, which again argues that the linker regions in our clusters are independent, nonhomologous occurrences. Thus, the evidence for the tendency of linkers to form specific conformation is not simply a result of distant evolutionary relatedness.

Are clusters of novel long linkers actually short linkers in disguise?

The novel clusters we have identified so far in this paper might arise due to poor definition of secondary structures, forming subsets of already defined clusters. For example, the E-L3-E linker with a motif AAL is similar to the E-L5-E linker with a motif EAAL, because the conformation of the end residues of the linker matches the flanking secondary structures. The cluster for the longer linker might arise due to the poor definition of ends of secondary structural elements. The linker H-L4-H, with L₁, L₂ in multiple states, L₃, L₄ as BB, might belong to H-L3-H with motif LBP. Similarly, the H-L7-E linker cluster might actually be the same as the H-L6-E linker cluster, if L₇ of the H-L7-E linker is indeed part of the flanking β -strand. However, most of the novel clusters are not subject to this potential confusion. The novel cluster of H-L6-H, with a motif L/BBBBAAA, cannot be categorized as an H-L3-H linker motif with the last three residues as part of a poorly defined α -helix because the secondary structural assignments at the last three positions of the linker are either 3_{10} -helix or β -turn, which are easily distinguishable from the exiting α -helix (the hydrogen bonding patterns are distinct).

Mainly α class of proteins with orthogonal architecture

The clustering of linkers in a nonredundant data set of proteins resulted in specific preferences for the main-chain torsions at var-

Fig. 4. Sequence logo of the linker H-L5-H with a motif LBBBB, with preference for Gly at L₁ position.

ious positions of the linker residues, irrespective of the orientation of the flanking elements. A closer look at the adjoining secondary structures revealed that these structures varied both in orientation and degree of interaction. In the context of loop modeling, we need to account not only for linker conformation, but also for the differing orientation of the flanking secondary structural elements. This is related to the problem of loop closure, which has been partly addressed (Kwasigroch et al., 1996) by superimposing the flanks from the same loop cluster. Thus, we were prompted to analyze linkers between successive helices of mainly α class of proteins with orthogonal architecture because we were assured of finding only a few well-defined angles relating the flanking helices.

A set of 49 proteins (see Materials and methods, Data set II) with pairwise sequence identity less than 45% was selected from the mainly α class of proteins having orthogonal architecture according to the CATH classification. Three-residue linkers were the most predominant, and hence we examined them closely. Six different clusters were identified (Table 2) using the same procedure as before (see Materials and methods). The helices on either side of the linker segment were represented by axial vectors (Geetha, 1996) and the angle between the helices was determined for each of the clusters (Table 2). The table suggests that the underlying premise of this analysis is indeed correct; the angle of orientation of the helices was well-defined in all the clusters, with relatively small variation. The first cluster with conserved LB/PP main-chain conformation had the flanking helices arranged orthogonal to each other. A structural superposition of the linkers (Fig. 5A) results in a mean RMSD of 0.43 (0.19) Å. The second cluster also has a conserved LB/PP backbone motif, but the flanking helices are oriented almost antiparallel to each other. A structural superposition of these linkers resulted in an RMSD of 0.44 (0.14) Å (Fig. 5B). The superposition includes the main-chain atoms of the three linker residues, as well as the two neighboring residues of the flanks. It can be seen that the interhelical orientation is well conserved apart from the main-chain conformation of the linker residues in both the clusters. Both the first and the second cluster have the same linker conformation, but have different orientation angles between the flanking helices. The parallel axis plot (Fig. 5C) shows the conserved ϕ, ψ at all the three positions of the linker, including both the first and the second cluster mentioned above.

The third cluster has a different linker conformation and the helices flanking the linkers of this cluster are antiparallel. Note the small deviation of the interhelical angle for the major clusters listed in Table 2. The ϕ, ψ torsions of the L_1 residue of the fourth cluster is in the "E" region of the ϕ, ψ plot, but with large variation

Table 2. H-L3-H linkers of mainly α proteins with orthogonal architecture

Cluster number	Number of examples	Main-chain conformation			Mean (SD) angle between helical axes (°)
		L_1	L_2	L_3	
1	17	L	B/P	P	97 (17)
2	5	L	B/P	P	157 (10)
3	7	A	A	Y	151 (8)
4	4	E	A	Y	115 (46)
5	3	L	L	Y	140 (26)
6	1	P	A	Y	158

Fig. 5. A: Structural superposition of some of the H-L3-H linkers from data set II along with the flanking helices, having LB/PP motif, showing the orthogonal arrangement of the helices. PDB entries along with chain identifier are 1hddC, 2cro_ (three examples), 1lmb3, 1lccA, 1r69_ (three examples), 1fiaA, 2wrpR. B: Structural superposition of H-L3-H linkers from data set II, along with the flanking helices (PDB entries are 1myt_, 3sdhA, 1lthA, 1colA, 1ccd_), having LB/PP motif, showing the antiparallel arrangement of the interacting helices. C: Parallel-axis plot of ϕ (solid line), ψ (dotted line) for H-L3-H linkers of data set II. Both the first and the second clusters (Table 2) are combined, because they have the same backbone conformation.

in the orientation angle. This cluster has not been observed earlier. These clusters, when compared with the earlier study on the H-L3-H loops, show a broad agreement of the preferred conformations for these linkers. The cluster AAB/Y, identified in an earlier study (Donate et al., 1996), had the flanking helices roughly orthogonal, but with large deviation, because it appeared to include a few flanks with antiparallel orientation in the tail of the angle distribution. The cluster AAY defined here (Table 2) shows much less variation in the distribution of the orientation angle between the flanks. A comparison of the clusters of this subset of proteins (Table 2) against those of the nonredundant data set (Table 1) shows a correspondence for LB/PP and AAY linker conformations.

Linkers containing the catalytic triad of serine proteases

It has long been known that the sequence as well as the structure around the catalytic triad residues is conserved in serine proteases (Perona & Craik, 1995; Bagley & Altman, 1996). This family of proteins is thought to exhibit convergent evolution toward a stable scaffold (Ollis et al., 1992). The strands are juxtaposed in a specific way to precisely position the side chains of the triad residues, responsible for catalytic activity (Bagley & Altman, 1996). Con-

Table 3. Catalytic triad of serine protease folds

PDB code ^a	His 57	Asp 102	Ser 195
1arb_	E-L7-H	E-L4-E	E-L14-E
2sga_	E-L1-H	E-L5-E	E-L10-E
2alp_	E-L9-E	E-L13-E	E-L15-E
1sgt_	E-L8-E	E-L5-E	E-L10-E
1gctA	E-L10-E	E-L13-E	E-L13-E
2pkaB	—	—	E-L14-E
4estE	E-L10-E	E-L15-E	E-L14-E
1thsH	—	E-L14-E	E-L18-E
1hneE	—	E-L12-E	E-L14-E
3rp2A	E-L7-E	E-L13-E	E-L15-E
1sgt_	E-L8-E	E-L11-E	E-L16-E
2pkaA	E-L8-E	E-L11-E	—
Conserved linker motif	E-L5-L(n-5)-E	E-L(n-5)-L5-E	E-L(n-8)-L8-E

^aPDB code includes chain ID and “_” denotes no chain ID.

formational clustering of linkers of same length fails to detect the structural conservation around the catalytic triad of proteins like serine proteases, because of the large variation in the linker length containing the catalytic triad residues. Aspartic proteinases, how-

Fig. 6. **A:** Structural superposition of E-L(n-8)-L8-E linkers containing Ser 195 of proteins from data set III (see the Materials and methods). A multiple β -turn around the catalytic residue is a conserved feature of this linker. **B:** Parallel-axis plot of ϕ (solid lines), ψ (dotted lines) for the conserved eight residue portion of the linker containing Ser 195 of proteins from data set III. The structural motif of the linker is E-L(n-8)-L8-E, where L(n-8) and L8 are the variable and conserved parts of the linker.

Fig. 7. Structure-based sequence alignment for the linker regions containing the catalytic triad residues of serine proteases. Arrows indicate the β -strands; solid bar indicates the linker positions.

ever, do show a conserved linker motif (Donate et al., 1996) in proteins having a pairwise sequence identity between 25 and 60%. Serine proteases are perhaps a more difficult family to study in this regard, owing to the larger sequence diversity (13–43% pairwise sequence identity in our data set). Our study now focuses on variable-length linkers containing the catalytic triad residues, in an attempt to find a structurally conserved portion of the linker and, hopefully, a canonical linker conformation.

The linkers containing the catalytic triads of serine proteases (see Materials and methods, Data set III) are mostly found to connect β -strands and are seen to have different lengths (Table 3). One member of the triad, Ser 195, is held between adjacent β -strands with linker lengths ranging from 10 to 18 residues. The strands are not interacting, but a structural superposition of the strands (Fig. 6A), along with the C-terminal region of the linker, resulted in an overall RMSD of 0.82 (0.24) Å for the main-chain atoms of the residues in the strands and that portion of the linker that is conserved. Apart from the orientation of the strands, the C-terminal part of this linker has a conserved multiple β -turn that connects to the exiting β -strand. The linker can therefore be split into a flexible portion that we denote as L(n-8) and a conserved part as L8. The full linker is therefore E-L(n-8)-L8-E, with the L8 having conserved conformation BPBPLAPL (Fig. 6B). The corresponding structure-based sequence alignment (Fig. 7) shows the conserved L8 linker containing the catalytic triad residue.

The second residue of the catalytic triad, Asp 102, is part of the loop linking two interacting strands that together form a β -sheet. It

is quite surprising to note the structural conservation of a portion of the linker region, once again with a conserved β -turn around Asp 102. The overall RMSD for the backbone atoms of a portion of the linker region, along with the residues of the strands, after structural superposition is 0.54 (0.12) Å (Fig. 8A). The conserved motif for the linker containing Asp 102 turns out to be five residues long, so the entire linker is denoted by E-L(*n*-5)-L5-E, where the main-chain conformation for the conserved L5 portion is L/A/B P/ELYB (Figs. 7, 8B). The first residue of the conserved part of the linker shows some variation, although the rest of them have a low degree of variation in the main-chain torsion angles.

The third catalytic residue, His 57, is part of a loop linking noninteracting β -strands with a conserved 3_{10} -helix as part of the linker region. A structural superposition of the main-chain atoms of the linker (Fig. 9A) results in an RMSD of 0.68 (0.23) Å for the N-terminus part of the linker region. The linker is therefore E-L5-L(*n*-5)-E and the motif of the conserved part of the linker is BPAAA for the first five residues at the N-terminus of the linker connecting the adjacent β -strands. The main-chain conformation of the five residues are plotted as a parallel axes as in Figures 7 and 9B. In summary, the structural conservation of the region of each linker containing the catalytic triad residues, along with the orientation of the adjoining secondary structures, has emerged as a direct consequence of analysis of linkers of length more than eight residues. We hypothesize that the nonconserved part of the linker might be responsible for the binding specificity of the proteases to different inhibitors or substrates.

Fig. 8. A: Structural superposition of E-L(*n*-5)-L5-E linkers containing Asp 102 of proteins from data set III, showing the conserved β -turn around the catalytic residue. B: Parallel-axis plot of ϕ (solid lines), ψ (dotted lines) for the conserved five-residue portion of the linker containing Asp 102 of proteins from data set III. The structural motif of the linker is E-L(*n*-5)-L5-E, where L(*n*-5) and L5 are the variable and conserved parts of the linker.

Fig. 9. A: Structural superposition of E-L5-L(*n*-5)-E linker containing His 57 of serine proteases from data set III, showing a conserved 3_{10} -helix around the catalytic residue. B: Parallel-axis plot of ϕ (solid lines), ψ (dotted lines) for the conserved five-residue portion of the linker containing His 57 of serine proteases from data set III. The structural motif of the linker is E-L5-L(*n*-5)-E, where L5 is the conserved part of the linker.

Our analysis of linkers of length two to eight residues has demonstrated that linkers take on distinct conformations and form regular motifs, adding support to the emerging notion of "nonrandomness of loops." Knowledge of these structural motifs should be of great assistance to knowledge-based loop modeling methods, even though loop closure still remains a problem. The linkers found in the nonredundant data set of proteins points to the existence of motifs irrespective of the orientation of the flanking structural elements. The broad agreement of the clusters identified here with those found earlier suggests the concept of structural motifs for the linkers to be largely independent of the secondary structural assignment program used. Analysis of proteins with less diverse orientation of the flanking elements, namely the mainly helical class of proteins, resulted in two broad conformations corresponding to the orthogonal and antiparallel orientation of the flanking helices.

Our focus on varying loop lengths, which maintained the geometry and orientation of the flanking secondary structural elements, has led to the definition of canonical loop structures for serine proteases. We have attempted to highlight the structural conservation as well as the flexibility of linkers. This analysis has allowed us to identify the conserved structural features for part of the linkers that contain the catalytic triad residues. The clustering and analysis of conformations of loop regions linking standard secondary structures has thus resulted in a library of distinct motifs that might be useful for modeling the loop regions in proteins.

Especially, given the orientation of the flanking helices between a defined linker length in cases like helical bundles, the table of linker motifs should greatly facilitate choosing the conformation of the linker for the purpose of designing the linker regions.

Materials and methods

Database of linkers

Data set I

A data set of 330 unique protein chains from the Protein Data Bank (Bernstein et al., 1977), which have $\leq 45\%$ pairwise sequence identity (Hobohm et al., 1992; Hobohm & Sander, 1994), with resolution better than 2.25 Å, and R -factor ≤ 0.25 is considered. Secondary structural regions are assigned according to the program STRIDE (Frishman & Argos, 1995). Linker regions of lengths from two to eight residues (L_n , $n = 2-8$), which connect secondary structures, helices, and strands, are isolated from the data set. The resultant linker database is represented in Figure 1. There are, in total, about 3,133 linkers, of which 446 are between two helices (H- L_n -H), 748 between a helix and a strand (H- L_n -E), 558 between a strand and a helix (E- L_n -H), and 1,381 between two strands (E- L_n -E).

Data set II

A set of mainly α proteins of orthogonal architecture, classified according to CATH-release of January 1995 (Orengo et al., 1993, 1994) with low pairwise sequence identity, were considered for analysis of linkers between two successive helices arranged in either orthogonal or antiparallel conformation. The proteins with PDB code along with the chain identifier are: 1eca_, 1myt_, 1mbd_, 3sdhA, 1hbg_, 1thbA, 2lhb_, 1pbxA, 1hdsB, 2mhba, 1pbxA, 1mba_, 1lhl_, 1lthA, 1cpcA, 1colA, 1lmb3, 1lccA, 1adr_, 2cro_, 1r69_, 1utg_, 1ccd_, 1aca_, 1fiaA, 2wrpR, 1hddC, 1ahdP, 1cmbA, 3mdsA, 3sdpA, 1abmA, 1rro_, 1omd_, 5pal_, 4cpv_, 1osa_, 5tnc_, 4icb_, 2sas_, 1ycc_, 351c_, 3c2c_, 155c_, 2mtaC, 1cc5_, 2cdv_, 1prcC. There are 42 three-residue linkers connecting adjacent helices.

Data set III

A set of serine proteases (Table 3), with low pairwise sequence identity from CATH-release of January 1995 (Orengo et al., 1993, 1994), were considered for analysis of linker regions and their role in structural conservation around the catalytic triad residues.

Clustering of linkers and analysis of conformations

Linkers of length $n = 2-8$ are described by various inter- α -carbon distances between residues in the linkers and the ends of the entering and exiting secondary structures. The α -carbon distances for a linker of size n are taken from C_{i+j-2}^α to C_{i+k-2}^α for $1 \leq j \leq k \leq n + 2$. An initial clustering procedure with all the relevant α -carbon distances is done (Geetha & Munson, 1997) by k -means clustering (Fischer & Van Ness, 1971) using commercially available software (JMPTM v. 3.0.1, 1994). These clusters are, in turn, clustered by k -means based on backbone torsions. The ϕ, ψ regions considered here (Fig. 2) are right-handed helix (α_R or A), left-handed helix (α_L or L), beta antiparallel (β_E or B), beta parallel (β_P or P), bridge region between α -helix and β -sheet regions (Y, named here for convenience), and epsilon (ϵ or E), occupied mostly by Gly (Efimov, 1993). The resulting clusters from the backbone conforma-

tional space are once again k -means clustered for the side-chain torsions (g^+ , g^- , and t), when it is possible to do so. We have not made any attempt to do a complete clustering of all of the data, but have eliminated data that lie outside the clusters.

Sequence preferences

Amino acid preferences within the various positions of the linkers are difficult to estimate, especially for the small number of examples within a cluster. This is true for relatively medium and long linker lengths. However, we made estimates of sequence preference for the clusters of the linkers, based on the sequence logos (Schneider & Stephens, 1990), the graphical display of consensus sequences with the most common residue stacked on top of other residues in a given position. The method also gives a measure of the information content at every position.

Acknowledgments

We thank Dr. Jean Garnier for critical reading of the manuscript. We also thank the referees for their useful criticisms and helpful suggestions.

References

- Bagley SC, Altman RB. 1996. Conserved features in the active site of nonhomologous serine proteases. *Folding and Design* 1:371-379.
- Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, 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.
- Chelvanayagam G, Roy G, Argos P. 1994. Easy adaptation of protein structure to sequence. *Protein Eng* 7:173-184.
- Cohen BI, Presnell SR, Cohen FE. 1993. Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci* 2:2134-2145.
- Donate LE, Rufino SD, Canard LHJ, Blundell TL. 1996. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Sci* 5:2600-2616.
- Efimov AV. 1986. Standard conformations of a polypeptide chain in irregular regions of proteins. *Mol Biol (USSR)* 20:250-260.
- Efimov AV. 1991. Structure of α - α hairpins with short connections. *Protein Eng* 4:245-250.
- Efimov AV. 1993. Standard structures in proteins. *Prog Biophys Mol Biol* 60:201-239.
- Fischer L, Van Ness JW. 1971. Admissible clustering procedures. *Biometrika* 58:91-98.
- Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. *Proteins Struct Funct Genet* 23:566-579.
- Geetha V. 1996. Distortions in protein helices. *Int J Biol Macromol* 19:81-89.
- Geetha V, Munson PJ. 1997. Analysis of linkers of regular secondary structural regions in proteins. In: Marshak DR, ed. *Techniques in protein chemistry, vol VIII*. Orlando, Florida: Academic Press. pp 667-677.
- Govindaraj S, Poulos TL. 1995. Role of the linker region connecting the reductase and heme domains in cytochrome P450_{BM-3}. *Biochemistry* 34:11221-11226.
- Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci* 3:522-524.
- Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci* 1:409-417.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.
- Kwasigroch JM, Chomilier J, Mornon JP. 1996. A global taxonomy of loops in globular proteins. *J Mol Biol* 259:855-872.
- Mattos C, Petsko GA, Karplus M. 1994. Analysis of two residue turns in proteins. *J Mol Biol* 238:733-747.
- Milner-White EJ, Poet R. 1986. Four classes of β -hairpins in proteins. *Biochem J* 240:289-292.

- Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J et al. 1992. The α/β hydrolase fold. *Protein Eng* 5:197–211.
- Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng* 6:485–500.
- Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–634.
- Perona JJ, Craik CS. 1995. Structural basis of substrate specificity in the serine proteases. *Protein Sci* 4:337–360.
- Rice PA, Goldman A, Steitz TA. 1990. A helix-turn-strand structural motif common in α - β proteins. *Proteins Struct Funct Genet* 8:334–340.
- Ring CS, Kneller DG, Langridge R, Cohen FE. 1992. Taxonomy and conformational analysis of loops in proteins. *J Mol Biol* 224:685–699.
- Rooman M, Rodríguez J, Wodak S. 1990. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol* 213:327–336.
- Scheerlinck JPY, Lasters I, Claessens M, De Maeyer M, Pio F, Delhaise P, Wodak S. 1992. Recurrent $\alpha\beta$ loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. *Proteins Struct Funct Genet* 12:299–313.
- Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res* 18:6097–6100.
- Sibanda BL, Thornton JM. 1985. β -Hairpin families in globular proteins. *Nature* 316:170–174.
- Sibanda BL, Thornton JM. 1991. Conformation of β -hairpins in protein structures: Classification and diversity in homologous structures. *Methods Enzymol* 202:59–82.
- Srinivasan N, Sowdhamini R, Ramakrishnan C, Balaram P. 1991. Analysis of short loops connecting secondary structural elements in proteins. In: Ramakrishnan CBP, ed. *Molecular conformation and biological interactions*. Bangalore: Indian Academy of Science, pp 59–73.
- Thornton JM, Sibanda BL, Edwards MS, Barlow DJ. 1988. Analysis, design and modification of loop regions in proteins. *Bioessays* 8:63–69.
- Tramontano A, Lesk AM. 1992. Common features of the conformations of antigen binding loops in immunoglobulins and application to modeling loop conformations. *Proteins Struct Funct Genet* 13:231–245.
- Wilmot CM, Thornton JM. 1988. Analysis and prediction of the different types of β -turns in proteins. *J Mol Biol* 203:221–232.
- Wintjens RT, Rooman MJ, Wodak SJ. 1996. Automatic classification and analysis of turn motifs in proteins. *J Mol Biol* 255:235–253.
- Wong WK, Gerhard B, Guo ZM, Kilburn DG, Warren RAJ, Miller RCJ. 1986. Characterization and structure of an andoglucanase gene *cenA* of *Cellulomonas fimi*. *Gene* 44:315–324.