Identification of β-Sheet Motifs, of ψ-Loops, and of Patterns of Amino Acid Residues in Three-Dimensional Protein Structures Using a Subgraph-Isomorphism Algorithm†

Peter J. Artymiuk,‡ Helen M. Grindley,§ Andrew R. Poirrette,§ David W. Rice,‡ Elizabeth C. Ujah,§ and Peter Willett*,§

Krebs Institute for Biomolecular Research, Department of Molecular Biology and Biotechnology, and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

Received June 24, 1993⊛

This paper discusses the use of graph-theoretical techniques for the representation and searching of 3-D protein structures that are represented by labeled graphs. Two types of graph are considered. The nodes in the first type of chemical graph describe the secondary-structure elements and the edges the geometric relationships between pairs of these elements. The use of Ullmann's subgraph isomorphism algorithm with these representations permits the identification of all occurrences of all possible β-sheet motifs in a 114-protein subset of the Protein Data Bank. A binary notation is used to describe the parallel or antiparallel characters of adjacent strands in a sheet, and it is shown that very few of the possible types of sheet are found to occur in practice. Similar conclusions are obtained using a more detailed notation that includes the sheets' topological connectivities. The latter notation enables searches to be carried out for ψ-loops and allows the identification of 20 proteins that had not previously been known to contain this type of loop. The nodes in the second type of protein graph describe the amino acid residues in protein structures and the edges the geometric relationships between pairs of these residues. The use of the Ullmann algorithm with these representations permits the identification of the occurrences of patterns of residues. The identification process is illustrated by means of searches for the two aspartate groups that are known to be involved in the catalytic mechanism of the aspartic proteinases.

1. INTRODUCTION

Searching techniques for databases containing the amino acid sequences of proteins are well established, and the databases are widely used in drug design, biotechnology, and protein engineering, *inter alia*.[1] The increasing availability of three-dimensional (3-D) protein structures has spurred the development of search algorithms that can process the geometric information contained within a protein structure, and a very wide range of database techniques has been reported for this purpose,[2] including relational theory,[3] object orientation,[4] logic programming,[5] geometric hashing,[6] and dynamic programming,[7] *inter alia*. A continuing program of research in the University of Sheffield[8-16] is evaluating the use of graph-theoretical algorithms for searching 3-D protein structures, using methods that are derived from those that are widely used for the representation and searching of 3-D small molecules.[17-19] Graph-based approaches have also been described by other workers.[20-23]

In this paper, we discuss the use of a subgraph-isomorphism algorithm to identify those proteins in the Protein Data Bank[24,25] that contain a user-defined query pattern. Two types of patterns are considered. The first allows searches to be carried out for patterns of β-strand secondary-structure elements (although the methodology is equally applicable to motifs that contain α-helices), and the second for patterns of amino acid residues. Sections 2 and 3 describe an exhaustive enumeration that we have carried out of all of the *β-motifs* in a 114-protein subset of the Protein Data Bank, where we use the term β-motif to refer both to complete β-sheets and to partial β-sheets that are subsets of larger β-sheets. Section 4 reports the use of these techniques for the identification of ψ-loops[26] in 20 proteins that had not previously been known to contain this structural motif. In section 5, we describe how it is possible to carry out searches for residue-based patterns, illustrating the approach with searches for the two aspartate groups that are known to be involved in the catalytic mechanism of the aspartic proteinases. The paper concludes with a brief summary of our major findings.

2. REPRESENTATION OF B-MOTIFS

The bulk of this paper involves the representation and searching of *supersecondary-structure motifs*, i.e., patterns of secondary-structure elements in 3-D space. The nodes in the graph representation of a protein that we have adopted are vectors drawn along the major axes of the α-helix and β-strand secondary-structure elements in a protein. The edges describe the geometric relationships between pairs of these nodes; specifically, each edge is a three-part data element that contains the angle between two vectors, the distance of closest approach and the distance between their midpoints (although the midpoint distances are not used in the work reported below, and we consider motifs that contain only β-strands). The reader is referred to refs 10 and 19 for a detailed account of the way in which such protein graphs are generated; they provide a compact, precise representation of the arrangement of the secondary-structure elements in 3-D space that can be processed with great efficiency by a subgraph-isomorphism algorithm in just the same way as a connection table is processed in a conventional substructure-search system.[17] The algorithm we have used is that due to Ullmann,[27] as implemented in the program PROTEP (PROtein Topographic Exploration Program).[28] This program has been extensively tested by means of searches of the Protein Data Bank for many different query motifs.[9,10] The results demonstrate that the system provides an extremely efficient

* To whom all correspondence should be addressed.
† Presented at the Third International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, 6th-10th June 1993.
‡ Department of Molecular Biology and Biotechnology.
§ Department of Information Studies.
⊛ Abstract published in *Advance ACS Abstracts*, January 15, 1994.

0095-2338/94/1634-0054$04.50/0

means of identifying all occurrences of supersecondary-structure motifs: in some cases, occurrences are identified that had not been recognized previously; e.g., we have demonstrated a striking resemblance in the tertiary folds of the *Salmonella typhimurium* CheY chemotaxis protein and of the GDP-binding domain of *Escherichia coli* (*E. coli*) elongation factor TU (EF TU).[11]

Here we categorize all of the β-motifs that occur in a 114-structure subset of the Protein Data Bank. The data set used in this work was derived from the February 1990 release of the Protein Data Bank, which contains 554 sets of coordinates. However, many of the proteins are represented by more than one structure: in such cases, only the highest resolution structure was selected. In some cases a decision had to be made on the inclusion or omission of a particular β-strand in a protein. Extended β-strands interspersed by short regions of indeterminate secondary structure were generally included if they were known to form hydrogen bonds with a neighboring strand in the sheet. Conversely, short strands which did not fall within hydrogen-bonding distances were almost always omitted, especially if they occurred at the edge of a sheet. The resulting data set contained 114 proteins, each of which had one or more β-motifs containing three or more strands: there were 190 such sheets *in toto*. The Protein Data Bank codes for these proteins are listed in the left-hand column of Table 1.

The parallel/antiparallel nature of the motifs, referred to subsequently as their *alignments*, are described using a binary notation, in which each β-strand in a motif is assigned a 1 or a 0. The first strand in the β-motif is always denoted by 1 and subsequent strands are assigned the values of 1 or 0, depending upon whether they are parallel or antiparallel, respectively, to the first strand. The notation thus represents the relative alignment of the β-strands in the order that they occur in the β-motif (and not their sequence order). In general, for a β-sheet consisting of N strands, there are $2^{(N-1)}$ possible β-motifs, although this total is reduced by considerations of symmetry. For example, the five-stranded motifs 10111 and 11101 are identical, since one is merely the reverse of the other: in such cases, the string with the numerically lower value is chosen to represent the motif (so that this particular five-stranded motif would always be denoted by 10111). Another example is provided by the set of three motifs 100, 101, 110, and 111 that can be generated from a three-stranded β-sheet; however, only three of these are unique, since 100 can be obtained from 110 (and *vice versa*) by a simple inversion of the relative strand alignment.

Queries were generated for all of the possible β-motifs that can be derived from sheets containing between 3 and 15 strands. These queries were then used to search the 114-structure data set described above. A query pattern consists of the angles and distances for neighboring strands and the distances to second and third neighbor strands. No angular value is set for the next-nearest neighbors because of the common phenomenon of twist occurring among the strands in β-sheets, which makes it very difficult to estimate the characteristic angular values for strands other than nearest neighbors. Consequently, an interstrand torsion angle of −25° is assigned to neighboring parallel strands and an angular value of +155° is assigned to neighboring antiparallel strands. Generalized closest-approach distances of 4.5, 9.0, and 13.5 Å define first-, second- and third-neighbor strands. Longer range distances are not defined as this would impose a restraint on the overall curvature of the sheet, and the degree of twist between neighboring strands along a sheet is seemingly unpredictable and variable from one globular protein to the next. The sequence order of the strands comprising a sheet was ignored in these initial searches, so as to focus on the strands' alignment in the sheet. A typical search matrix, that for the four-strand motif 1011, is shown in Table 2.

The binary notation provides a concise representation of the parallel/antiparallel nature of a motif, but takes no account of the motif's *connectivity*, i.e., the form of the connection between pairs of adjacent β-strands in a sheet. There are two main types of connection: in a *hairpin* connection the backbone chain re-enters the sheet on the same side as it left, whereas it re-enters on the opposite side in a *crossover* connection. In both cases the backbone loop between any two β-strands is of variable length and can take any conformation, except that it may not include another β-strand that is part of the same β-sheet. The two connected β-strands may be nearest neighbors in the β-sheet, or one or more other strands may lie between them. We have defined the connectivities using the notation of Richardson,[29] where each connection is described in terms of the number of strands it moves over in the sheet and in which direction, starting from the N-terminal strand, and where an "x" is added for crossover connections. Thus, a "+1" is a hairpin and a "+1x" a crossover connection between nearest neighbor strands; a "+2" and a "+2x" are hairpin and crossover connections, respectively, that skip past one intervening strand in the β-sheet, and so on. The absolute values of the signs are not meaningful since the β-sheet can be "read" upside down, e.g., the five-stranded sheet of actinidin (represented here by 2ACT) can be described either as +4,−3x,+1,+1, or as −4,+3x,−1,−1. This particular sheet is illustrated in Figure 1. Once a motif has been identified by an alignment search, the corresponding connectivity information is extracted automatically from the Protein Data Bank.

The outputs from the alignment and connectivity searches are detailed in Table 1, which thus provides an exhaustive catalogue of all β-motifs in the 114 proteins that we have selected for analysis.

3. SEARCHING FOR β-MOTIFS

3.1. Analysis of Alignments. Table 3 summarizes the alignments that were actually identified in the searches, and compares this with the total numbers of alignments that are possible for sheets containing N strands.

It will be seen that a vast number, 18 432, of β-motifs are feasible, for $3 \leq N \leq 15$, but that only 96 (just over 0.5%) are actually found to be present in the Protein Data Bank structures. In general, the smaller motifs are retrieved from a greater number of proteins than the larger motifs (as would be expected as many of the smaller motifs are subsets of larger β-motifs). Small motifs that are less prevalent in the Protein Data Bank structures occur less frequently as subsets of larger motifs, or in some cases are completely non-existent. All of the possible β-motifs occur at least once up to and including the five-stranded motifs. However, as shown in Table 3, one-quarter of all of the possible six-stranded motifs are not found, and the fraction of undetected motifs rises extremely rapidly thereafter. The small number of structures in the Protein Data Bank is clearly responsible for many of the absences; there is, however, a clear pattern to the results in that there are at least two well-defined classes of motif that are noticeably underrepresented. The first disfavored group seems to comprise adjacent units of three or more parallel strands with each of the units in antiparallel alignment, e.g., the motif 111000. The second disfavored group appears to be composed of a unit of three or more strands that are parallel to each other, immediately adjacent to a unit of three or more strands

Table 1. Dictionary of All Occurrences of β-Motifs in a Set of 114 Proteins in the Protein Data Bank[a]

protein	alignment	connectivity
1AAT	1011111	+6x,−1x,−1x,−1x,−2,+1
1ABP	11111	−1x,+2x,+1x,+1x
	111110	+1x,−2x,−1x,−1x,+1
2ABX	101	+1,−2x
2ACT	1010	+2x,+1,−2x
	10100	+4,−3x,+1,+1
1ACX	101	+1,+1
	101	−1,+2x
8ADH	101	+1,−2x
	111111	−1x,−1x,+3x,+1x,+1x
3ADK	11111	−2x,+1x,+2x,+1x
1ALC	101	+1,+1
2ALP	101	+1,+1
	101010	−1,−1,+3,+1,+5
	1010010101	+5x,−1,−3x,+1,+1,+6,−1,−1,−1
2APE	1010	+1,+2x,−1
	1010	−1,−1,+3
	1010	+1,−2x,−1
	1010	+1,−2x,−1
	10010	+2,+1,−2x,−2
	101010	−2x,+1,−4x,+2x,−1
	101100	+1,+1,+2,+1x,−2
7API	101010	+5,−1,−1,−1,−1
	110101	−1x,+2,+3,−1,−1
3APP	1010	+1,−2x,−1
	1010	+1,−2x,−1
	1010	−1,−1,+3
	101010	−2x,+1,−4x,+2x,−1
	110010	+2,+3x,−1,−1,−2
	101100	+1,+1,+2,+1x,−2
3APR	1010	+1,−2x,−1
	101100	+1,+1,+2,+1x,−2
	110010	+2,+3x,−1,−1,−2
	1010101	−1,+3,−1,+4x,−2x,+1
4ATC	101	+1,+1
	1111	+1x,−2x,−1x
	111111	−1x,−1x,+3x,+1x,+1x
2AZA	101	+1,+1
	11010	+3x,+1,−2x,−1
1B5C	11011	+3x,−1,+2,−3x
3BCL	1010	+1,+2x,−1
	10110101010101010	−1,−3,−1,−1,−1,−1,−1,−2x,+1,+7, +1,+3x,+2x,−1
1BDS	101	+2x,−1
1BLM	10101	−1,+4x,−1,−1
1CA2	1010100100	+8x,−1,−1,−1,−1,+5,−7x,+1,−2x
8CAT	10101010	+1,+1,+1,+1,+1,+1,+1,+1
2CBH	101	+2x,−1
1CBP	1010	+1,−2x,−1
	1010	+1,+2x,−1
6CHA	101010	−5,+1,+3,−1,−1
	101010	−5,+1,+3,−1,−1
1CMS	101	+1,+1
	1010	+1,+2x,−1
	1010	−1,+2x,+1
	1100	+2,+1x,−2
2ENL	10111111	+1x,+1x,+1x,+1x,+1x,+1x,+1x
6EST	101010	−5,+1,+3,−1,−1
	101010	−5,+1,+3,−1,−1
1ETU	111110	−2,+1,+2x,+1x,+1x
1FCR	11111	−1x,+2x,+1x,+1x
1FXB	1111	−1x,+2x,+1x
3GAP	1010	+1,+2x,−1
	10101	+3,−1,+2x,−3
2GBP	11111	−1x,+2x,+1x,+1x
	111110	−1x,+2x,+1x,+1x,+1
2GCR	1010	−1,+2x,+1
	1010	−1,+2x,+1
2GN5	101	+1,+1
1GOX	11111111	−1x,−1x,−1x,−1x,−1x,−1x,−1x
1GP1	1000010	−1,+4,−1x,−1x,+3,+3,+1
1GD1	1011111111	−1x,−3x,+1,+1,+3x,+1x,+2x,−1x
	1101001	−2,+1,+2,+3x,−1x,−1
3GRS	1101	+3x,−1,−1
	1101	+1x,+1,+1
	11111	+2x,+1x,+1x,−3x
	10101	+1,+3,−1,−1
	1111110	−1x,−1x,+3x,+2,−1
1HIP	1011	+1,+1,+1x
2YHX	101	+1,+1
	10111	−1,−1,+3x,+1x
	101000	+3,−1,−1,+3x,+1x
1HLA	10101010	−1,−1,−1,+4x,+1,+1,+1
3HMG	111	+1x,+1x
	1010	+2x,+1,−2x
	1010	+2x,+1,−2x
	10101	+1,+1,+1,+1
1HOE	101	−1,+2x
	101	+1,+1
3HVP	101100101	−1,−2,−3x,+1,+6x,+1,−6x,+2
2IGN	1010101	+1,+3,−1,−1,+3,+1
	101010101	+1,+5,−1,−1,−1,−1,+5,+1
4INS	101	+2x,−1
2PKA	1010101	−1,−1,+3,+1,−6x,+1
	1010101	+2x,−1,+5,−3,+1,+1
2LBP	11111	−1x,+2x,+1x,+1x
	1111101	−1x,+2x,+1x,+1x,+1,+1
8LDX	111	−1x,−1x
	111111010	−1x,−1x,+3x,+1x,+1x,+1,+1,+1
2LIV	11111	−1x,+2x,+1x,+1x
	1111101	−1x,+2x,+1x,+1x,+1,+1
3LYM	1010	+1,+1,+1
4MDH	111111010	−1x,−1x,+3x,+1x,+1x,+1,+1,+1
2MEV	1010	+2x,+1,−2x
	10101	+2x,+1,−2x,+3
2MON	101	+1,−2x
	10101	+1,+1,−4x,+1
1NXB	101	−1,+2x
2OVO	101	−1,+2x
2P21	111110	+2,−1,−2x,−1x,−1x
2PAB	1001	−1,+2,+1
	1010	−3,+1,+1
1PHH	101	+1,+1
	1011	−3x,+1,+1
	11111	+1x,+1x,−3x,−1x
	1001010	−1,+5x,−3,+1,+1,+2x
1PHY	10101010	−1,−1,−1,+6,−1,−1,−1
2PLV	10101010	−1,+3,−5,−1,+7,−5,+3
1PRC	101	+1,+1
	101001	−2x,−1,−1,+3,+1
2PRK	1111111	−2x,+1x,+2x,+1x,+1x,+1x
1PSG	101	+1,+1
	1010	+1,+2x,−1
	1010	+1,+2x,−1
	1010	+1,−2x,−1
	101010	−1,+4x,−2x,+1,−4x
	101100	+1,+1,+2,+1x,−2
	110010	+2,+3x,−1,−1,−2
1PTE	10001	+1,+1x,+1x,+1
1PYK	11111111	+1x,+1x,+1x,+1x,+1x,+1x,+1x,+1x
2PYP	1011	+2,−1,+2x
1RHD	11111	+3x,+1x,−2x,−1x
4RHV	1010	+1,+1,−3
	1010	+1,+1,−3
1RN3	10100	−1x,−1,−1,−1
3RP2	101010	−1x,+3,+1,+1,−3
2RSP	1011001010	−1,+2,−3x,+1,+6x,+1,−6x,+2
3RUB	11111111	+1x,+1x,+1x,+1x,+1x,+1x,+1x,+1x
4RXN	101010	−1,−1,−1,−1,−1
1SBC	1111111	−2x,+1x,+2x,+1x,+1x,+1x
4SBV	1010	+2x,+1,−2x
	1010	+2x,+1,−2x
4SGA	1010	+1,+2x,−1
	10101	−1,−2x,+1,+3
	11010	+3x,+1,−2x,−1
1SN3	1010	+3,−1,−1
2SNS	10101	−1,−1,+3,+1
2SOD	10101010	+1,+1,+1,+3,−1,−1,+3,−1
1SRX	10111	−2x,+1x,−2,−1
2SSI	10101	−1,−1,+3,+1
1STV	1010	+2x,+1,−2x
	1010	+2x,+1,−2x
3LZM	101	+1,+1
2TAA	101	+2x,−1
	10101010	+1,+1,+5,−3,+1,+1,−3
	11111111	+1x,+1x,+1x,+1x,+1x,+1x,+1x,+1x
2TBV	1010	+2x,+1,−2x
	101101	−1,+5,−3,+2x,−1
1TEC	1010	+2x,+1,−2x
	1011	+3x,−1x,−1
	1111111110	−1x,−1x,−1,+4,+1x,+1x,+1x, +2x,−1x
1TLD	101010	−5,+1,+3,−1,−1
	101010	−5,+1,+3,−1,−1
1THI	10010	−1,+3x,+1,−2x
	101010	+3,−1,−1,+3,+1
2TIM	11111111	+1x,+1x,+1x,+1x,+1x,+1x,+1x,+1x
3TMN	10001	+1,+1x,+2,−1

Table 1 (continued)

1CMS	100101	−1,+2,+4,+1,−4	1PPD	1010	−2x,1,+2x	3TMN	10100	−1,−1,−1,+4
	101010	−1,+4x,−2x,+1,−4x		10100	+4,−3x,+1,+1	2TMV	1010	+1,−2x,−1
3CNA	101010	−1,−1,+4x,+1,+2x	2PAZ	10011010	+1,−2,+5,+1,−7,+5,−1		1010	+1,−2x,−1
	1010101	−1,−1,+4x,+2x,−1,−2x	6PCY	10100110	−1,+2,−5,−1,+7,−5,+1	1TNF	10101	+1,+2x,+1,−2x
4CPA	10111100	+1,+2,−1x,+2x,+2,+1x,−2	4PFK	1000001	−1x,−1,+3,+1x,+1x,+1		10101	+1,+2x,+1,−2x
1CRO	101	+1,+1	2PGK	101	+1,+1	1TON	1010101	+1,+1,−3,−1,+6x,−1
1CTF	101	−1,+2x		111111	−1x,−1x,+3x,+1x,+1x		1010101	+1,+1,−3,−1,+6x,−1
1CTX	101	−1,+2x		111111	−2x,−1x,−2x,+1x,+3x	2TS1	111110	+3,+1x,+1x,−3x,−1x
1DPI	11011	+3x,−1,−1,+3x	1PGI	1111	+1x,+1x,+1x	1UBQ	10010	−1,+3x,−2x,−1
	1010101	+2x,−1,−2x,		111111	−1x,+2x,+1x,+2x,−1x	3WGA	1010	+1,−3,+1
4DFR	11111101	−2x,−1x,−1x,+3x,+2x,+2x,−1	3PGM	101111	+2x,+1x,−2x,−2,−1	1WSY	11111111	+1x,+1x,+1x,+1x,+1x,+1x,+1x
5EBX	10101	+1,+3,−1,+2x	1PHH	101	+2x,−1	2XIA	11111111	+1x,+1x,+1x,+1x,+1x,+1x,+1x

[a] The alignment describes the parallel/antiparallel nature of the constituent β-strands, while the connectivity describes the form of the connection between adjacent β-strands in a sheet.

Table 2. Search Matrix for the β-Motif 1011[a]

strand	1		2		3		4	
1	0	0.0	155	4.5		9.0		13.5
2	155	4.5	0	0.0	155	4.5		9.0
3		9.0	155	4.5	0	0.0	−25	4.5
4		13.5		9.0	−25	4.5	0	0.0

[a] The first and second figure in each element of the table correspond to the interline angle and to the closest-approach distance, respectively.

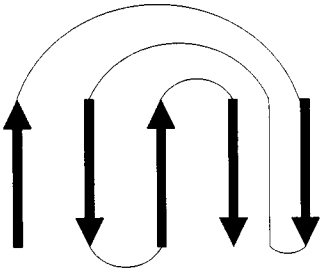


Figure 1. Diagrammatic representation of the five-stranded β-sheet in actinidin (2ACT). The alignment of this sheet is of type 10100, and its connectivity is of type +4,−3x,+1,+1.

Table 3. Potential and Actual Occurrences of β-Motif Alignments

no. of strands (N)	no. of possible motifs (2^{N-1})	no. of possible unique motifs	no. of actual motifs found
3	4	3	3
4	8	6	6
5	16	9	9
6	32	20	15
7	64	36	17
8	128	72	18
9	256	142	11
10	512	288	8
11	1024	576	5
12	2048	1152	1
13	4096	2304	1
14	8192	4608	1
15	16384	9216	1
totals	32764	18432	96

Table 4. List of the Most Prominent N-Strand β-Motifs

	type of protein			
N	all	β	α + β	α/β
3	101	101	101, 100	111
4	1010	1010	1010, 1011	1111
5	10101, 11111	10101	10101, 10100	11111
6	101010, 111111	101010	101010	111111
7	1010101, 1111111	1010101	1010101	1111111
8	10101010, 11111111	10101010	10101010	11111111

in which the strands are antiparallel to each other, e.g., 1000101. Although generally disfavored, a few examples of this class of motif were found in the larger proteins, e.g., the nine-strand motif 111111010 occurs in both lactate dehydrogenase (8LDX) and malate dehydrogenase (4MDH).

Just as some classes of motif are disfavored, so other classes appear to occur more frequently than would be expected. The most common motifs for each value of N are listed in the second column of Table 4, headed "all". No motifs are listed for $N > 8$, since most of the motifs here occurred with approximately equal frequencies and since the total occurrences were far too low to provide accurate statistics. For the smaller motifs, the 101 antiparallel motif is consistently the most frequent, and then the 111 parallel motif. In addition, we have classified the domains containing each of the β-motifs into the three classes of β, α + β, and α/β as defined by Levitt and Chothia.[30] Pure β-sheet domains are almost exclusively

Table 5. Occurrence Frequencies of Each Connection Type in β-Sheets Consisting of Three or More β-Strands

connection		parallel sheets	antiparallel sheets	mixed sheets	total
10	±1		229	130	359
11	±1x	107		66	173
1*0	±2			38	38
1*1	±2x	18	72	26	116
1**0	±3		38	9	47
1**1	±3x	6		29	35
1***0	±4			8	8
1***1	±4x		11		11
1****0	±5		14	6	20
1****1	±5x			2	2
1*****0	±6			1	1
1*****1	±6x		3	5	8
1******0	±7		1	3	4
1******1	±7x			1	1
1*******0	±8				
1*******1	±8x			1	1
totals		131	368	325	824

β-sheet; α + β domains are principally composed of α-helices and β-sheets that are segregated along the polypeptide chain; while α/β domains essentially comprise α-helix and β-sheet alternating along the polypeptide chain such that the resulting tertiary structure fold is a sandwich of alternate layers of α-helix and β-sheet structures. Analysis of the most frequently occurring motifs in each of these three classes, as summarized in Table 4, reveals that the β-sheet domains are characterized by large numbers of antiparallel sheets, whereas the α + β domains are chiefly composed of antiparallel sheets followed by repeating units of the 100 motif, a finding that is in agreement with previous studies.[31,32] α/β-proteins are characterized by the presence of parallel β-motifs. An obvious example of this behavior is provided by the eight-stranded β-sheet, in which the parallel strands are arranged close together to form a β-barrel.

3.2. Analysis of Connectivities. Inspection of the connectivity data in Table 1 reveals a total of 102 distinct β-motifs. It is of interest to compare these with the 36 such motifs identified by Richardson in her 1977 survey.[29] In the present study, 34 of Richardson's topologies were identified. The two remaining motifs are a four-stranded mixed sheet in lactate dehydrogenase (8LDX) (+1,+1,−3,+4x) and a six-stranded mixed sheet in papain (1PPD) (+2x,+1,+1,−4x). The present study identified two β-sheets in lactate dehydrogenase: these consisted of three (−1x,−1x) and nine (−1x,−1x,+3x,+1x,-1x,+1,+1,+1) β-strands, respectively. Neither of these sheets correspond remotely to Richardson's assignment. Two sheets were identified in papain; however, these were a four-stranded antiparallel sheet (−2x,−1,+2x) and a five-stranded mixed sheet (+4,−3x,+1,+1), both of which are significantly different from Richardson's assignments. These differences may be due to improved sets of coordinates becoming available since the previous survey or from the use of the Kabsch–Sander algorithm for the identification of the secondary-structure elements. Thus, the exhaustive nature of the graph-matching algorithm used here has allowed us to identify over 60 new β-motifs in the Protein Data Bank.

Table 5 lists the occurrence frequencies for each of the individual connection types that make up the β-sheets. The table reveals the presence of just three connection types in parallel β-sheets, whereby almost four-fifths (107 out of 131) of the connections are ±1x. Conversely, all possible connection types are found in mixed sheets in varying frequencies (with the exception of the ±4x crossover and the ±8 hairpin). Antiparallel β-sheets exhibit relatively fewer connection types.

Table 6. Occurrence Frequencies for Consecutive Pairs of Connection Types in β-Motifs Consisting of Three or More β-Strands

connection pair		parallel sheets	antiparallel sheets	mixed sheets	total
101	+1,+1		63	41	104
100	+1,+1x			16	16
10*1	+1,+2			7	7
110	+1,−2			13	13
10*0	+1,+2x		36	7	43
101	+1,−2x		56	9	65
10**1	+1,+3		19	4	23
1*10	+1,−3		38	4	42
10**0	+1,+3x			7	7
1*01	+1,−3x			17	17
1**10	+1,−4			5	5
1**01	+1,−4x		13		13
1***10	+1,−5		10	4	14
10****1	+1,+5		5	2	7
1***01	+1,−5x			2	2
1****10	+1,−6			1	1
10*****1	+1,+6			1	1
1****01	+1,−6x		6	2	8
10*****0	+1,+6x			4	4
10******1	+1,+7			2	2
1*****10	+1,−7		1	2	3
1*****01	+1,−7x			1	1
1******01	+1,−8x			1	1
111	+1x,+1x	63		19	82
11*0	+1x,+2			7	7
011	+1x,−2			7	7
11*1	+1x,+2x	20		9	29
111	+1x,−2x	14		9	23
11**0	+1x,+3			2	2
1*00	+1x,−3			1	1
11**1	+1x,+3x	6		6	12
1*11	+1x,−3x	5		7	12
0**11	+1x,−4			1	1
1****11	+1x,−6x			1	1
1*0**1	+2,+3			1	1
1*0***1	+2,+4			1	1
1**1*0	+2,−5			2	2
01*0	+2,−3x			1	1
0***1*0	+2,−6x			2	2
1*1*0	+2x,+2			3	3
1*1*1	+2x,+2x			1	1
01*1	+2x,−3		2	1	3
1*1**1	+2x,+3x			2	2
1*1***1	+2x,+4x		1		1
1*1*1	+2x,−4x		4		4
10*1	+3x,−2			1	1
01**1	+3x,−4			2	2
1*1****0	+5,−7		1	2	3
0*1****0	+5,−7x			1	1

This is entirely expected due to the stricter pattern constraints that apply to purely antiparallel topologies, whereby connections which do not promote antiparallelism are excluded but may well be included in mixed sheet topologies. For instance, the ±4 hairpin (1***0) can never be part of a purely antiparallel sheet as the resulting strand alignment is unsuitable for this topology; for the same reason, it could also never be part of a purely parallel β-sheet. Moreover, parallel sheets consist only of crossover connection types. In the case of mixed sheets, the ±4 hairpin is readily incorporated into the β-sheet topology, e.g., −1,−1,−1,+4 in thermolysin (3TMN) and +4,−3x,+1,+1 in actinidin (2ACT) and papain (1PPD).

Table 6 lists the occurrence frequencies for consecutive pairs of connection types. The +1,+1 connections are by far the most common connection pair (104 occurrences), followed closely by the +1x,+1x connection pair (82 occurrences). Not surprisingly these strand connectivities primarily promote pure-antiparallel sheet and pure-parallel sheet topologies, respectively, thereby confirming the previous analysis of the predominant β-motif patterns. The remaining types of

connection pairs are distributed unevenly among the antiparallel and mixed β-sheets; overall, more types of connection pairs are found in mixed-sheet topologies than in any other β-sheet topology.

From this study, parallel β-sheets exhibit the least assortment of connection types (3) and connection pair types (5). The parallel sheet topologies are primarily composed of the +1x,+1x (63 occurrences) connection pair and then, in progressively lower frequencies, the +1x,+2x (20 occurrences), +1x,–2x (14 occurrences), +1x,+3x (6 occurrences), and +1x,–3x (5 occurrences) connection pairs. The limited variety of topological patterns in parallel sheets implies that the occurring sheet topologies are not merely a random arrangement but are subject to constraints of simplicity of folding and of structural stability, whereby all the connectivities are relatively short, (<±3x). Parallel sheets are present in a wide range of proteins, including those as functionally diverse as L-arabinose binding protein (1ABP; –1x,+2x,+1x,+1x) and adenylate kinase (3ADK; –2x,+1x,+2x,+1x). The former is a periplasmic binding protein isolated in gram-negative bacteria with specificity for sugars, amino acids, and vitamins, whereas the latter is a nucleotide-binding protein. Overall, these parallel sheets are located in differing structural environments within each of their respective proteins; yet, despite this, their local sheet topologies are quite similar, consisting of a combination of ±1 and ±2 connections. It thus appears that these parallel β-sheets isolated from the Protein Data Bank portray a rigorous repetition of limited connectivities, which implies that any parallel β-sheet structures would resemble existing topologies, both in their alignment and in their connectivity, irrespective of whether or not the parent structures are related. For instance, in the so-called α/β barrels, the ±1x,±1x,... topology is the most stable pattern exhibited by all members of the group. Since all crossover connections should be right-handed,[33] and since they cannot go down the center of the barrel, because of structural constraints, the repetitive ±1x topology is a very stable feature.

Distinctive topological patterns can also be readily identified in the ten different types of antiparallel sheets detailed in Table 5. With a wider range of connection types present, many more patterns occur than with the purely parallel sheets. The most prevalent topologies occur in four-stranded sheets: +1,+2x,–1 (8 occurrences), +1,–2x,–1 (10 occurrences), and +2x,+1,–2x (11 occurrences). Other distinctive topologies are the simple up-and-down β-meander pattern[34] of +1,+1+,+1,... (which is found, for example, in 3LYM, 4RXN, and 1CAT), and the "Greek key" pattern of +3,–1,–1,+3 or just +3,–1,–1 (which is found, for example, in 3GRS, 2SNS, 2SSI, 4RHV and 1THI). Indeed the ±1 connection is by far the predominant connectivity in antiparallel sheets, followed by the lesser frequencies of the ±2 and the ±3 connections. This observation is reinforced by Table 6, which reveals the +1,+1 pair as the most frequently occurring connection pair, followed by the +1,–2x pair and then by the +1,–3 and +1,+2x pairs. Longer connectivities, i.e., four or more intervening strands between sequential β-strands, are observed in antiparallel sheets, albeit with quite low frequencies of occurrence.

Mixed sheets can accommodate a combination of parallel and antiparallel β-motifs within a single sheet, and they thus allow a range of connection types much wider than the "pure" β-sheets. Nonetheless, around three-quarters of the 26 different mixed-sheet topologies occur only once, and none of the other motifs occur more than three times. Longer connectivities are also more common among mixed sheets, e.g., one occurrence of +8x and –7x in 1CA2 and three occurrences of ±7 in 2PAZ, 6PCY, and 3BCL.

Table 7. ψ-Loops in the Protein Data Bank[a]

protein	no. of ψ-loops	strand connectivities	residue nos.
endothiapepsin (4APE)	4	+2,+3,–1,–1,–2	196–303
penicillopepsin (3APP)	4	+2,+3x,–1,–1,–2	31–124
rhizopuspepsin (3APR)	4	+2,+3x,–1,–1,–2	212–300
pepsinogen (1PSG)	2	+2,+3x,–1,–1,–2	28–123
elongation factor TU (1ETU)	1	+2,–1	9–81
pyrophosphatase (1PYP)	1	+2,–1	119–159
thermolysin (7TLN)	1	+2,–1	98–124
α-antitrypsin (6API)	1	–1x,+2	110–195
carboxypeptidase A (5CPA)	1	+2,–1x	47–110
	2	+2,+1x,–2	190–271
glyceraldehyde 3-phosphate dehydrogenase (1GD1)	2	–2,–1,+2	169–245
chymosin (2CMS)	2	+2,+1x,–2	31–126
HIV proteinase (3HVP)	2	+2,+3x,–1,–1,–2	22–79
rous sarcoma virus protein (2RSP)	2	+2,+3x,–1,–1,–2	34–110
pepsinogen (1PSG)	2	+2,+1x,–2	209–303
endothiapepsin (4APE)	2	+2,+1x,–2	29–123
penicillopepsin (3APP)	2	+2,+1x,–2	208–299
rhizopuspepsin (3APR)	2	+2,+1x,–2	30–125
pseudoazurin (2PAZ)	1	+1,–2	1–35
plastocyanin (6PCY)	1	+1,–2	1–32
P21 protein catalytic domain (2P21)	1	+2,–1	1–58

[a] Those from endothiapepsin to glyceraldehyde 3-phosphate dehydrogenase have been reported previously by Hutchinson and Thornton.

4. ANALYSIS OF ψ-LOOPS

ψ-loops were first identified in the aspartic proteinase family[26] and are characterized by a "+2" hairpin connection between two sequentially adjacent strands in the same β-sheet. A single ψ-loop is composed of three β-strands: the two outer loops are capable of hydrogen bonding to each other, but they allow the third strand to intervene between them. There are four types of ψ-loop.[35] Type 1 is the simplest ψ-loop and is composed of three consecutive strands in a +2,–1 topology. A type 1X ψ-loop is characterized by a +2,–1x topology, with a crossover connection instead of the –1 hairpin. Chain reversals produce type 1′ and type 1X′ ψ-loops, which are related to types 1 and 1X and which have the topologies +1,–2 and +1x,–2, respectively.

The connectivity results detailed previously were analyzed to identify all occurrences of ψ-loops in our subset of the Protein Data Bank, and the resulting occurrences of the motif were checked visually using the FRODO program. A list of the proteins found to contain ψ-loops and their associated strand connectivities is shown in Table 7.

The first part of Table 7, down to and including glyceraldehyde 3-phosphate dehydrogenase (1GD1), contains the ψ-loops that have been identified previously.[35] The simplest ψ-loop (type 1) was found in just three proteins: elongation factor TU (1ETU), thermolysin (7TLN), and pyrophosphatase (1PYP). A type 1X ψ-loop was identified in the N-terminal region of carboxypeptidase A (5CPA) and its reverse topology, type 1X′, in α_1-antitrypsin (6API). Paired ψ-loops were found in glyceraldehyde 3-phosphate dehydrogenase, carboxypeptidase A, and four members of the aspartic protease family. Glyceraldehyde 3-phosphate dehydrogenase has the topology –2,+1,+2 and consists of a type 1 single ψ-loop, coupled, by means of a fourth strand added to the C-terminal end of the sheet, to a type 1X ψ-loop at the same end of the sheet. The C-terminal region of carboxypeptidase A also has one type 1

Table 8. Additional ψ-Loops Identified in the April 1992 Release of the Protein Data Bank

protein	no. of ψ-loops	strand connectivities	residue nos.
cholesterol oxidase (1COX)	1	+2,−1	283–473
mannose-binding protein (1MSB)	1	+2,−1	180–208
mesentericopeptidase (1MEE)	1	+2,−1	197–268
chloramphenicol acetyltransferase (1CLA)	1	−1,+2	89–151
flavocytochrome (1FCB)	1	−1,+2	25–85
fructose 1,6-biphosphatase (1FBP)	1	−1,+2	93–115
scorpion neurotoxin (1NRD)	1	−1,+2	263–296
isoenzyme 3–3 glutathione-5-transferase (1GST)	1	−1x,+2	2–64
acetylcholinesterase (1ACE)	1	+2,−1x	95–147
serine carboxypeptidase (2SC2)	1	+2,−1x	31–95
leucine aminopeptidase (1LAP)	1	+2,−1x	227–301
	2	+2,+1x,−2	352–450
purine nucleoside phosphorylase (2PNP)	2	+2,+1x,−2	110–241

and one type 1X′ ψ-loop; however, they occur at opposite ends of the sheet and thus give rise to a different overall topology (+2,+1x,−2). Coupled type 1–type 1X′ ψ-loops were also identified in the aspartic proteases. However this ψ-loop pair is created by the insertion of three strands between the second and third strands of a single type 1 ψ-loop. The resulting topology is +2,+3x,−1,−1,−2 and is found in the N- and C-terminal domains of endothiapepsin (4APE), penicillopepsin (3APP), and rhizopuspepsin (3APR). Pepsinogen (1PSG) has just one domain containing a similar double ψ-loop.

The second part of Table 7, from chymosin (2CMS) onward, details ψ-loops that have not been reported previously. Additional ψ-loops were found in p21 protein catalytic domain (2P21), chymosin, rhizopuspepsin, penicillopepsin, pseudoazurin (2PAZ), and plastocyanin (6PCY), these additional loops increasing the number of occurrences of the motif from 7 to 10 unrelated proteins. The simplest ψ-loop (strand connectivities +2,−1), was found in p21 protein catalytic domain. ψ-loops occurring in pairs were identified in three additional members of the aspartic protease family: chymosin, HIV protease (3HVP) and Rous sarcoma virus protein (2RSP). The pair consists of one type 1 and one type 1X′ ψ-loop and has the overall topology +2,+1x,−2. The remaining two ψ-loops occur in the N-terminal regions of plastocyanin and psueodazurin; these loops are of type 1′, which has not previously been detected in the Protein Data Bank.

Since the work reported in Table 7 was carried out, the search has been repeated on the entire April 1992 release of the Protein Data Bank, with the resulting identification of 16 ψ-loops in 13 additional proteins. These proteins and their ψ-loop topologies are detailed in Table 8. The type 1 loop (+2,−1) was found in eight proteins (flavocytochrome (1FCB), chloramphenicol acetyltransferase (1CLA), cholesterol oxidase (1COX), mannose-binding protein (1MSB), leucine aminopeptidase (1LAP), fructose 1,6-bisphosphatase (1FBP), mesentericopeptidase (1MEE), and scorpion neurotoxin (1NRD)). Type 1X loops were identified in proteins from four different families (isoenzyme 3–3 glutathione-5-transferase (1GST), leucine aminopeptidase, serine carboxypeptidase (2SC2), and acetylcholinesterase (1ACE)), although the striking structural similarity that we have recently identified[13] between carboxypeptidase A and leucine aminopeptidase may be indicative of a very remote, divergent relationship between the families of aminopeptidases and carboxypeptidases. Additionally, paired type 1–type 1X loops (+2,+1x,−2) were identified in two unrelated proteins (leucine aminopeptidase and purine nucleoside phosphorylase (2PNP)).

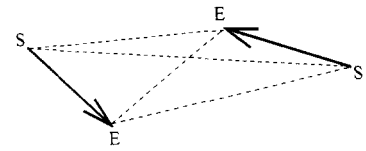


Figure 2. Representation of the SS, SE, ES, and EE distances (dashed lines) between the vectors representing two residues.

Table 9. Representation of the Geometrical Relationships between the Asp-32 and Asp-215 Residues in Endothiapepsin (4APE)[a]

	Asp-32	Asp-215
Asp-32	0 0 0 0	76 60 62 43
Asp-215	76 62 60 43	0 0 0 0

[a] Each entry is a distance (Å) and then multiplied by 10 and rounded to the nearest integer.

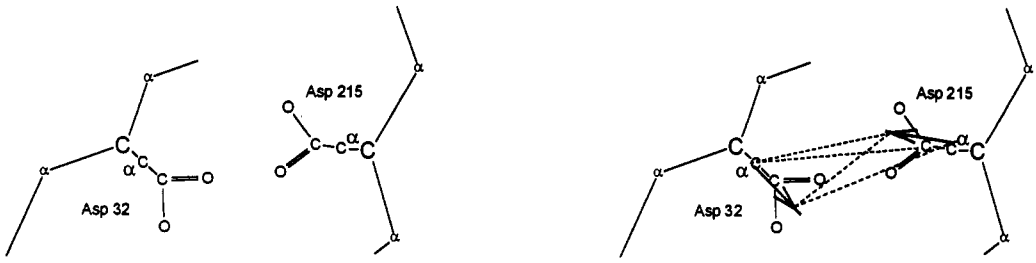
Taking the results in Tables 7 and 8 together, therefore, ψ-loops were identified in 16 unrelated proteins. As the April 1992 release of the Protein Data Bank contains representatives of approximately 120 different protein families,[36] our findings would suggest that the ψ-loop is a comparatively rarely occurring structural feature.

5. REPRESENTATION AND SEARCHING OF RESIDUES

The work that has been discussed thus far involves the use of the Ullmann algorithm with graphs in which the nodes denote the α-helix or β-strand secondary-structure elements. We are currently extending our programs so that subgraph-isomorphism searches can be carried out on graphs in which the nodes denote individual amino acid side chains (with the edges again denoting the internode geometric relationships), thus allowing the retrieval of all of the sets of residues in a 3-D protein that provide a geometric fit with a comparable set of query residues.

A simplified representation of the structure is used (although this may be made more complex later) in which each side chain is represented by two pseudo-atoms, one near the start and the other near the end of each side chain. Similar residues are represented in a similar fashion, so that, for example, glutamate residues and aspartate residues can be compared directly if desired. Thus, in addition to the 20 standard amino acids, generic amino acid types (acidic, amide, basic, aromatic, small and large hydrophobic) can also be defined by the user as appropriate to a particular search. The locations of the two pseudo-atoms are used to generate a vector, and the geometric relationships between pairs of residues are defined in terms of distances between the corresponding vectors. Specifically, if we let S and E denote the start and the end, respectively, of a vector, then the graph edges contain four parts, these being the SS, SE, ES, and EE distances; these distances are illustrated schematically for two vectors in Figure 2. As a specific example, consider the two aspartyl residues (Asp-32 and Asp-215) in endothiapepsin, the structure of which (4APE) was extracted from chestnut blight fungus and resolved at 2.1 Å. The distance matrix for these two residues is illustrated in Table 9, the geometric arrangement of the two residues in Figure 3a and the two residues with the superimposed distances in Figure 3b. It is also possible to include the middle (M) points of the vectors, thus allowing the specification of a further, MM distance; however, we have not used this distance for the searches that are detailed below.

This vectorial representation is clearly an extremely simple description of the relative orientations of the side chains in a



a) b)

Figure 3. Two aspartyl residues in endothiapepsin that are involved in the catalytic action of the enzyme: (a) the residues alone and (b) residues with the vectors and distances superimposed.

3-D protein structure. It does, however, have the advantage that it does not overdefine the orientations of the ends of side chains, as could occur if a more precise representation was to be used that was based directly on the individual atomic coordinates in the Protein Data Bank. This is a useful feature for at least three reasons: in medium-resolution protein-crystallographic studies, it is often difficult to get the final torsion angle value correct and so the fine details of the side chain orientations may be in doubt; the identifications of the individual atoms in a residue can often be ambiguous, e.g., the ND2 and OD1 atoms of an asparagine residue, and side chains can often move or twist, for example on binding substrates.

We shall illustrate the use of the program here by means of a search for a pattern based on the two aspartate groups that are known to be involved in the catalytic mechanism of the aspartic proteinases.[37] The aspartic proteinases are a group of enzymes which catalyze the hydrolysis of proteins; in their mechanism of action, they are specific for peptide bonds located between large hydrophobic residues and they all contain two essential aspartyl groups.[38] The search pattern here consisted of the two aspartyl residues from endothiapepsin that have been discussed above. A series of searches was carried out to investigate the extent to which the user-defined distance tolerances affected the search output. For any given search pattern, an increase in the tolerance will generally lead to the retrieval of a greater number of sites: however, the relevance of the higher tolerance sites will have to be considered more carefully than those retrieved in a low-tolerance, high-precision search. The aspartic proteinase pattern was searched at four different tolerances, ±0.2, ±0.4, ±0.6, and ±0.8 Å.

Each retrieved site was identified as belonging to one of three predefined classes: (1) known aspartic proteinase catalytic sites; (2) known metal binding sites; (3) any hit residues not covered by classes 1 or 2. Thus, for example, a hit on residues Asp-32 and Asp-215 (the active-site residues) in pepsinogen (1PSG) is a class 1 hit; a hit on residues Asp-51 and Asp-53 (the Calcium-ion binding site) in oncomodulin (1OMD) is a class 2 hit; and a hit on residues Asp-11 and Asp-34 (unknown relevance) in ferredoxin (1FXI) is a class 3 hit. A total of 19 different aspartic proteinase compounds were identified in the January 1992 version of the Protein Data Bank that was used in the searches, the codes of these compounds being as follows: 1CMS, 1MVP, 1PSG, 2APR, 2ER0, 2ER6, 2RSP, 3APP, 3PEP, 4APE, 4APR, 4CMS, 4HVP, 4PEP, 5APR, 5CMS, 5HVP, 5PEP, and 6APR. The active site of the aspartic proteinases is highly symmetrical,[37] this implying a maximum possible of 38 class-1 sites in the search file.

The results for the searches at the four different tolerances are shown in Table 10. The figures in the table show the number of sites that were retrieved in each class at each tolerance, with percentages in brackets. The table shows clearly that the number of pairs of residues able to satisfy the query constraints increases as the tolerance increases, from a total of 14 at ±0.2 Å to 135 at ±0.8 Å. In addition, an increase in the tolerance is accompanied by a decrease in the percentage of class-1 hits (and thus by an increase in the relative percentages of the other two classes). For example, at the lowest tolerance, 79% of the hits belong to class 1 (and none to class 3 at all) and the search pattern is clearly highly specific for sites with the same catalytic nature. At the highest tolerance, conversely, only 28% of the hits come from class 1 with 35 and 37% from classes 2 and 3, respectively. The actual number of class-1 hits increases by only 2 (from 36 to 38) in moving from ±0.6 to ±0.8 Å, which indicates that the majority of acid proteinase active sites have been identified at the lower of these two tolerances. In fact, it was established that all of the 19 sites had been identified at the ±0.6 Å tolerance and in both of the two possible orientations (with two exceptions, these being chymosin (1CMS) and pepsin (5PEP)).

Table 10. Numbers (Percentages in Brackets) of Sites Retrieved in Searches for the Aspartic Proteinase Pattern at Different Distance Tolerances

	no. of sites for given tolerance (Å)			
class	0.2	0.4	0.6	0.8
1	11 (79)	28 (80)	36 (53)	38 (28)
2	3 (21)	6 (17)	15 (22)	47 (35)
3	0 (0)	1 (3)	17 (25)	50 (37)
totals	14	35	68	135

The aspartic proteinase results demonstrate clearly the power and flexibility of the technique. In this case, the search pattern consists of just two residues, and yet the program is still able to retrieve all of the known aspartic proteinases in the January 1992 release of the Protein Data Bank. Even with the largest tolerance tested here, ±0.8 Å, almost two-thirds of the retrieved pairs of residues are either known aspartic proteinase active sites or known metal-binding sites of one sort or another. There are sufficiently few of the remaining "sites of unknown character" to permit a detailed analysis of them; in this particular case, no particularly interesting structural features were identified (but searches with a range of other patterns have revealed large numbers of unexpected, and potentially interesting, side chain patterns that are now under intensive study in our laboratories).

6. CONCLUSIONS

In this paper, we have discussed the use of a subgraph isomorphism algorithm for the identification of patterns of

secondary-structure elements and patterns of amino acid residues in the 3-D structures in the Protein Data Bank.

The use of a graph-matching approach has allowed us to carry out an exhaustive survey of all of the β-motifs that occur in a 114-protein subset of the Protein Data Bank. This survey has revealed that only a very small fraction of the possible β-motifs occur in practice, and the observed distribution of the motifs' frequencies of occurrence suggest that this will continue to be the case even when very large numbers of 3-D protein structures become available. We have also been able to identify 20 proteins that had not previously been known to contain the ψ-loop motif, which consists of three β-strands. When taken with the many previous studies we have carried out (see, e.g., refs 11, 13, and 16), we believe that these results provide a firm basis for the application of graph-theoretical approaches to the analysis of the secondary and tertiary structures of proteins.

We are currently extending our programs to encompass the representation and searching of amino acid side chain patterns. The aspartic proteinase results demonstrate clearly that it is possible to carry out high-precision searches, despite the great simplicity of the representation that we have used. Since these results were obtained, we have carried out searches for several other side chain patterns: this work has resulted in the identification of many previously unrecognized structural resemblances and will be reported shortly.

ACKNOWLEDGMENT

We thank Pfizer Central Research, the Science and Engineering Research Council and Tripos Associates for funding. P.J.A. is a Royal Society University Research Fellow; D.W.R. is a Lister Institute Research Fellow. This paper is a contribution from the Krebs Institute for Biomolecular Research, which is a designated Centre for Molecular Recognition Studies under the Molecular Recognition Initiative of the Science and Engineering Research Council.

REFERENCES AND NOTES

(1) Artymiuk, P. J.; Rice, D. W. Database Systems in Molecular Biology. In *Chemical Structure Systems*; Ash, J. E., Warr, W. A., Willett, P., Eds.; Ellis Horwood: Chichester, U.K., 1991; pp 299–328.
(2) Thornton, J. M.; Gardner, S. P. Protein Motifs and Database Searching. *Trends Biochem. Sci.* **1989**, *14*, 300–304.
(3) Islam, S. A.; Sternberg, M. J. E. A Relational Database of Protein Structures Designed for Flexible Enquiries about Conformation. *Protein Eng.* **1989**, *2*, 431–442.
(4) Gray, P. M. D.; Paton, N. W.; Kemp, G. J. L.; Fothergill, J. E. An Object-Oriented Database for Protein Structure Analysis. *Protein Eng.* **1990**, *3*, 235–243.
(5) Rawlings, C. J.; Taylor, W. R.; Nyakairu, J.; Fox, J.; Sternberg, M. J. E. Reasoning about Protein Topology Using the Logic Programming Language PROLOG. *J. Mol. Graphics* **1985**, *3*, 151–157.
(6) Fischer, D.; Bachar, O.; Nussinov, R.; Wolfson, H. An Efficient Automated Computer Vision Based Technique for Detection of Three-Dimensional Structural Motifs in Proteins. *J. Biomol. Struct. Dyn.* **1992**, *9*, 769–789.
(7) Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; Vriend, G. A Database of Protein Structure Families with Common Folding Motifs. *Protein Sci.* **1992**, *1*, 1691–1698.
(8) Brint, A. T.; Davies, H. M.; Mitchell, E. M.; Willett, P. Rapid Geometric Searching in Protein Structures. *J. Mol. Graphics* **1987**, *7*, 48–53.
(9) Artymiuk, P. J.; Mitchell, E. M.; Rice, D. W.; Willett, P. Searching Techniques for Databases of Protein Secondary Structures. *J. Inf. Sci.* **1989**, *15*, 287–298.
(10) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins. *J. Mol. Biol.* **1990**, *212*, 151–166.
(11) Artymiuk, P. J.; Rice, D. W.; Mitchell, E. M.; Willett, P. Structural Resemblance between the Families of Bacterial Signal Transduction Proteins and of G Proteins Revealed by Graph Theoretical Techniques. *Protein Eng.* **1990**, *4*, 39–43.
(12) Artymiuk, P. J.; Grindley, H. M.; Rice, D. W.; Ujah, E. C.; Willett, P. Searching Techniques for the Tertiary Structure of Proteins in the Protein Data Bank. In *Recent Advances in Chemical Information*; Collier, H., Ed.; Royal Society of Chemistry: Cambridge, U.K., 1991; pp 91–106.
(13) Artymiuk, P. J.; Grindley, H. M.; Park, E. J.; Rice, D. W.; Willett, P. Three-Dimensional Structural Resemblance between Leucine Aminopeptidase and Carboxypeptidase A Revealed by Graph-Theoretical Techniques. *FEBS Lett.* **1992**, *303*, 48–52.
(14) Artymiuk, P. J.; Bath, P. A.; Grindley, H. M.; Pepperrell, C. A.; Poirrette, A. R.; Rice, D. W.; Thorner, D. A.; Wild, D. J.; Willett, P.; Allen, F. H.; Taylor, R. Similarity Searching in Databases of Three-Dimensional Molecules and Macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 617–630.
(15) Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm. *J. Mol. Biol.* **1993**, *33*, 707–721.
(16) Artymiuk, P. J.; Grindley, H. M.; Kumar, K.; Rice, D. W.; Willett, P. Three-Dimensional Structural Resemblance Between the Ribonuclease H and Connection Domains of HIV Reverse Transcriptase Revealed Using Graph-Theoretic Techniques. *FEBS Lett.* **1993**, *324*, 15–21.
(17) Ash, J. E., Warr, W. A., Willett, P., Eds. *Chemical Structure Systems*; Ellis Horwood: Chichester, U.K., 1991.
(18) Martin, Y. C. 3D Database Searching in Drug Discovery. *J. Med. Chem.* **1992**, *35*, 2145–2154.
(19) Willett, P. *Three-Dimensional Chemical Structure Handling*; Research Studies Press: Taunton, U.K., 1991.
(20) Kaden, F.; Koch, I.; Selbig, J. Knowledge-Based Prediction of Protein Structures. *J. Theor. Biol.* **1990**, *147*, 85–100.
(21) Koch, I.; Kaden, F.; Selbig, J. Analysis of Protein Sheet Topologies by Graph Theoretical Methods. *Proteins: Struct., Funct., Genet.* **1992**, *12*, 314–323.
(22) Subbarao, N.; Haneef, I. Defining Topological Equivalences in Macromolecules. *Protein Eng.* **1991**, *4*, 877–884.
(23) Kasinos, N.; Lilley, G. A.; Subbarao, N.; Haneef, I. A Robust and Efficient Automated Docking Algorithm for Molecular Recognition. *Protein Eng.* **1992**, *5*, 69–75.
(24) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, M.; Tasumi, M. The Protein Data Bank: a Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535–542.
(25) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. Protein Data Bank. In *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*; Allen, F. H., Bergeroff, G., Sievers, R., Eds.; Data Commission of the International Union of Crystallography: Cambridge, U.K., 1987; pp 107–132.
(26) Tang, J.; James, M. N. G.; Hsu, I. N.; Jenkins, J. A.; Blundell, T. L. Structural Evidence for Gene Duplication in the Evolution of the Acid Proteases. *Nature* **1978**, *271*, 618–621.
(27) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach.* **1976**, *16*, 31–42.
(28) PROTEP is distributed by Tripos Associates Inc., St. Louis, MO.
(29) Richardson, J. S. β-Sheet Topology and the Relatedness of Proteins. *Nature* **1977**, *268*, 495–500.
(30) Levitt, M.; Chothia, C. Structural Patterns in Globular Proteins. *Nature* **1976**, *261*, 552–557.
(31) Chothia, C. Principles that Determine the Structure of Proteins. *Annu. Rev. Biochem.* **1984**, *53*, 537–572.
(32) Richardson, J. S. The Anatomy and Taxonomy of Protein Structure. *Adv. Protein Chem.* **1981**, *34*, 167–339.
(33) Richardson, J. S. Handedness of Crossover Connections in β-Sheets. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73*, 2608–2623.
(34) Schulz, G. E.; Schirmer, R. H. *Principles of Protein Structure*; Springer: New York, 1979.
(35) Hutchinson, E. G.; Thornton, J. M. HERA-a Program to Draw Schematic Diagrams of Protein Secondary Structures. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 203–212.
(36) Chothia, C. One Thousand Families for the Molecular Biologist. *Nature* **1992**, *357*, 543–544.
(37) Pearl, L.; Blundell, T. The Active Site of Aspartic Proteinases. *FEBS Lett.* **1984**, *174*, 96–101.
(38) Fersht, A. *Enzyme Structure and Mechanism*; 2nd ed.; W. H. Freeman: San Francisco, 1985.