Analysis, clustering and prediction of the conformation of short and medium size loops connecting regular secondary structures.

Stephen D. Rufino, Luis E. Donate, Luc Canard and Tom L. Blundell.

The Imperial Cancer Research Fund,
Unit of Structural Molecular Biology,
Department of Crystallography,
Birkbeck College, University of London,
Malet Street, London WC1E 7HX, UK.

Loops are regions of non-repetitive conformation connecting regular secondary structures. They are both the most difficult and error prone regions of a protein to solve by X-ray crystallography and the hardest regions to model using knowledge-based procedures. While the core of a protein can be straight forwardly modelled from the structurally conserved regions of homologues of known structure, loops must be modelled from a selected homologue or from a loop chosen from outside the family. Here we present a loop prediction procedure that attempts to identify the conformational class of the loop rather than to select a specific loop from a database of fragments. The structures of some 2083 loops of one to eight residues in length were extracted from a database of 225 protein and protein domain structures. For each loop, the relative disposition of its bounding secondary structures is described by the separation between the tips of their axes, the angle and dihedral angle between their axes. From the clustering of the loops according to the root mean square deviation of their spatial fit, a total of 162 loop conformational classes, including 79% of loops, were identified. One-hundred and eight of these, involving 66% of the loops, were populated by at least four non-homologous loops or four loops sharing a low sequence identity. Another 54 classes, including 13% of the loops, were populated by at least three loops of low sequence similarity from three or fewer non-homologous groups. Most of the previously described loop conformations were found among the populated classes. For each class a template was constructed containing both sequence preferences and the relative disposition of bounding secondary structures among member loops. During comparative modelling, the conformation of a loop can be predicted by identifying a loop class with which its sequence and disposition of bounding secondary structures are compatible.

1. Introduction

During the comparative modelling of the structure of a protein, its structurally conserved regions, SCRs, are defined by the superposition of homologues of known structure. They are then used to determine the framework of the model. The backbone conformation in regions outside the SCRs, the structurally variable regions or SVRs, have to be predicted from their sequence, the spatial disposition of their surrounding framework fragments and the overall framework. The sequence of the SVR provides residue conformational preferences, the surrounding framework fragments limit the conformations by specifying the SVRs' termini and the overall framework allows the elimination of loop conformations that would clash with it. Additional hints may be found in other known structures which have core regions of similar disposition linked by a fragment of the desired length. The length of an SVR is an important factor in the likelihood of predicting its correct conformation and even in the potential to attempt a prediction. Since an SVR can even contain inserted domains we will restrict further discussion to short and medium size loops linking secondary structure elements.

When modelling a loop region the simplest case is that of finding a loop of identical length and of similar sequence among the homologues of known structure. If no loop is found within the family, then other protein structures can be searched for fragments whose three first and last residues superpose well with the termini of the framework bounding the loop[1,2]. All identified fragments are then least-square fitted to the framework and those whose backbone clash with that of the framework are eliminated. If several fragments still remain, Blundell *et al.*[2] suggest determining an averaged backbone trace from the fitted fragments and then selecting the one that best fits the trace. Alternatively, Topham *et al.*[3] propose a scoring scheme based on the compatibility of the loop sequence with templates derived from the fragments. They obtain the template of a fragment by applying conformationally constrained environmental amino acid substitution tables to its sequence. The tables were generated from observed substitutions in a set of protein families. So as to evaluate the fit of a loop sequence to the conformation of a fragment and not to possible conformations derived from it through substitutions, only substitutions that maintain residue backbone conformation were considered.

Although irregular, loops have been shown by many studies not to have completely random backbone conformations[4,5,6,7,8,9,10,11]. For example, Efimov describes in his work[7,8] several typical αα-loop and ββ-loop classes. Out of the 106 β-hairpins they analysed, Sibanda and Thornton[4,6] identified a total of 49 that belonged to five conformational classes. In a study of βαβ units, Edwards, Sternberg and Thornton[5] found that out of 129 loops of less than 12 residues only 18 fell within four conformational classes with distinctive sequence patterns. The conformation of a loop is at least partially determined by its sequence and the spatial relations of its bounding secondary structures. The sequence of a loop affects its conformation through the need to satisfy residue conformational preferences, burial of hydrophobic residues and pairing or exposing of charged residues. 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[4,6] identified the residue preferences as glycine, asparagine and aspartate for the first position in an α_L conformation and glycine for the second position in an γ_L conformation. The burial of a large hydrophobic residue, usually Val, Ile and Leu, at the fourth position of the L1 loop of the immunoglobulin light chain variable Vκ domain, within a cavity between the domain's two β-sheets, was identified by Tramontano, Chothia and Lesk[12] as an important determinant in the loop's conformation. Similarly, the two major determinants of the conformation of the Vκ domain L3 loop were the burial and hydrogen bonding of a polar residue, glutamine or asparagine, and the presence of a cis-proline.

Following the work of Sun and Blundell[11], we have developed an approach to loop selection based on the identification of preferred loop conformational classes. In this paper, we describe how the favoured loop conformations were identified from the 2083 short and medium size loops in a database of 225 protein structures. We then show how for each loop class, sequence preferences and average disposition of bounding secondary structures can be determined. Some previously described loop conformations are compared to the loop classes. Finally, we illustrate how the conformation of a loop can be predicted by comparing its sequence and disposition of bounding secondary structures to those observed for the loop classes.

2. Methods

2.1. Protein structure database.

225 protein and protein domain structures with resolutions better than 2.5Å were selected from the Brookhaven Protein Data Bank[13]. These included 182 structures from a database of 66 protein families[14], structurally aligned using the MNYFIT[15] and COMPARER[16] programs, and 43 unique structures[17].

2.2. Identification of secondary structure elements.

The secondary structure assignment of each of the proteins in our database was determined following the hydrogen-bond pattern based procedure of Kabsch and Sander[18] as encoded in the program SSTRUC of Smith, D. K. and Thornton, J.M. Secondary structure elements were defined as continuous fragments of identical secondary structure assignment. Short 3_{10}-helices containing only three residues, which are often found to play a role in polypeptide chain reversals or redirections, were not considered as secondary structure element. Although 3_{10}-, α- and π

-helices were grouped in a general helical class, contiguous 3_{10}-,α- or π-helices were not merged.

2.3. Identification of loop motifs.

Loops were defined as regions of non-repetitive secondary structure between secondary structure elements. A loop motif was defined to include both a loop and its pair of sequentially adjacent 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. that bordering the central loop) of β-hairpin β-strands were determined using only the hydrogen bonding patterns within the loop motif itself.

2.4. Associated vectors.

For each loop motif two vectors were determined representing the direction and position of the proximal termini of its two secondary structures. Vectors were fitted to the proximal termini of secondary structures so as to avoid large deviations between vectors and the local axis at the termini of their associated secondary structures, a problem associated with curved or kinked secondary structures. Vectors associated with β-strands, 3_{10}-, α- and π-helices were calculated using all their residues up to a maximum of 6, 7, 8 or 9 proximal residues respectively. If a β-strand of only two residues in length was bounded at its distal terminus (i.e. that most distant from the central loop) by neither an amino-acid chain terminus nor a main-chain break then an extra distal residue was used in the determination of its associated vector. Vectors associated with secondary structures of four or more residues were defined so as to minimise the sum of the squares of the distances from the vector to the Cα atoms of the corresponding secondary structure[19,10]. 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 Cα atoms with each of the two terminal Cα atoms contributing half the weight of the central Cα atom. The vector associated with a β-strand containing only two residues was defined as the virtual bond linking its two Cα atoms.

For each loop motif, its inter-vector separation, angle and dihedral angle were calculated. The separation between the two vectors of a motif was defined as the distance between the projections of the proximal residues of their respective secondary structures. The vector linking these projections in sequential order was used to determine both the direction of the inter-vector angle and the dihedral angle.

2.5. Backbone conformation.

For each motif the backbone conformation of the residues in its loop region was determined and described using a reduced notation of seven residue conformational classes[3] (figure 1): α-helical (a), anti-parallel 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). These seven classes were reduced to five by the merging of p and b as well as the g and l conformational classes into the larger classes b and g respectively.

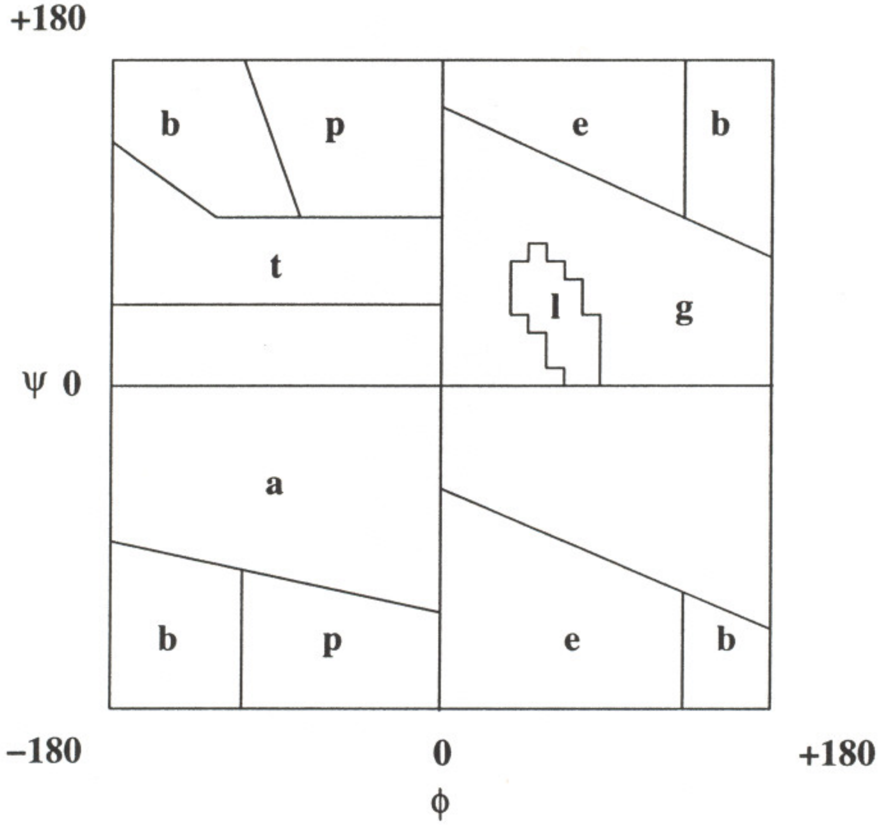

Fig. 1: Ramachandran plot of residue ϕψ angles indicating the seven conformations considered (taken from Topham *et al.* [3]).

2.6. Loop database.

2083 motifs with loops containing between one and eight residues were identified from the 225 protein structures in our database. These motifs were first classified according the type of their bounding secondary structures into four groups: HH; EH; HE; EE. Further classification according to the length of their loop regions yielded a total of 32 groups.

2.7. Identification of loop conformational classes.

To identify motifs with similar loop conformations all loops within a group were superposed[20] pairwise using their carbonyl oxygen and carbon atoms, their Cα atoms and their amide nitrogen and hydrogen atoms. For each amide bond a hydrogen atom was generated 1Å away from the amide nitrogen, in the amide bond plane, along the line which bisects the angle formed by the carbonyl carbon, amide nitrogen and Cα atoms and most distant from the carbonyl carbon and Cα atoms. So as not to over emphasise the fitting of the backbone atoms versus the ϕ ψ angles the carbonyl oxygen and the amide nitrogen were given weights of 1.5 while all other atoms were given weights of 1.

For each of the 32 motif groups a dissimilarity matrix was generated from all pairwise superpositions using the distance measure:

$- \ln ({}^{1}/_{1\,+\,\mathrm{r.m.s.d.}}) = \ln (1 + \mathrm{rmsd})$

derived from the root mean square deviation, r.m.s.d., of individual pairwise superpositions. All loops within a group were then clustered according to their dissimilarity matrix, using the hierarchical clustering program KITSCH[21], which is part of the Phylogenetic Inference Package, PHYLIP. For a given dissimilarity matrix, the dendrogram selected by this procedure is one that minimise the difference between the distances in the dendrogram and those in the dissimilarity matrix[22]. The distance between two loop motifs is represented in a dendrogram by the distance along the horizontal axis from the tip of their respective branches to their first shared branching point.

From the dendrograms, clusters containing at least four non-homologous loops or with loops differing in sequence by at least a quarter of their residues were identified as a loop conformational class. Additionally, clusters with at least three loops, differing in sequence by at least a quarter of their residues, from three or fewer non-homologous groups were identified and labelled as such (figures 2, 3 and 4).

2.8. Description of residue types and environments.

21 residues types were considered, the 20 amino-acids encoded by the genetic code with the addition of the cystine, an oxidised cysteine residue involved in a disulphide bridge. A total of 216 residue environments were defined in terms of secondary structure, backbone conformation, relative side-chain solvent accessibility and side-chain hydrogen bonding[3].

The secondary structure and backbone conformation of a residue is evaluated as either β-sheet or helical, including 3_{10}, α and π-helical, if it belonged to a secondary structure as defined by Kabsch and Sander[18] or as one of the 7 backbone conformations already described if it was in a coil region.

Three classes of relative side-chain accessibility[23], based on a spherical probe with a radius of 1.4Å, were considered: less than 7%; between 7 and 40%; greater than 40%.

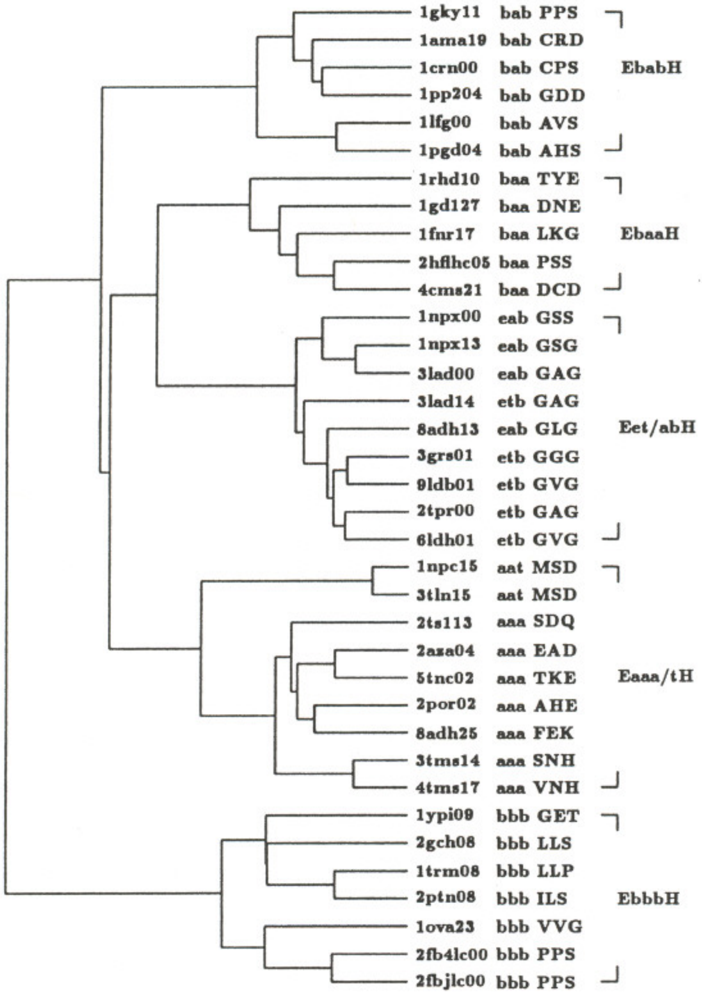

Fig. 2: Clustering of loops of length 3 linking an extended β-strand to an α -helix. Loops not belonging to a cluster were omitted for the sake of readability. The loop names, conformations, sequences and classes are indicated.

Three types of side-chain hydrogen bonds were considered: to other side-chain or heteroatom; to main-chain carbonyl oxygen; to main-chain amide hydrogen. Hydrogen bonds with preceding and following side-chains were excluded. Side-chain hydrogen bonding was defined using only donor-acceptor distance[14] and not angles since side-chain atoms are not always reliably positioned by X-ray crystallography. Considering that a residue side-chain can have either no hydrogen bond, hydrogen bonds of one, two or all three types there is a total of eight side-chain hydrogen bonding classes.

2.9. Environment dependent substitution tables.

The conformationally constrained environmental amino-acid substitution tables used here are fully described in Topham *et al.*[3]. The substitution tables were generated from 311,422 observed substitutions, that maintained the backbone conformation of the involved residues, in a database of 352 protein and protein domain structures grouped into 98 homologous families. The tables describe the probability with which a residue of a given type in one of the 216 defined environments is mutated to another selected residue type with an identical backbone conformation but in an unspecified environment. The oxidation state of cysteine residues in loops of unknown structure was presumed to be unknown and so although the initial residue of a substitution could be of 21 different types, including cystine, the residue mutated to could only be of the 20 types in the genetic code.

2.10. Generation of templates for loop conformational classes.

For each identified loop class a sequence template was generated by applying the conformationally constrained environmental amino-acid substitution tables of Topham *et al.*[3] to the sequences of its member loops. At each position in a template the probability of finding a specific residue was determined from the probabilities of the residues in the corresponding position of each member loop being substituted to the residue in question. The contribution of a loop to a template was weighted by the inverse of the number of its homologous member loops, so that for example, five homologous aspartic proteinase loops would each contribute with a weight of 1/5 to the template while a unique serine proteinase loop would contribute with a weight of 1.

In addition to its sequence template each loop class template contains the average value and standard deviation for the distributions of the inter-vector separation, angle and dihedral angle of the member loop motifs (figure 5 and 6).

2.11. Scoring a loop motif against a loop template.

When comparing a loop motif to a loop templates several estimators were considered. The sequence score, S_{seq} derived from the probability, P_{seq}, of the sequence of the motif's loop matching the sequence template was defined as:

$$S_{seq} = 100 * (P_{seq})^{1/n} = 100 * (P_1 * P_2 * P_3 *...* P_n)^{1/n},$$

where P_1, P_2, P_3 and P_n, are the probabilities of respectively the first, second, third and last residue of the motif's loop at the corresponding positions of the sequence template. For each of the three inter-vector relations, namely separation, angle and dihedral angle, the difference between their values in the motif and their average values in the template were measured in absolute terms and in number of standard deviations.

3. Results and discussion

3.1. Identification and clustering of loop motifs.

From the database of 225 protein and protein domain structures of 2.5Å or better resolution 2083 loops of length of one to eight residues were identified. A loop was defined as a region of non-repetitive secondary structure bounded by two secondary structures, while its motif was defined to include both the loop and its adjacent secondary structures. Secondary structure was identified as suggested by Kabsch and Sander[18] with two small modifications. Firstly, 3_{10}-helices of only three residues in length were considered as belonging to their surrounding loop region. Secondly, to avoid erroneous classification of β-hairpin motifs due to hydrogen bonding with β-strands outside the motif, the proximal termini (i.e. that bordering the central loop) of β-hairpin β-strands were determined using only the hydrogen bonding patterns within the loop motif itself. For each secondary structure in a loop motif an associated vector[19,10] was calculated representing the local axis of its proximal termini. The spatial relations between the two vectors of a loop motif were defined by the separation of their proximal termini, their angle and their dihedral angle.

The 2083 loops were grouped into a total of 32 groups according to their length and to the type of their bounding secondary structures, which were either helical or β-strand. The loops within each group were pairwise superposed as described previously. Within a group, populated loop backbone conformation classes were then identified, by clustering the loops according to their pairwise structural similarity using the hierarchical clustering program KITSCH[21] (figure 2 and 3). The name of a loop class was derived from the type of its bounding secondary structures and the five state backbone conformation of its member loops. Where necessary, alternate conformations are indicated separated by a backslash. If several clusters were found to have the same conformation then a simple numbering scheme was used to distinguish them. For example, Haab/tH describes a loop class of length three, conformation aab or aat, linking two helices and Eaa-2-E describes one of several loop classes of length two, conformation aa, linking two β-strands. Two types of loop conformation classes were identified (figure 4), those whose clusters contained at least four non-homologous loops or loops differing in sequence by at least a quarter of their residues and those whose clusters contained at least three loops, differing in sequence by at least a quarter of their residues, from at the most three non-homologous groups. This second type of class, the "small" classes indicated by a final "s" in the naming scheme, are often but not always specific to a protein family as is for example the case in the immunoglobulin loop classes Eab/tbbaaabEs, EbaeEs, Ebb/ag/tbbaaEs, Ebba/bb/ebEs, EbbgbbEs, EbgbbaaabEs, Eg/bbbaaabEs and Egbb/taaabbEs. Of the 2083 motifs studied 79% clustered into a total of 162 loop conformational classes with 66% in 108 populated classes and an additional 13% in 54 "small" classes. Fewer of the longer loops were found to cluster when compared to shorter loops with, for example, among the loops connecting two strands 82.3% of those of length two as compared to 48.1% of those of length eight.

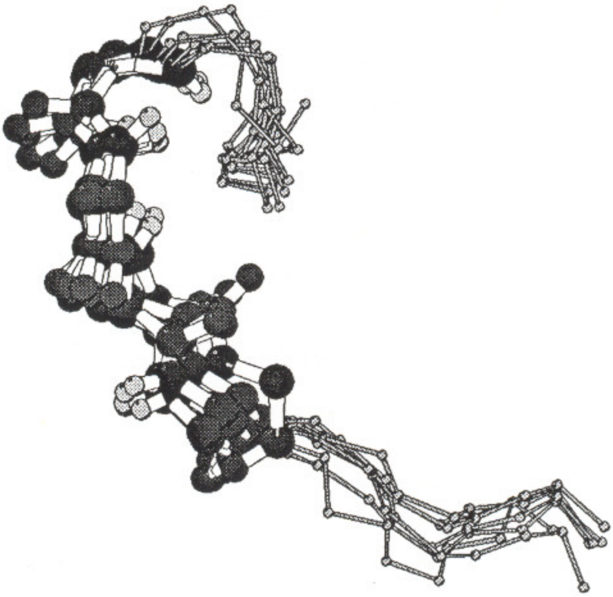

Fig. 3: Superposition of some loop of length 4 of the HagbbE class.

Among the identified classes most of the previously described loop conformations were found. In their study of 106 β-hairpins Sibanda and Thornton[4,6] identified 49 loops that fell within four loop conformational classes, the 2:2 β-hairpins of type I, I' and II', the 4:4 β-hairpins of type I and the β-hairpins 3:5 of type I. For each of these, the conformation described by the authors were $\alpha_R\gamma_R$, $\alpha_L\gamma_L$, $\varepsilon\gamma_R$, $\alpha_R\alpha_R\gamma_R\alpha_L$ and $\beta\alpha_R\gamma_R\gamma_L\beta$ respectively while the corresponding class identified in the present work were Eaa-2-E, EggE, EeaE, EaaagE and EbaagbE respectively. In an analysis of 129 βαβ unit loops Edwards, Sternberg and Thornton[5] identified four conformational classes, which together contain a total of only 18 loops. They considered adjacent βαβ units whose β-strands hydrogen bond to one another and non-adjacent βαβ units whose β-strands are separated by intervening β-strands. Excluding the non-adjacent βα class of length zero that they describe all three other classes are also found in this study although the boundaries of the loops vary due to the slightly different secondary structure definitions used. The loops in the adjacent αβ class of length one were found in the HagbbE class, those in the adjacent αβ class of length three where found in the Hg/abaE class and those in the adjacent βα class of length three where found in the Eet/abH class.

3.2. Sequence templates

For each loop conformational class a template was constructed containing information about sequence preferences and relative secondary structure disposition of its member loop motifs. The sequence template of a loop class was determined from the sequences of its member loops. The probabilities of finding each of the 20 amino-acids at a specific position of a sequence template were derived by applying conformationally constrained environmental substitution tables[3] to each of the residues found at that position in member loops. Each loop class template also contained information about the distribution of the inter-vector separation, angle and dihedral angle of its member loops measured terms of average value and standard deviation.

From the sequence template of a loop conformational class, specific sequence preferences can be identified. For example, the sequence template of the EggE class (figure 5), which corresponds to the type I' 2:2 β-hairpins, shows in the first position a marked preference for glycine with a probability of 0.358, followed by asparagine and aspartate with respective probabilities of 0.169 and 0.118. In the second position a very strong preference for glycine is shown with a probability of 0.795 followed distantly by asparagine with a probability of 0.048. These preferences are similar to the sequence patterns identified by Sibanda and Thornton[4,6]. In the case of the EeaE class (figure 6) corresponding to the type I' 2:2 β-hairpins preferences are glycine, aspartate and asparagine at the first position with probabilities of 0.788, 0.048 and 0.045 respectively and aspartate, serine, asparagine, proline, threonine and glycine at the second position with probabilities of 0.189, 0.140, 0.099, 0.092, 0.079 and 0.077 respectively. The preferences at the second position differ from the patterns found by Sibanda and Thornton[4,6] which included only serine, threonine and glycine.

EE

1 a; b; t-s
2 ba; ea; aa-1; aa-2; gb-s; bb; b/eb; gg
3 bae-s; tae/b-s; aag; aaa; bgg/a/b; gga; abg-s; bbb; a/g/tbb
4 abbb/a/e; t/bggb; agab; bbgb-s; aaag; t/a/g/baab; baab; babb-s; eaag-s
5 baagb; bbgbb-s; abaab; bba/bb/eb-s; bbggb-s; agagb-s; aabab-s; bbgbt-s; b/taaab/a
6 baagab; aabb/ab/tb; baaagb; baaagb-s; b/taabab-s
7 b/tbb/a/eg/t/a/ebbb; bb/ag/tbbaa-s; g/bbbaaab-s; a/tb/abbbba/b-s; baaa/ta/bgb/a
8 bbaagbba; bbbgataa-s; ab/tbbaaab-s; bgbbaaab-s; gbb/taaabb-s; bb/ea/eb/abb

EH

1 b/p; b; b-HTAL; b-MKEW; b-GSDP; a; t
2 bb/e; ab; bb/t; bg-s
3 aaa/t; bab; baa; et/ab; bbb
4 b/aaab/a; a/babb/a; b/t/ebbb/e-s; bba/tb/e; aeaa-s; aggb-s
5 b/e/a/gg/b/tbab; b/ebbgb/t-s; a/baab/tb; b/aaaab/e/t; b/ab/tga/bg/t/b-s; a/t/g/bbb/t/g/abb; b/abaaa
6 a/b/taaba/tb/a/t; bbbgag-s; bbaabb-s
7 a/b/gbb/abaaa

HE

1 b-QLRS; b-G; e-s; g-G-s; g-SR-s
2 aa; gb; bb-s; ba; ag
3 aga/t; agb; bgb; aab/t/a; bab-s; g/aba; g/abb; bb/tb
4 t/eaab/a; b/a/gabb/a/e; agbb; agba; a/gagb/a; gbt/bb; agag-s
5 b/abbgb; agb/tb/ab/t/g; a/taaba/b; agaba-s; b/aaaga-s
6 agabab-s; gbbaaa-s; taagag; aabb/ag/ab/a; ag/agabb/a/e; g/agb/taab-s; t/aaaaba-s; ab/taaba-s
7 t/baagagb; t/aaabab/at-s; agab/tabb/a/t-s; agab/ag/b/aa/bb/a

HH

1 e-s; b-s; b-SQYT; b-DWN; t-FYAT; t-DHN; a-GDTY; a-VIAF
2 at/a; a/gb; bb; ag
3 g/abb; aab/t; aga/t; taa-s
4 a/gbaa/b; a/g/babb; baba-s; agbb; a/e/b/taat/a/b; g/a/bba/b/t
5 t/a/eaab/t; gbbbb; aagbb; e/baat/bb; b/aab/tbb-s
6 agbb/a/tb/tb/a; t/g/aaaat/bb-s; aagbba-s
7 agbbaaa/b-s; gbbbbbb-s
8 aabbgabb-s

Fig. 4: Loop conformational classes ordered by length. Positions with alternate conformations are indicated using a slash. Clusters with identical conformations are differentiated by a simple numbering scheme or their residue preferences. Classes of the "small" type are indicated by an "s".

pos.	conf.	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	g	1.3	0.1	11.8	2.6	0.4	35.8	7.4	0.5	2.6	1.7	0.4	16.9	0.2	2.0	2.3	8.7	0.5	0.2	2.2	2.5
2	g	0.5	0.4	2.0	1.4	0.1	79.5	4.5	0.0	0.8	0.1	0.1	4.9	0.0	0.4	0.3	4.3	0.4	0.1	0.2	0.3

	average	standard dev.	minimum	maximum
separation	5.32Å	0.42Å	4.50Å	6.43Å
angle	147.57°	14.13°	110.60°	167.26°
dihedral	153.85°	11.00°	126.80°	169.49°

Fig. 5: EggE class template including the probabilities of finding each of the 20 amino acid types at the two positions of the loop multiplied by 100 and the observed distributions of the three inter-vector relations.

pos.	conf.	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	e	0.7	0.0	4.8	2.0	0.3	78.8	1.1	0.0	0.9	0.3	0.1	4.5	0.0	1.0	2.0	1.9	0.9	0.3	0.0	0.7
2	a	4.3	0.1	18.9	3.8	0.9	7.7	4.4	0.8	3.4	1.0	0.5	9.9	9.2	6.0	3.4	14.0	7.9	1.1	1.0	1.8

	average	standard dev.	minimum	maximum
separation	5.28	0.37	4.74	6.17
angle	154.34	13.98	115.38	171.92
dihedral	158.80	10.84	128.22	172.47

Fig. 6: EeaE class template including the probabilities of finding each of the 20 amino acid types at the two positions of the loop multiplied by 100 and the observed distributions of the three inter-vector relations.

Some of the identified classes are found to have very restricted inter-vector separation, angle and dihedral angles while other show more flexibility in the relative disposition of the secondary structures that their member loops connect. The EaaagE class for example, which corresponds to the type I 4:4 β-hairpins identified by Sibanda and Thornton[4,6], is found to have standard deviations for its distribution of inter-vector separation, angle and dihedral angle of 0.32Å, 13.4° and 11.3° respectively. On the other hand, the HagbE class is found to have greater standard deviations for its distribution of inter-vector separation, angle and dihedral angle with values of 1.13Å, 34.7° and 30.5° respectively.

The sequence template of each loop class was tested against each of its member loops so as to identify loops whose sequences scored poorly. Most of such cases were confined to short loop classes of one or two residues in length 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 valine, isoleucine, alanine or phenylalanine and the second with a glycine or a charged or polar residue such as aspartate, threonine or lysine. In such cases the loop classes were split, with the HaH class being separated into the Ha-VIAF-H and Ha-GDTY-H classes.

3.3. Illustrating the prediction of loop conformation

During the modelling process, the information available for the prediction of the conformation of a loop consists of the loop's sequence and the spatial disposition of its bounding secondary structures. To identify the conformation of a loop, its motif was compared in turn to the templates of all the identified loop classes of matching length and bounding secondary structure types. Each comparison was evaluated in terms of both sequence match and fit of bounding secondary structure vectors. Sequence compatibility was measured in terms of a score derived from the probabilities of the loop's residues at each of the corresponding positions of the sequence template. The fit of each of a loop's inter-vector relations to the corresponding distribution observed among the member loops of a class was measured in terms of distance to the average and number of standard deviations. A loop's predicted conformation was selected as that of the class with the highest sequence score once those classes for which at least one of the loop's inter-vector relations diverged by more than three standard deviations from the average had been eliminated.

Sequence	Structure	Prediction	S_{seq}	Separation		Angle		Dihedral	
SWD	EabbE	none							
G	HbH	HbHs	77.2	1.3Å	1.25σ	-58.7°	-1.96σ	-20.3°	-0.68σ
HPEV	HtaabE	Ht/eaab/aE	14.5	1.1Å	0.75σ	-31.6°	-1.00σ	-14.0°	-0.44σ
Y	HtE	Hb-QLRS-E	0.9	-1.3Å	-1.32σ	-8.8°	-0.17σ	0.8°	0.01σ
DS	EaaE	Eaa-1-E	13.0	-0.2Å	-0.18σ	-26.8°	-0.88σ	-24.6°	-0.82σ
LID	EabaE	EabgEs	2.8	2.6Å	2.64σ	-0.4°	-0.01σ	97.0°	1.75σ
DDDLNIN	HbaabttaE	Ht/aaabab/atEs	7.0	-1.5Å	-0.78σ	2.7°	0.09σ	14.9°	0.50σ
APSENN	EtbbaaaH	Ea/b/taaba/tb/a/tH	6.3	0.5Å	0.30σ	-2.0°	-0.03σ	81.0°	1.11σ
KDYIN	HaaabaE	Ha/taaba/bE	4.6	2.0Å	2.02σ	26.7°	0.89σ	20.1°	0.67σ
YHPHK	HabbgaE	Ha/taaba/bE	5.8	-0.5Å	-0.53σ	-17.0°	-0.57σ	-29.9°	-1.00σ
TD	EabH	Ebb/tH	10.6	-1.9Å	-1.88σ	150.1°	2.05σ	152.6°	2.10σ
		EabH	9.0	0.0Å	0.03σ	88.3°	2.67σ	93.4°	3.11σ
NKIT	HgabbH	Ha/g/babbH	5.2	0.9Å	0.91σ	-81.3°	-1.48σ	-62.1°	-1.15σ
TFSLP	HagabaE	Ha/taaba/bE	2.9	-0.21Å	0.67σ	48.7°	1.62σ	26.8°	0.89σ
		HagabaEs	1.4	1.3Å	0.95σ	4.0°	0.13σ	-8.7°	-0.29σ
N	EbH	EtH	11.6	0.5Å	0.52σ	111.8°	2.01σ	131.8°	2.65c
		EbH	5.4	-0.8Å	-0.50σ	92.8°	1.35σ	115.9°	1.67σ

Fig. 7: Conformational class predictions for the loops of narbonin. In the three cases where the correct conformational class was not identified, it is shown after the predicted class. The difference between the inter-vector relations of each loop and the distribution in the loop classes is indicated in absolute terms and in number of standard deviations.

The conformational classes identified in this study are probably only a subset of all preferred loop conformations. This means that at the best, predictions for a maximum of 79% of loops might be possible. It is therefore important to provide not only a prediction but also some assessment of the validity of the prediction.

To illustrate the loop prediction protocol a protein structure, not included in the initial database of 255 structures, was selected, the plant seed protein narbonin an αβ-barrel solved at a resolution of 1.8Å (1NAR)[25]. Of the 14 loops with lengths of one to eight residues in narbonin (figure 7) eight were found to have a conformation belonging to one of the 162 identified classes when transitions between the transition (t) and helical (a) conformations or between the transition and β-sheet (b) conformations were allowed. Out of the eight, the correct conformation was identified solely based on sequence score in four cases with scores ranging from 4.6 to 14.5: HtaabE (figure 8); EaaE; HaaabaE; HgabbH. In another case, HbaabttaE, the elimination of loop classes with a divergence of at least three standard deviations between one of the loop's inter-vector relations and the classes' average, led to the prediction of the correct conformation with a sequence scores of 77.2. The conformation of the other three loops, EabH, HagabaE and EbH, were erroneously predicted with sequence scores of 10.6, 2.9 and 11.6 while the correct conformational classes had sequence scores of 5.4, 1.4 and 9.0 respectively. Of the six loops whose conformations did not belong to a loop class, one, EabbE, had no predicted conformation because the disposition of its bounding secondary structures was incompatible with all loop classes. Four, HtE, EabaE, EtbbaaaH and HabbgaE, had erroneously predicted conformations with sequence scores of 0.9, 2.8, 6.3 and 5.8, The last loop HbaabttaE had a predicted conformation with sequence scores of 7.0 which corresponded to the classes with the lowest r.m.s.d. Ht/aaabab/atEs. Considering a sequence score cut-off for a valid prediction in the range 3.0 to 4.0, six of the 14 loops would have been predicted with satisfactory conformations, four would have had no predicted conformations and four would have been predicted with erroneous conformations. Through this example it is possible to see how loop length affects sequence score, with shorter loop showing an overall trend toward higher scores. Additionally, several classes of short loops show similar residue preferences making it difficult to predict loop conformation. This is an expected result since for example only twenty different sequences are available to loops of one residue in length. On the other hand, a higher proportion of short loops cluster as compared to longer loops. These two facts indicate that loop conformation prediction should be most accurate for loops with intermediate lengths of three to six residues.

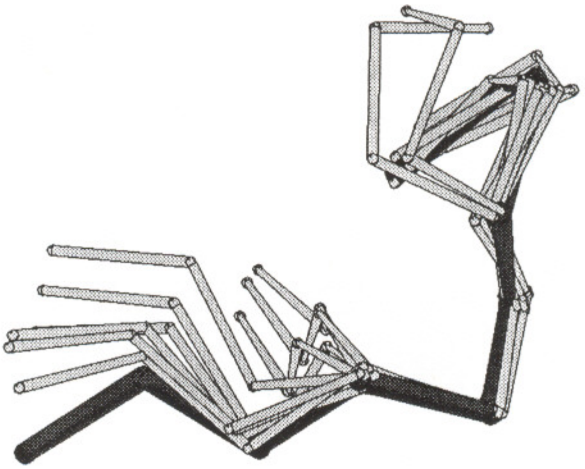

Fig. 8: Superposition of the HtaabE loop of narborin (dark grey) with the member loops of the predicted loop conformational class, Ht/eaab/aE (light grey). The backbone trace of the loops themselves plus three residues on either side are shown.

3.4. Further developments

The assignment of a secondary structure terminus can vary by one residue if a single hydrogen bond is broken or if one extra hydrogen bond is considered. This makes secondary structure termini susceptible to protein structure resolution, refinement protocol and local temperature factors. The same loop extracted from two different structure determinations of the same protein may differ in length and therefore be classified in different loop conformational classes. To take account of this, loop termini should be allowed to move somewhat both when identifying loop classes and when comparing a loop to a class template. Alternatively, compatible loop classes of differing length could be merged.

Sequence score shows a negative correlation to loop length and the general prediction cut-off of 3.0 to 4.0 should be replaced by a length-dependent cut-off. This should be defined from the distributions of sequence scores of both correct predictions and of erroneous predictions. To segregate efficiently correct predictions from erroneous predictions the r.m.s.d. distribution for each loop length studied has to be defined.

4. Conclusions

Of the 2083 loops of one to eight residues in length 66% clustered into 108 populated loop conformation classes, which included most of those previously described. When the less populated classes were also considered, 79% of the loops were found to cluster into a total of 162 classes. This suggests that although the distribution of the conformation of short to medium length loops might be continuous, it is certainly not random, some conformations being favoured over others. This is probably due to steric hindrance and the need for a sequence compatible with the different constraints, such as backbone conformation and side-chain burial. A greater number of conformations is available to the longer loops than to short loops which is reflected in the fact that a lower percentage of the longer loops were found to cluster as compared to the short loops. Having more conformational freedom may also mean that restrictive conformations, typical of many short loop classes, are more easily avoided.

The upper bound of the prediction rate is 79%, corresponding to the percentage of clustered loops in the studied database. The highest attainable prediction rate should be some what less than this since the identified clusters most certainly represent a lower percentage of all loops. In the case of short loops of one to two residues in length different conformational classes can have similar residue preferences making prediction difficult. Allowing for the prediction of the most similar loop class in cases where the loop to be predicted does not belong to a predicted class should on the other hand increase the prediction rate upper bound. The "short" loop classes contain less information about sequence preferences and acceptable inter-vector relations than the populated loop classes and should therefore identify fewer of the compatible loops.

The 14 examples taken from narbonin show a prediction rate of around 43% or 6 loops, with around 29% or 4 erroneous predictions and 29% or 4 none predictions. Although these figures, which are within the range found for other test cases, may seem low they should be viewed in the context of the observed mobility of many loop regions as shown by N.M.R. and X-ray temperature factors.

Acknowledgements

We would like to thank Dr. Chris Topham for the use of his substitution tables, 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 our friends and colleagues at the ICRF unit of structural molecular biology and Birkbeck College for their kind help.

L.E.D. is a recipient of a Human Capital and Mobility fellowship from the European Union, L.C. was funded at the time of this work by Dorian and S.D.R. is a recipient of an Imperial Cancer Research Fund bursary.

References

1 Jones, T.A. and Thirup, T. (1986). Using known substructures in protein model building and crystallography. EMBO J. **5**, 819-822.

2 Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L. and Sutcliffe, M. (1988).

Knowledge-based protein modelling and design. Eur. J. Biochem. **172**, 513-520.

3 Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S. and Blundell, T.L. (1993). Fragment ranking in modelling of protein structures. J. Mol. Biol. **229**, 194-220.

4 Sibanda, B.L. and Thornton, J.M. (1985). β-hairpin families in globular proteins. Nature **316**, 170-174.

5 Edwards, M.S., Sternberg, M.J.E. and Thornton, J.M. (1987). Structural and sequence patterns in the loops of βαβ units. Protein Engineering **1**, 173-181.

6 Sibanda, B.L., Blundell, T.L. and Thornton, J.M. (1989). Conformation of β-hairpin in protein structures. J. Mol. Biol. **206**, 759-777.

7 Efimov, A.V. (1991). Structure of α-α-hairpins with short connections. Protein Engineering **4**, 245-250.

8 Efimov, A.V. (1991). Structure of coiled β-β-hairpins and β-β-corners. FEBS Letters **284**, 288-292.

9 Srinivasan, N., Sowdhamini, R., Ramakrishnan, C. and Balaram, P. (1991). Analysis of short loops connecting secondary structural elements in proteins. Molecular conformation and biological interactions (ed. Ramakrishnan, C. and Balaram, P.), 59-73.

10 Sowdhamini, R., Srinivasan, N., Ramakrishnan, C. and Balaram, P. (1992). Orthogonal ββ Motifs in proteins. J. Mol. Biol. **223**, 845-851.

11 Sun, Z. and Blundell, T.L. (1995).The pattern of common supersecondary structure (motifs) in protein database. Proceedings of the twenty-eighth annual Hawaii international conference on system sciences.

12 Tramontano, A., Chothia, C. and Lesk, A.M. (1989). Structural determinants of the conformation of medium-sized loops in proteins. Proteins **6**, 382-394.

13 Bernstein, F.C., Koetzle, T.F., William, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. **112**, 535-542.

14 Overington, J., Johnson, M.S., Šali, A. and Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. R. Soc. Lond. B. **241**, 132-145.

15 Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. (1987). Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng. **1**, 377-384.

16 Šali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial constraints. J. Mol. Biol. **234**, 779-815.

17 Sowdhamini, R. and Blundell, T.L. (1995). An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. Protein Science **4**, 506-520.

18 Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**, 2577-2637.

19 Chou, K.-C., Nemethy, G., and Scherega, H.A. (1984). Energetic approach to the packing of α-helices. 2. General treatment of nonequivalent and nonregular helices. J. Am. Che. Soc. **106**, 3161-3170.

20 Kearsley, S.K. (1989). On the orthogonal transformation used for structural comparisons. Acta Cryst. A **45**, 208-210.

21 Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**, 783-791.

22 Fitch, W.M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science **155**, 279-284.

23 Richmond, T.J., and Richards, D.C. (1978). Packing of alpha-helices: geometrical constraints and contact areas. J. Mol. Biol. **119**, 537-555.

24 Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. **213**, 859-883.

25 Hennig, M., Pfeffer-Hennig, S., Dauter, Z., Wilson, K.S., Schlesier, B. and Nong, V.H. and (1995). Crystal-structure of narbonin at 1.8-angstrom resolution. Acta Cryst. D **51**, 177-189.