Article No. jmbi.1999.2581 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 286, 1673–1691

JMB

Three-dimensional Structure Analysis of PROSITE Patterns

Atsushi Kasuya[1] and Janet M. Thornton[1,2]*

[1]*Biomolecular Structure and Modelling Unit, Department of Biochemistry & Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK*

[2]*Crystallography Department Birkbeck College, Malet Street London, WC1E 7HX, UK*

**Corresponding author*

Pattern matches for each of the sequence patterns in PROSITE, a database of sequence patterns, were searched in all protein sequences in the Brookhaven Protein Data Bank (PDB). The three-dimensional structures of the pattern matches for the 20 patterns with the largest numbers of hits were analysed. We found that the true positives have a common three-dimensional structure for each pattern; the structures of false positives, found for six of the 20 patterns, were clearly different from those of the true positives. The results suggest that the true pattern matches each have a characteristic common three-dimensional structure, which could be used to create a template to define a three-dimensional functional pattern.

Keywords: protein sequence motifs; protein family; protein function; structure; conformation

Introduction

In amino acid sequences of proteins with a common function, the occurrence of particular clusters of residue types is often observed. Such recurring clusters, called patterns, motifs or fingerprints, are not only common to homologous proteins, but also in some cases, common to distantly related proteins in which sequence similarities are hardly identified (Bork & Koonin, 1996). Therefore, these characteristic features of protein sequences are thought to be important especially when predicting the function of uncharacterised proteins from their primary structure (Rastan & Beeley, 1997; Andrade & Sander, 1997).

For various protein families, such sequence patterns have been identified and accumulated in databases. The patterns are often represented either as regular expressions (''patterns'') of possible combinations of amino acids, or as tables of probabilities of each amino acids for each position (''profiles'').

PROSITE (Bairoch, 1993; Bairoch *et al*., 1997), which links with the SWISS-PROT protein sequence database (Bairoch & Apweiler, 1997), is one of the most widely used and comprehensive pattern databases. Although some entries are defined as profiles, in the form of weight matrices, the majority are described as patterns. The patterns in PROSITE were derived from analyses of the sequences, based on the literature for proteins of certain functions.

While the PROSITE patterns are defined in principle as the shortest discriminating sequences, in order to improve the recognition power for distant relatives, multiple sequence motifs for a family have been introduced. Attwood *et al*. (1994, 1998) have developed a database PRINTS, which contains multiple sequence motifs (''fingerprints'') derived from the results of multiple sequence alignments. A similar approach has been made by Henikoff & Henikoff (1994) in their database, Blocks. The Blocks database contains multiple alignments of the conserved regions (''blocks'') in protein families. Protein or nucleotide sequences can be searched against blocks. A profile is computed from each block, and used to score a query sequence (Henikoff & Henikoff, 1994; Henikoff *et al*., 1998).

Generally, these sequence patterns are considered to include those residues which play important roles in the function of the protein or in the formation of the core structure of the protein. Therefore, it is expected that for each conserved sequence pattern there is a common three-dimensional structure, which can be considered characteristic of the function. Indeed, some of the sequence patterns have been identified from analysis of three-dimensional structure as well as from sequence analysis. For example, the sequence of the EF-hand, which was one of the first recognised

Abbreviations used: PDB, Protein Data Bank; ; rmsd, root-mean-square deviation; 3D, three-dimensional.

E-mail address of the corresponding author: thornton@biochem.ucl.ac.uk

0022-2836/99/101673–19 $30.00/0

functional motifs in proteins, is constrained to form the helix-turn-helix structure, deduced from the analysis of the three-dimensional structures (Strynadka & James, 1989; Marsden *et al*., 1990). The structural requirement for the function puts constraints on the sequence of the protein.

Although three-dimensional structures have sometimes been used to define new patterns, pattern matches in the sequences of unknown three-dimensional structures have not been generally used in modelling their structures. No general rules have so far been obtained concerning how the structures of the matches for each pattern are conserved, and to what extent one can rely upon them when modelling the structures. Recently, because of the accelerated rate of sequence determination, information on sequence patterns, including many new patterns, have been deduced from the sequence analyses (Bork *et al*., 1995) and used to predict the function of uncharacterised proteins. The three-dimensional structures of such new patterns have not always been examined. If a sequence pattern match is found in a protein sequence of unknown structure, will the three-dimensional structure of the matched fragments be similar to those of other matched sequences?

Here we report the results of the analysis of the three-dimensional structures of the sequence patterns. Sequence matches for each pattern in PROSITE were searched in a sequence library of proteins of known three-dimensional structures. The structures of sequence matches were analysed for structure distinctions between true and false positives, and for common structural features among the true positives.

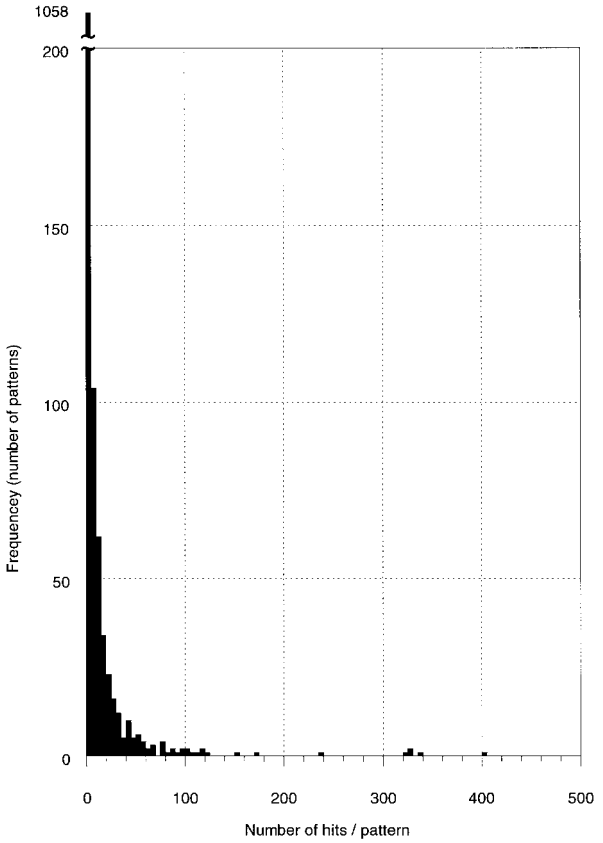

Figure 1. Distribution of number of hits per pattern. Each column represents frequency of patterns having sequence matches in the 3D-sequence library. The bin width is 10. There were 1058 patterns having numbers of hits smaller than 10, including 712 patterns with no hits.

Results

Pattern searches

The initial search through the PDB sequences (Bernstein *et al*., 1977) for matches against 1265 PROSITE patterns (Bairoch *et al*., 1997) gave 9493 matches in all, with 553 of the patterns having one or more hits and the remaining 712 patterns having none. Figure 1 shows the distribution of number of hits per pattern. There were 14 patterns that gave more than 100 matches and 38 patterns that gave more than 50. Table 1 shows the 20 patterns with largest numbers of matches.

The numbers of hits include false positives as well as true positives, as described in the following section. Excluding false and unidentified hits (see below), the number of true positive hits was 8788. For 16 patterns, only false or unidentified hits were found: one or more true positives were found for 537 patterns. This indicates that, three-dimensional structures of 42% of 1265 patterns in the PROSITE database are known with an average number 16 structures per pattern, although many of these are just the structures of the same protein solved in different states or under different conditions.

Most of the frequently occurring patterns were those for active/binding sites. There were also patterns defining disulphide bonds and domain structures. For some protein functions, two or more patterns are defined: in Table 1, both TRYPSIN_HIS and TRYPSIN_SER are for serine proteases, XYLOSE_ISOMERASE_1 and XYLOSE_ISOMERASE_2 are for xylose isomerases, and LECTIN_LEGUME_ALPHA and LECTIN_LEGUME_BETA are for legume lectins. The matches for these patterns tend to overlap one another; sequences that were hit by one pattern were usually also hit by the other pattern, except in some cases of false negatives and mutated sequences. Thus, of the 325 TRYPSIN_HIS matches 321 also matched TRYPSIN_SER, of the 112 LECTIN_LEGUME_ALPHA and 118 LECTIN_LEGUME_BETA matches 110 were common to both, and of the 116 XYLOSE_ISOMERASE_2 matches 108 were also matched XYLOSE_ISOMERASE_1.

Although 120 matches were found in the search for the INTRADIOL_DIOXYGENAS pattern, all of them were from a single protein, protocatechuate 3,4-dioxygenase, present in ten individual structure

Table 1. PROSITE patterns with largest numbers of matches in the 3D-sequence library

PROSITE ID[a]	Accession number[a]	Number of chain hits (PDB entries)	Pattern in regular expression[b]	Length (res.)	Description[a]
ATP_GTP_A	PS00017	402(256)	[AG].{4}GK[ST]	8	ATP/GTP-binding site motif A (P-loop)
IG_MHC	PS00290	336(185)	[FY].C.[VA].H	7	Immunoglobulins and major histocompatibility complex proteins signature.
TRYPSIN_HIS	PS00134	325(285)	[LIVM][ST]A[STAG]HC	6	Serine proteases, trypsin family, histidine active site
ASP_PROTEASE	PS00141	325(159)	[LIVMFGAC][LIVMTADN][LIVFSA]D[ST]G[STAV][STAPDENQ].[LIVMFSTNC].[LIVMFGTA]	12	Eukaryotic and viral aspartyl proteases active site
TRYPSIN_SER	PS00135	321(282)	[DNSTAGC][GSTAPIMVQH].{2}G[DE]SG[GS][SAPHV][LIVMFYWH][LIVMFYSTANQH]	12	Serine proteases, trypsin family, serine active site
EF_HAND[c]	PS00018	237(93)	D.[DNS][^ILVFYW][DENSTG][DNQGHRK][^GP][LIVMC][DENQSTAGC].{2}[DE][LIVMFYW]	13	EF-hand calcium-binding domain
LACTALBUMIN_LYSOZYME	PS00128	170(154)	C.{3}C.{2}[LF].{3}[DEN][LI].{5}C	19	α-lactalbumin/lysozyme C signature
CYTOCHROME_C	PS00190	152(97)	C[^CPWHF][^CPWR]CH[^CFYW]	6	Cytochrome *c* family heme-binding site signature
INTRADIOL_DIOXYGENAS	PS00083	120(10)	[LIVM].G.[LIVM].{4}[GS].{2}[LIVM].{4}[LIVM][DE][LIVMFY].{6}G.[FY]	29	Intradiol ring-cleavage dioxygenases signature
LECTIN_LEGUME_BETA	PS00307	118(47)	[LIV][STAG]V[EQV][FLI]D[ST]	7	Legume lectins beta-chain signature
XYLOSE_ISOMERASE_2	PS00173	116(53)	[FL]HD.D[LIV].[PD].[GDE]	10	Xylose isomerase signature 2
LECTIN_LEGUME_ALPHA	PS00308	112(46)	[LIV].[EDQ][FYWKR]V.[LIV]G[LF][ST]	10	Legume lectins alpha-chain signature
XYLOSE_ISOMERASE_1	PS00172	108(50)	[LI]EPKP.{2}P	8	Xylose isomerase signature 1
ANNEXIN	PS00223	101(17)	[TG][STV].{8}[LIVMF].{2}R.{3}[DEQNH].{7}[IFY].{7}[LIVMF].{3}[LIVMF].{11}[LIVMFA].{2}[LIVMF]	53	Annexins repeated domain signature
AA_TRNA_LIGASE_II_2	PS00339	100(59)	[GSTALVF][^DENQHRKP][GSTA][LIVMF][DE]R[LIVMF].[LIVMSTAG][LIVMFY]	10	Aminoacyl-transfer RNA synthetases class-II signature 2
DNA_POLYMERASE_X	PS00522	99(99)	G[SG][LFY].R[GE].{3}[SGCL].D[LIVM]D[LIVMFY]{3}.{2}[SAP]	20	DNA polymerase family X signature
EUK_CO2_ANHYDRASE	PS00162	95(92)	SEH.[LIVM].{4}[FYH].{2}E[LIVM]H[LIVMFA]{2}	17	Eukaryotic-type carbonic anhydrases signature
PEROXIDASE_1	PS00435	93(73)	[DET][LIVMTA].{2}[LIVM][LIVMSTAG][SAG][LIVMSTAG]H[STA][LIVMFY]	11	Peroxidases proximal heme-ligand signature
COPPER_BLUE	PS00196	89(53)	[GA].{0,2}[YSA].{0,1}[VFY].C.{1,2}[PG].{0,1}H.{2,4}[MQ]	12-17[c]	Type-1 copper (blue) proteins signature
INSULIN	PS00262	87(44)	CC[P].{2}C[STDNEKPI].{3}[LIVMFS].{3}C	15	Insulin family signature

[a] Pattern ID, accession number and description in the PROSITE database (Bairoch *et al.*, 1997).

[b] PROSITE patterns, described using the following shorthand: amino acids are indicated by the standard one-letter code; letters within square brackets [. . .] indicate the acceptable amino acids for the position, or the unacceptable amino acids if the first letter is ^ (i.e. [^ . . .]); the symbol . is used to indicate that any amino acid is acceptable for the position; repetition of an element is indicated by a following numerical value (number of repetition) or a numerical range (minimum and maximum acceptable number) within curly brackets { . . . }.

[c] Range of lengths of the actually matched sequences.

entries in the PDB. The structure of the enzyme has six protomers in the asymmetric unit; each protomer consists of alpha and beta chains (Ohlendorf *et al.*, 1994; Orville *et al.*, 1997a,b). The pattern matched both the alpha and beta chains in each protomer, giving 120 matches in all from the ten PDB entries.

True and false identifications

Within the total of 9493 sequence matches, 8788 (93 %) were identified as true positives and 359 (4 %) as false positives. Two matches were indicated as ''unknown'' in the databases. There were 653 matches which could not be identified by solely following links in the PDB, SWISS-PROT and PROSITE databases. In those, 304 true and five false matches were identified by manual inspection and included in the above numbers. For all patterns in Table 1, all of the matches were identified as true or false except for the ATP_GTP_A pattern.

Table 2 shows the results of the true and false identifications. About 70 % of the matches for the ATP_GTP_A pattern could not be clearly identified. For all patterns except AA_TRNA_LIGASE_II_2, either all matches or the majority of matches were true positives, confirming the recognition power of the PROSITE patterns.

Structure comparisons of pattern matches

The group rmsd values of the true hits were calculated for all 466 patterns having more than one true match. The results are shown in Figure 2. Although there were several patterns having exceptionally high rmsd values, the distribution of the rmsd values had clear tendency to concentrate below 0.5 Å. There were 276 patterns (59 %) having rmsd values smaller than 0.5 Å, and 376 patterns (81 %) smaller than 1 Å.

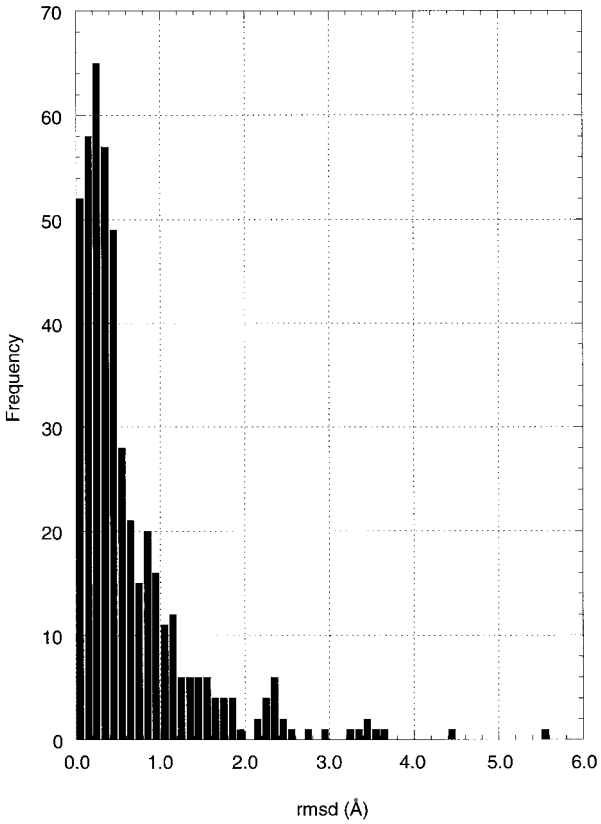

Figure 2. Distribution of rmsd values for the true hits. The rmsd was calculated from all true hits eliminating false and unidentified hits for each of the 466 patterns having more than one true hit.

Table 2 shows the rmsd values for the structure fragments corresponding to the pattern matches.

Table 2. True/false identification of the pattern matches

PROSITE ID	Number of C^{α}	Number of hits (rmsd(Å)) Total	True	False
ATP_GTP_A[a]	4	401(2.16)	113(0.44)	4(2.53)
IG_MHC	7	336(0.41)	336	0
TRYPSIN_HIS	6	325(0.22)	325	0
ASP_PROTEASE	12	325(0.78)	319(0.63)	6(0.30)
TRYPSIN_SER	12	321(0.79)	321	0
EF_HAND	13	237(2.32)	220(1.49)	17(3.72)
LACTALBUMIN_LYSOZYME	14	170(0.43)	170	0
CYTOCHROME_C	6	152(1.36)	145(0.93)	7(1.81)
INTRADIOL_DIOXYGENAS	15	120(0.40)	120	0
LECTIN_LEGUME_BETA	7	118(0.29)	118	0
XYLOSE_ISOMERASE_2	10	116(0.37)	116	0
LECTIN_LEGUME_ALPHA	10	112(0.35)	112	0
XYLOSE_ISOMERASE_1	8	108(0.14)	108	0
ANNEXIN	20	101(1.35)	101	0
AA_TRNA_LIGASE_II_2	10	100(2.93)	20(0.37)	80(2.52)
DNA_POLYMERASE_X	20	99(0.32)	99	0
EUK_CO2_ANHYDRASE	13	95(0.17)	95	0
PEROXIDASE_1	11	93(2.03)	85(0.75)	8(0.41)
COPPER_BLUE	8	89(1.30)	89	0
INSULIN	15	87(0.97)	87	0

[a] 284 pattern matches were not identified for the ATP_GTP_A pattern (see the text).

For the patterns with false hits, the rmsds of true and false hits are shown separately. As one might expect, the rmsd for the true hits was small for all patterns. The largest true-hit rmsd in Table 2 is 1.49 Å for EF_HAND. Other patterns with large rmsd values are ANNEXIN and COPPER_BLUE with 1.35 and 1.30 Å rmsd, respectively. As described in the following sections, the structures of these three patterns were found to cluster into two or more groups, each of them having a smaller rmsd value. Interestingly, the number of C^{α} atoms used in the structure comparison for each pattern (shown in Table 2) appeared to have little influence on the rmsd value.

Six of the patterns in Table 2 included false hits. The multidimensional scaling plots of the rmsd values of these patterns are shown in Figure 3. Here, the true hits are shown as filled circles and false hits as open rectangles. The unidentified structures in ATP_GTP_A are also included in Figure 3(a), indicated by + symbols. The clusters of true and false hits are labelled as T (or T1 and T2) and F (or F1, F2 and F3), respectively.

Figure 4 shows the results of the multidimensional scaling for the true hits of the 20 patterns, excluding any false hits. For some patterns, there was more than one cluster in the plot. The clusters which differ significantly from each other are

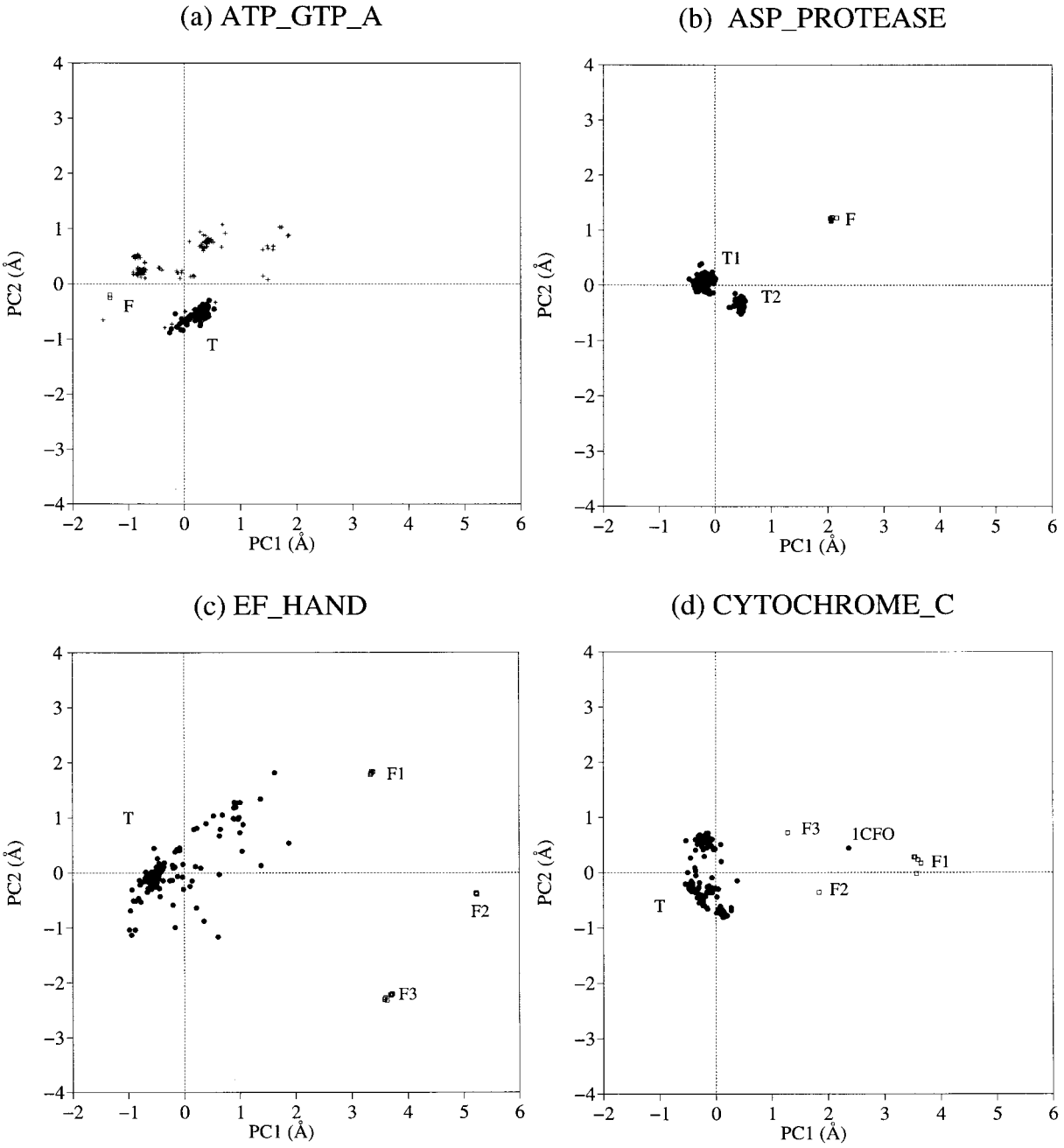

Figure 3(a-d) (*legend overleaf*)

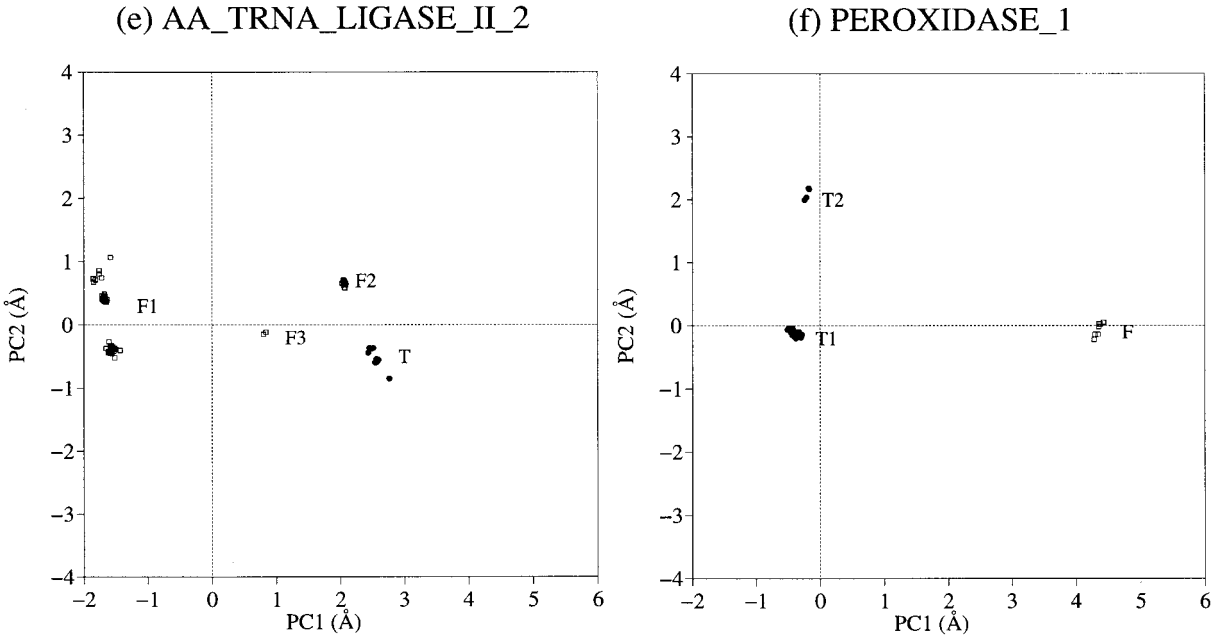

Figure 3. The multidimensional scaling plots for six patterns with false hits. The true and false hits are indicated with filled circles and open squares, respectively. The unidentified structures in ATP_GTP_A are also included in (a), indicated by + symbols. For ATP_GTP_A, only the 290 structures within 1.5 Å rmsd of the true hits were included (i.e. 111 were excluded) for clarity; all pattern matches were included for the other patterns. The clusters of true and false hits are labelled as T (or T1 and T2) and F (or F1, F2 and F3), respectively.

labelled as T1, T2, etc. The outlying structures that differ largely (more than about 1 Å) from the others are labelled by their PDB entry codes in the plots.

In the following sections, the results for some of the 20 patterns are discussed in some detail. The patterns not covered below all had only true hits, with no false positives, and all these hits clustered into a single group (as shown in Figure 3) with an overall rmsd as given in Table 2.

ATP_GTP_A

The pattern ATP_GTP_A is defined for the ATP/GTP binding loop motif A, or ''P-loop'' (Walker *et al.*, 1982; Saraste *et al.*, 1990). The pattern is known to match many, but not all, of the ATP/GTP binding proteins. Some of the ATP/GTP binding proteins do not have the loop structure. The pattern also matches a number of proteins which are not thought to bind a nucleotide, although the experimental confirmation has yet to be obtained.

Of the 402 matches from 256 PDB entries, 134 were identified as true hits but about 70% of the matches could not be definitely identified as true or false. We did not identify any hits as false except four matches from mutants of trypsin (1AMH; Perona & Craik, 1995), chemotaxis protein CheY (1HEY; Cronet *et al.*, 1995) and lysozyme (1TAY; Muraki *et al.*, 1992). The others were left unidentified, though nucleotide-binding function has not been reported for most of them. Further inspection was done after the cluster analysis, only for unidentified hits near the cluster of the true hits.

Excluding one of the matches for which there were no coordinates for one of its C^{α} atoms in the PDB, 401 structures were compared in an all against all comparison. The overall rmsd was 2.16 Å, however, the true hits had very similar structures (rmsd 0.44 Å) and all the true hits were in a discrete ''true'' cluster (labelled T in Figure 3(a)). The cluster included some of the structures not identified as true or false (shown by + symbols). Among them were five structures from human and pig lipase (G311-T318 of 1LPA (van Tilbeurgh *et al.*, 1993), 1LPB (Egloff *et al.*, 1995) and 1GPL (Withers-Martinez *et al.*, 1996), A312-T319 of 1ETH (Hermoso *et al.*, 1996)). Although they have similar loop structures to the P-loop, no GTP-binding function was reported for the lipase. The other seven unidentified structures were known to bind ATP or GTP, although they were not mentioned explicitly having the P-loop structure in the PROSITE and SWISS-PROT database. From this three-dimensional structure analysis, it was clear that those seven structures have the loop structure associated with nucleotide binding.

ASP_PROTEASE

The ASP_PROTEASE is a pattern for 12 residues in the active site of eukaryotic and viral aspartyl proteases (Rawlings & Barrett, 1995). The eukaryotic aspartyl proteases are monomeric enzymes consisting of two domains, which are similar and have probably arisen by gene duplication, while most of the viral aspartyl proteases are homodimers (Rao *et al.*, 1991). The aspartyl proteases have

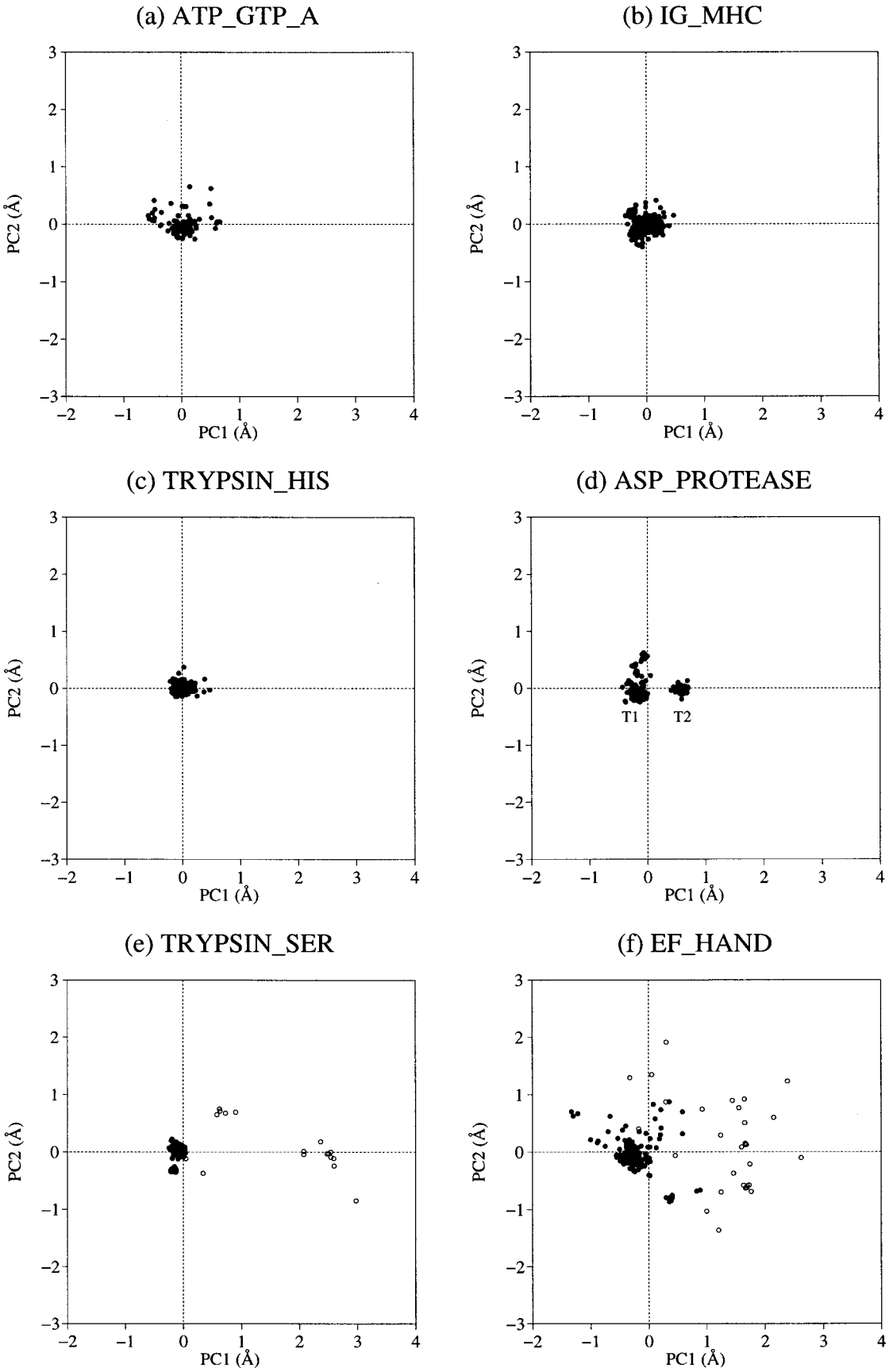

Figure 4 (*legend on page* 1682)

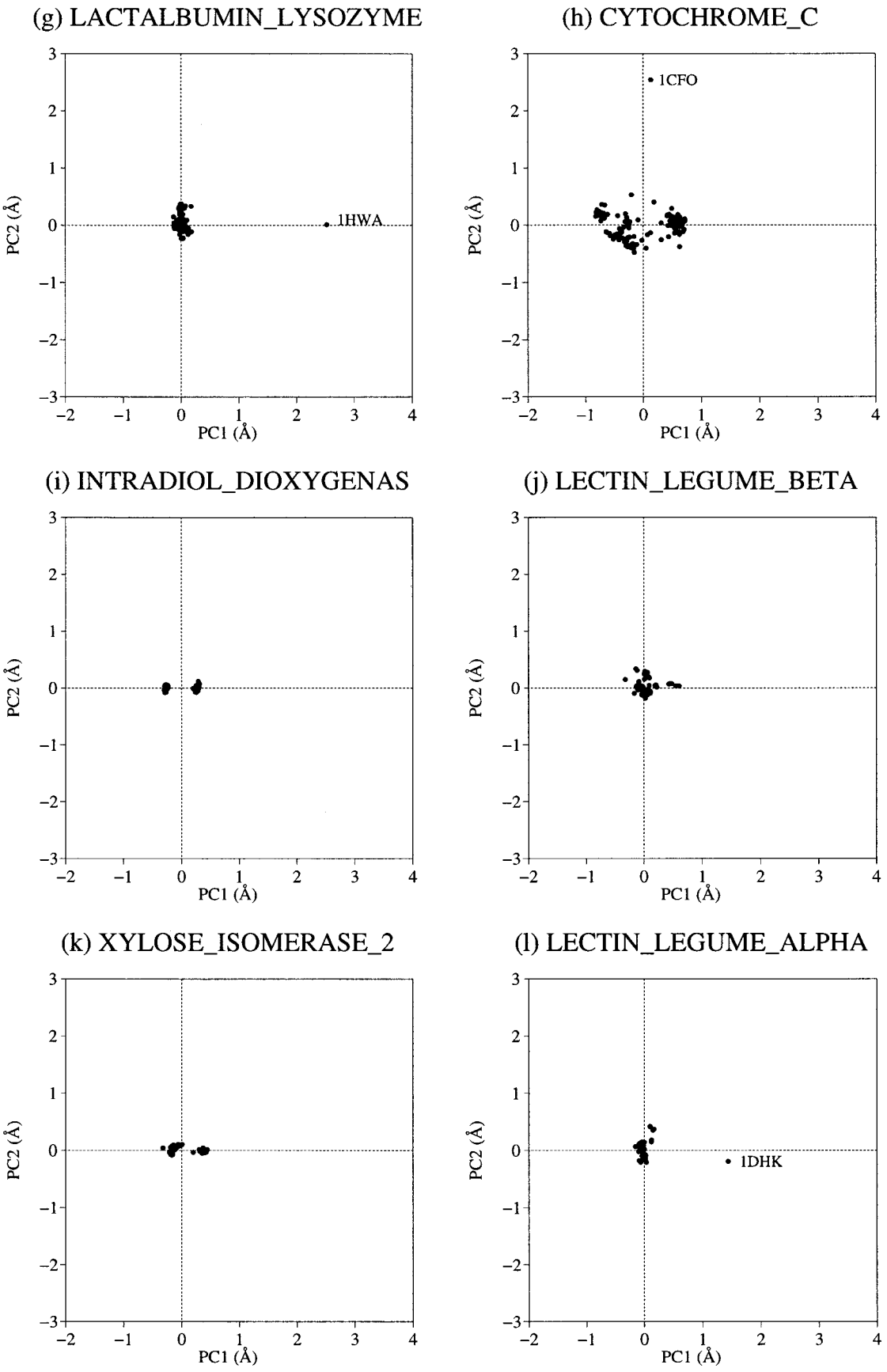

Figure 4 (*legend on page* 1682)

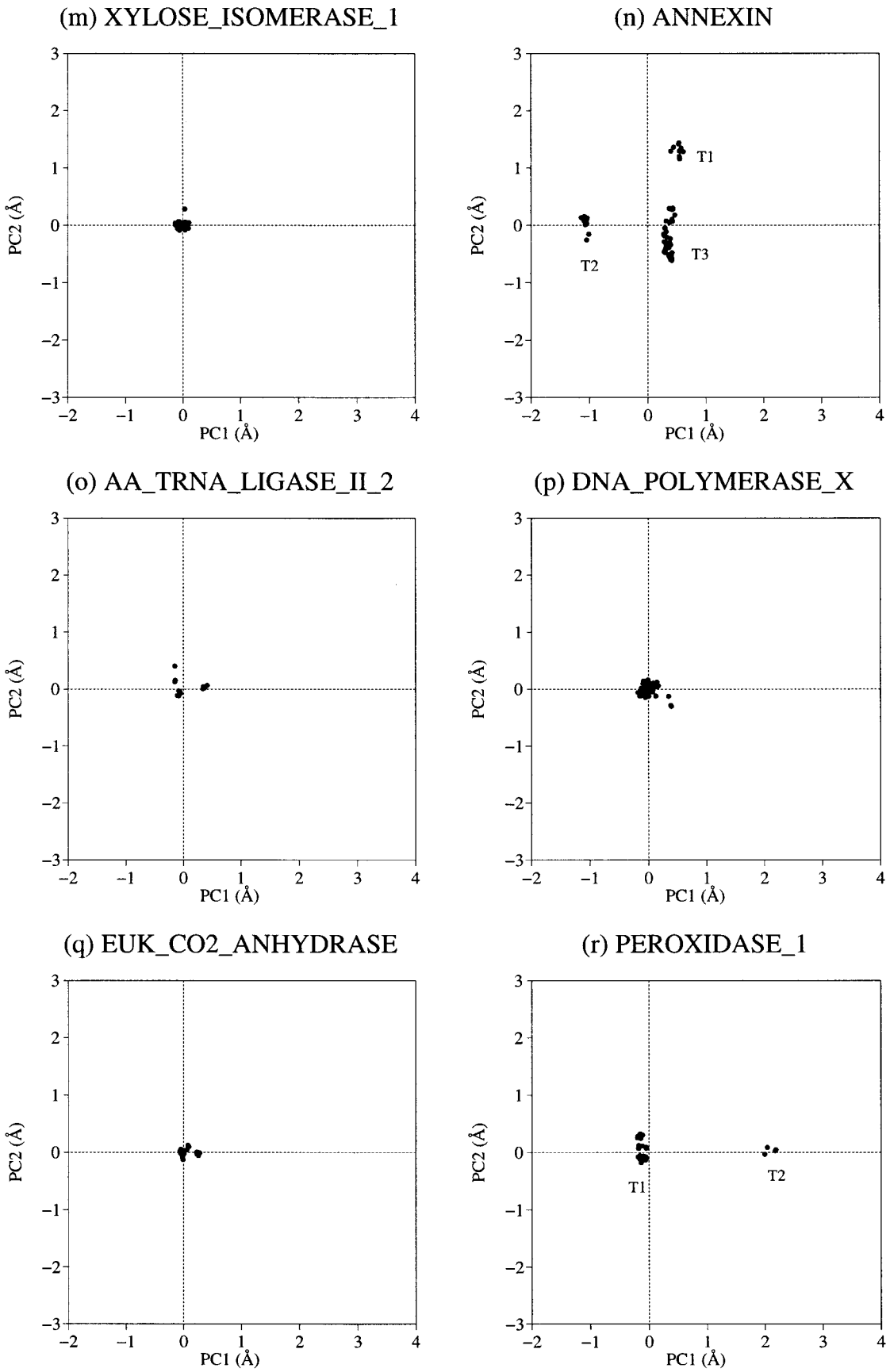

Figure 4 (*legend overleaf*)

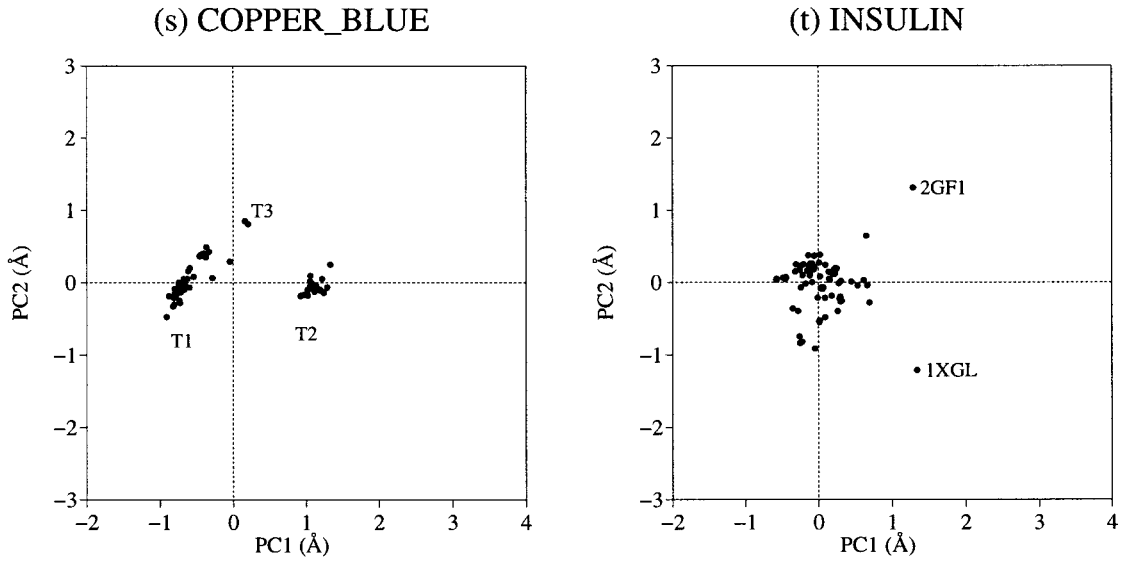

Figure 4. The multidimensional scaling plots for 20 patterns with the largest number of hits. Only the true hits are included for the six patterns with false hits.

two catalytically essential aspartate residues. In the eukaryotic proteases, each domain supplies one of the two aspartate residues, while in the viral proteases, the aspartate residues come from each of the two monomers. The pattern defines the conserved residues around these catalytic aspartate residues. It occurs in each domain in the eukaryotic enzymes and in each monomer in the viral dimer.

Matches for the pattern were found in 325 protein sequences. There were six false hits, all of which were found in the sequence of β-lactamase (the residue range 47-59). The others were true hits that included sequences of HIV-1 protease, HIV-2 protease, endothiapepsin, renin and pepsin. The rmsd value calculated over all hits was 0.78 Å, but this reduced to 0.63 Å for the true hits. The false hits were distinct from the true hits (see Figure 3(b)) and all clustered together since they came from one protein, giving an internal rmsd of only 0.30 Å. The distance between the true and false clusters (D_{TF}) was 2.12 Å.

There were two clusters of the true hits. The larger cluster (T1) consists of 240 structures, 162 from dimeric viral proteases and 78 from C-domains of eukaryotic aspartyl proteases. The smaller (T2) cluster consists of 78 structures from the N-domains of eukaryotic aspartyl proteases (Figure 4(d)). One outlier in the true hits was the structure from 2HVP (Navia *et al.*, 1989), which differs by more than 0.9 Å from both T1 and T2. In Figure 4(d), the point for the 2HVP structure is in the position overlapping the cluster T1.

EF_HAND

The pattern for the EF-hand motif (EF_HAND) in PROSITE defines the consensus sequence of the whole loop region and the first residue which follows the loop (Strynadka & James, 1989; Nakayama & Kretsinger, 1994). There were 237 matches containing this pattern in our 3D-sequence library. Of these, 220 were true hits and 17 were false hits. Figure 3(c) shows the results of the multidimensional scaling analysis of the structures. The true hits show a much broader spread than previous examples, with an rmsd of 1.49 Å. However, the false hits were still notable outliers. The false hits clustered into three distinct groups: the residue range 403-415 of glucoamylase (F1, six hits), 75-87 of galactose oxidase (F2, three hits) and 20-32 of ribulose-1,5-bisphosphate carboxylase/oxigenase (F3, eight hits). The distance between the true and false clusters (D_{TF}) was 2.87 Å.

Figure 4(f) shows the plot of only true hits of the pattern. Note that with the false hits excluded the multidimensional scaling gives an altered spread of data points when projected onto the two principal axes. Among the true hits, 186 structures had bound metal ions (filled circles) and the remaining 34 did not (open circles). The unbound structures include those of low affinity binding sites (such as site I and site II of the N-terminal regulatory domain of troponin C) as well as those of high affinity sites in calcium-free conditions. There is a marked tendency for the unbound structures to be spread more widely than the bound structures, suggesting that the EF-hand structures are more flexible without bound metal ions. The rmsd of the 186 metal bound structures was 0.90 Å.

CYTOCHROME_C

The pattern CYTOCHROME_C is defined for the heme-binding site of the cytochrome *c* family (Mathews, 1985; Ambler, 1991). The six-residue pattern includes two cysteine residues that are covalently bonded to the heme group, and the histidine residue that is one of the ligands of the heme iron.

As shown in Table 2 and Figure 3(d), 152 hits were found with an overall rmsd of 1.36 Å. Seven false hits were found, from the residue range 30-35 of tumour necrosis factor receptor (F1, 5 hits), 45-50 of putidaredoxin (F2), and 74F-79 of a lysozyme mutant structure (F3, 1LMT; Yamada *et al*., 1995), and their structures differed significantly from those of the true hits. In the plot of CYTOCHROME_C (Figures 3(d), and 4(h)), one of the true hits, the structure of the third match (62-67) in cytochrome c_7 (1CFO; Banci *et al*., 1996), adopted an exceptional structure. This structure, derived using NMR, is not well determined in this region (Banci *et al*., 1996). The other 144 true structures, including the other two matches from the entry 1CFO, were clustered into a small region with an rmsd of 0.87 Å. The distance between the true and false clusters (D_{TF}) was 1.57 and 1.79 Å, including and excluding 1CFO, respectively, showing easy discrimination between true and false hits.

AA_TRNA_LIGASE_II_2

The pattern AA_TRNA_LIGASE_II_2 is defined for class-II aminoacyl-tRNA synthetases (Lévêque *et al*., 1990; Cusack *et al*., 1991). Although 100 pattern matches were found, only 20 structures were true hits (labelled as T in Figure 3(e)). The large rmsd (2.93 Å) and distribution in Figure 3(e) is very different from the previous plots. The other 80 false matches were from dihydrofolate reductase (F1, 58 matches), phospholipase C (F2, 20 matches) and enol-acyl carrier protein reductase (F3, two matches). The structures of true hits were clustered into a single group with a very low rmsd of 0.37 Å. The false hits were scattered giving a high rmsd (2.52 Å). The distance between the true and false clusters (D_{TF}) was 1.20 Å, which is the distance between the clusters T and F2.

PEROXIDASE_1

For the PEROXIDASE_1 pattern (Henrissat *et al*., 1990) of 11 residues defined for the region including the histidine that is one of the ligands to the heme iron (the ''proximal'' histidine), eight false hits from cathepsin B were found. Most of the structures of true hits were clustered into single group (labelled as T1 in Figure 3(f)). In the true hits, there were four outlying structures from human (1MHL; Fenna *et al*., 1995) and canine (1MYP; Zeng & Fenna, 1992) myeloperoxidases (T2). However, these peroxidases are not thought to be related to the other peroxidases in the PDB (see below). The rmsd of the other 81 structures was 0.35 Å. The distance between the true and false clusters (D_{TF}) was 4.58 Å.

TRYPSIN_SER

The TRYPSIN_SER pattern, together with the TRYPSIN_HIS pattern, is defined for the active site of the trypsin serine protease family. The TRYPSIN_SER is a pattern of 12 residues including the active site serine residue (Rawlings & Barrett, 1994).

The TRYPSIN_SER pattern was found in 321 sequences from 282 PDB entries. All of the sequences were identified as true hits, coming from thrombin, trypsin, chymotrypsin, α-lytic protease, elastase, and other trypsin-like serine proteases. Figure 4(e) shows the results of the multidimensional scaling analysis of the structures of the TRYPSIN_SER pattern. In this case there is one major group of 304 fragments. The outlying 17 fragments were found to be from proenzymes, including trypsinogen and chymotrypsinogen. In Figure 4(e), proenzyme structures including the 17 fragments are plotted as open circles.

In these proteins, the structure of the pattern matches (the residue range 189-200) overlaps one of the regions which are known to undergo structural changes on activation; the regions, having flexible conformation in the proenzymes, from rigid structures in the active enzymes. Some of the residues in the pattern matches, especially Asp194, play important roles in the structural change on the activation (Huber & Bode, 1978).

Also in the major group, there are five trypsinogen (Walter *et al*., 1982; Marquart *et al*., 1983; Bode *et al*., 1984) and two chymotrypsinogen (Hecht *et al*., 1991) in complex with pancreatic trypsin inhibitors, which are known to have structures very similar to those of trypsin and chymotrypsin. The rmsd of the 304 structures of the major group was 0.38 Å. Thus these proenzymes have been induced to adopt the conformation of active enzymes by binding the inhibitors (Huber & Bode, 1978; Bode *et al*., 1978; Marquart *et al*., 1983).

ANNEXIN

Annexin has a characteristic motif in which 70 amino acid residues are repeated usually four times to form four homologous domains (I, II, III and IV) arranged in a cyclic array (Liemann & Huber, 1997). Annexin VI has the repeat eight times and is known to have two copies of the four-domain motif (Benz *et al*., 1996; Kawasaki *et al*., 1996). The pattern ANNEXIN is defined for the repeated regions, stretching over 53 residues (Fiedler & Simons, 1995).

In the structures of annexin I, III, IV, V, VI and XII in the PDB, the pattern matches for the ANNEXIN pattern were found in all repeats except for domain IV of annexin XII and domain III of the second half of annexin VI. The structure comparison was performed using only 20 C^{α} atoms, eliminating four long ''spacer'' regions. The structures were clustered into three groups. The first group (labelled T1 in Figure 4(n), 27 structures, rmsd 0.42 Å) consists of domain IV from all annexins. The second group (T2, 11 structures, 0.58 Å) consists of all structures of domain III of annexin IV and some structures of domain III of annexin V, and the third group consists of all the other struc-

tures (T3, 63 structures, 0.82 Å). The rmsd for all these structures is 1.39 Å, reflecting the presence of different clusters.

COPPER_BLUE

The eight-residue pattern COPPER_BLUE is defined for the copper ligand sites, including the cysteine and the histidine that are two of the ligands to the copper atom, of type-1 copper proteins (Rydén & Hunt, 1993). For COPPER_BLUE, the structures of the true hits were clustered into three distinct groups, as shown in Figure 4(s): 55 structures from azurin and rusticyanin (labelled as T1, rmsd 0.52 Å), 32 structures from plastocyanin and amicyanin (T2, 0.40 Å), and two structures from cucumber basic protein and cucumber stellacyanin (T3, 0.32 Å). Again the overall rmsd (1.30 Å) reflected the presence of distinct subgroups.

Discussion

For many of the patterns, the rmsd values for the true hits were small enough to consider that the true hits have a single common structure. When examined in detail, the subtle structural differences between groups of true hits were sometimes distinguishable, as shown in Figure 4. The three-dimensional structures of the EF_HAND pattern were found to have less ordered structures without the metal ion bound. For the TRYPSIN_SER pattern, the structures of proenzymes deviated from those of matured enzymes. The structures of the ASP_PROTEASE and ANNEXIN patterns in different domains were clustered into separate groups.

The search gave false hits for six patterns. In the plots for those patterns (Figure 3) the structures of the true hits were clearly separated from those of false hits. No false hits, except for the unidentified hits for the ATP_GTP_A pattern, were found to have similar structures to true hits, although the false hits have sequence similarity to some extent to the true hits within the matched fragments. The average structures of the clusters of the true hits or the false hits of some of the patterns are shown in Figure 5. The structures in Figure 5 were calculated by averaging the main-chain coordinates of all structures of each cluster. The results suggest that the structures of the true positives are characteristic for each pattern.

Although false positives were found in many cases, most of the pattern matches (ca. 87%) were true positives. This confirms that searching sequence patterns is a powerful method for identifying functions of proteins from their sequences. Excluding the ten ''common'' patterns in PROSITE (see Materials and Methods), the pattern ATP_GTP_A matched the largest number of sequences from many proteins. About 70% of the matches could not be identified definitely as either true or false by referring to the databases, although most of the unidentified matches are probably false hits. Therefore, it can be considered as a rather non-specific pattern.

For some of the other patterns, a few structural outliers were found in the true matches. As mentioned above, outliers were found for ASP_PROTEASE, CYTOCHROME_C and PEROXIDASE_1. Such outlying structures were also found in the true matches for LACTALBUMIN_LYSOZYME, LECTIN_LEGUME_ALPHA, and INSULIN, as shown in Figure 4.

For LECTIN_LEGUME_ALPHA, there was one outlying structure in the pattern matches (Figure 4(l)). The structure was from lectin-like α-amylase inhibitor (1DHK; Bompard-Gilles *et al.*, 1996), which has similar structure to the other legume lectins. However, in this case, the large structural difference of the preceding loop region causes considerable shifts in the positions of the atoms of the first residue (Val171) in the pattern match of the inhibitor structure, resulting in the large deviation from the other structures.

In the true hits for PEROXIDASE_1, there were four outlying structures from human (1MHL; Fenna *et al.*, 1995) and canine (1MYP; Zeng & Fenna, 1992) myeloperoxidases. In the pattern matches of myeloperoxidases, the residue corresponding to the proximal histidine in the pattern is His250, which, however, is actually not a ligand of the heme; the proximal histidine in these proteins is His336. Therefore, in the sequences of myeloperoxidases, and probably also in the sequences of the other vertebrate peroxidases, the pattern matches found for PEROXIDASE_1 should be considered as false hits from the aspect of three-dimensional structures.

In the two structural outliers for INSULIN, one is from an engineered protein, a disulphide isomer of human insulin (A6-20 of 1XGL; Hua *et al.*, 1995). This structure has two non-native disulphide bonds (A7-A11 and A6-B7 instead of native A6-A11 and A7-B7) and one native-like (A20-B19) disulphide bond, and the structure of the pattern match is largely perturbed by the non-native pairing of the disulphide bonds (Hua *et al.*, 1995).

The other outlier for INSULIN was from human insulin-like growth factor 1 (the residue range 47-61 of 2GF1; Cooke *et al.*, 1991). This structure was from solution structures determined by NMR experiments. In this structure, the region including the pattern match has irregular or not well-determined structure, as a consequence of the small number of distance constraints obtained from the NMR experiment. This is also the case for most of the other outlying structures, including the residue range 55-63 of hen lysozyme (1HWA; Smith *et al.*, 1993) for LACTALBUMIN_LYSOZYME, and the residue range 62-67 of cytochrome c_7 (1CFO; Banci *et al.*, 1996) for CYTOCHROME_C. This may in part reflect the pragmatic approach adopted here of using only one structure from an NMR ensemble.

Except for the rare outliers, the structures corresponding to sequence matches were clustered into

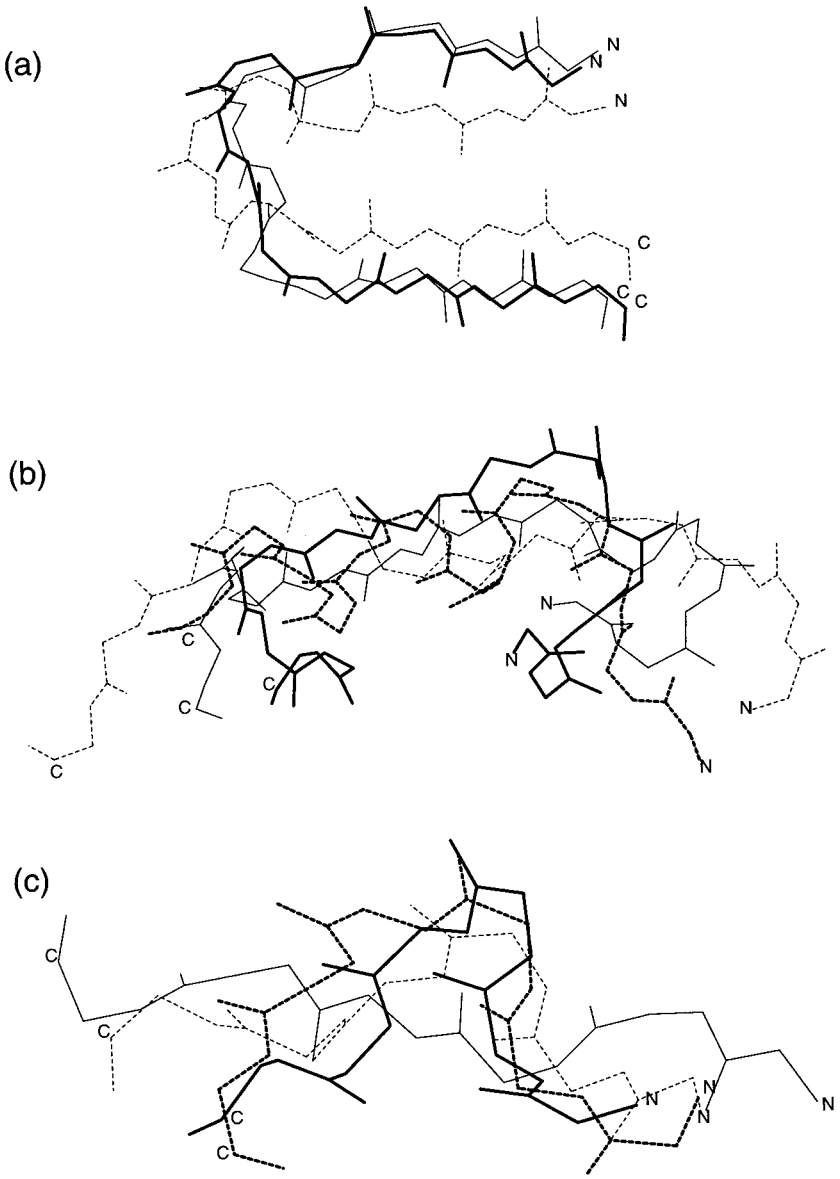

Figure 5. The average structures of the clusters. (a) ASP_PROTEASE: T1 (bold), T2 (thin) and F (broken); (b) EF_HAND: T (bond), F1 (thin), F2 (broken thin) and F3 (broken bold); (c) CYTOCHROME_C: T (bold), F1 (thin), F2 (broken thin) and F3 (broken bold).

discrete groups for each pattern. The structures in each group had a common structure with a small rmsd value. The common structure for each group can be used as a structural template for the protein function, in modelling and analysing the protein structures. The representative structures of the 20 patterns were shown in Figure 6.

The PROSITE patterns, defined mostly for enzyme active sites and ligand binding sites, are closely connected to the protein function, and often used in predicting these functions from sequence. By extending the patterns to include structural information, it will be possible to improve their efficiency in both detecting distantly related proteins and differentiating between true and false positives. The results of this study provide a structural basis for such modifications. For example, the structures of true hits for some patterns were found to cluster into more than one group. A structural motif could be derived separately for each cluster. The conserved structures may be helpful in modelling protein structures and validating modelled structures. Three-dimensional templates of the functional regions can be derived, which can be used as reference structures for the purpose.

This work clearly provides the detailed analysis required as the basis for creating a library of such structural templates based on PROSITE patterns. True matches were found in the PDB sequences for 537 patterns (42%) of the 1265 PROSITE patterns. As shown in Figure 6, those patterns are represented as structures in the PDB. Although 359 false hits (4% of the total 9493 hits) were found at the sequence level, it was almost always possible to differentiate true and false hits from the structural data; the structures of the false hits clustered into a separate group from the true structures. The next stage would be to explore if these 3D templates can uniquely identify the functional site in a protein structure, even if the sequence is modified, and to extend the functional site if the local structure is conserved.

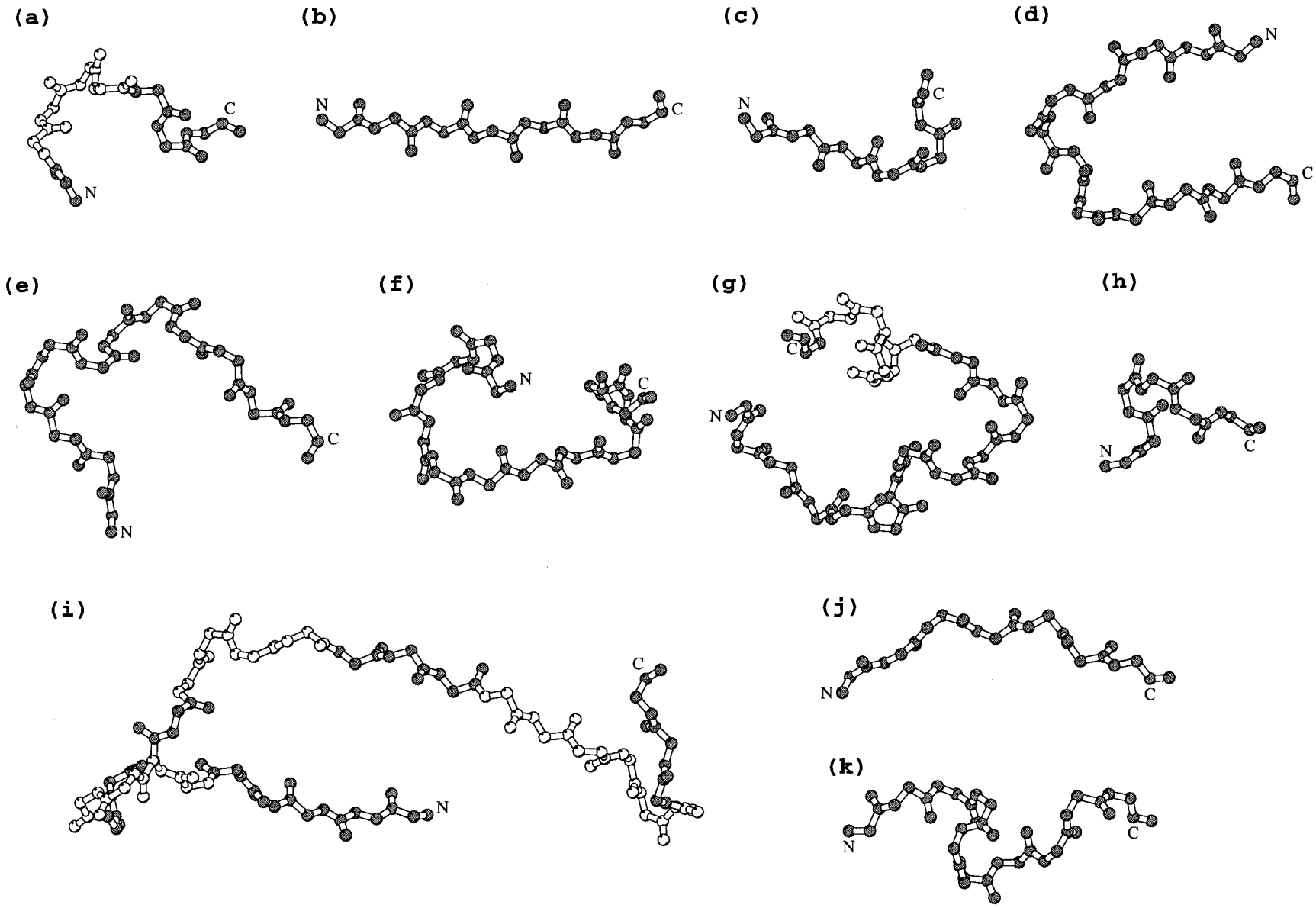

Figure 6 (*legend on page* 1688)

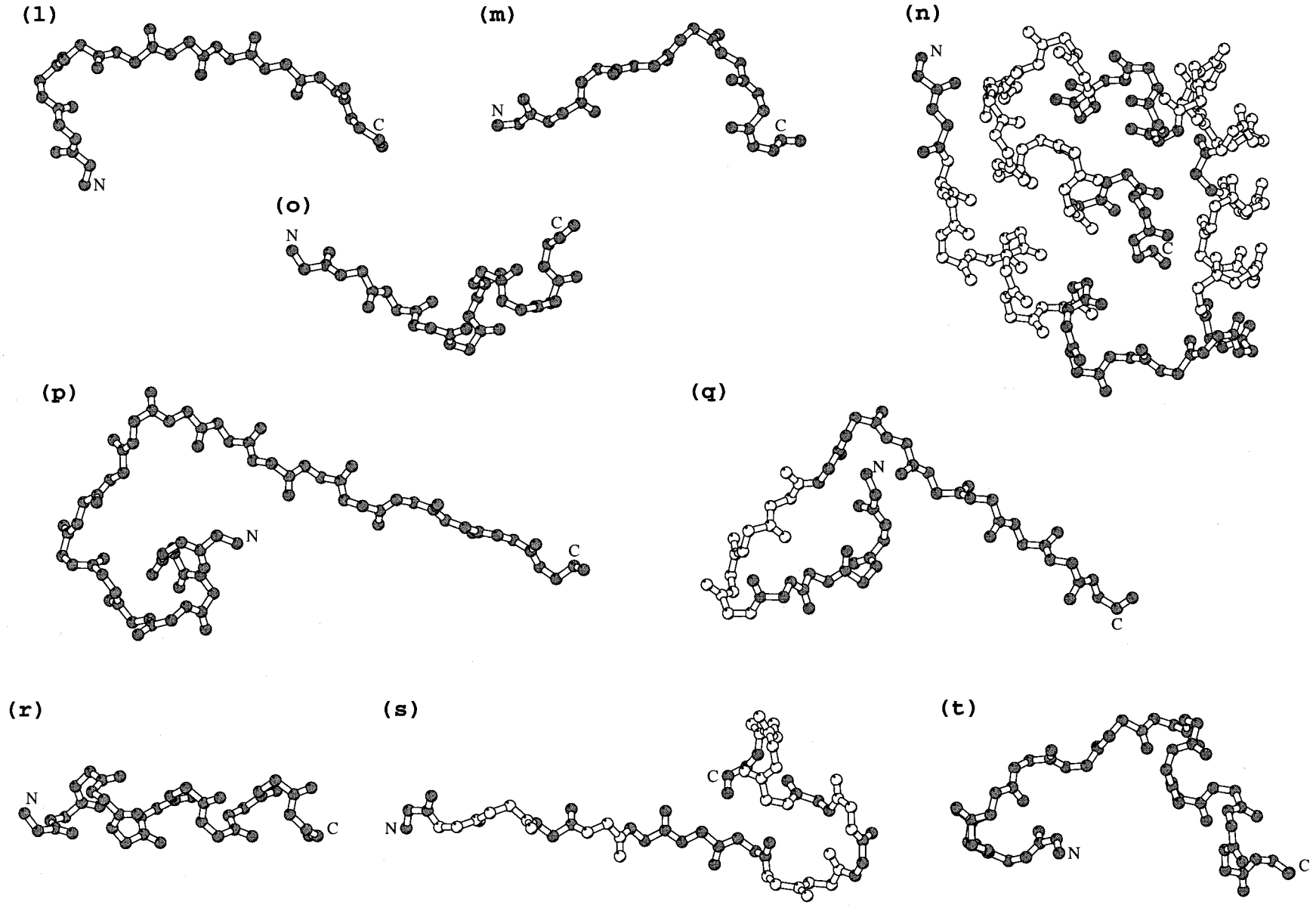

Figure 6 (*legend opposite*)

Materials and Methods

Pattern search

A sequence library, consisting of protein structures in the current PDB (Bernstein *et al*., 1977) was created. All protein structures were used including NMR structures, except for crystal structures determined to lower than 3.5 Å resolution, and theoretical models. Structures of the same protein from independent experiments under different conditions or in different states were included. Sequences of 11,432 polypeptide chains were taken from 6226 PDB entries into the 3D-sequence library and used for searching sequence patterns.

PROSITE release 14.0 (Bairoch *et al*., 1997) contained 1335 entries, 1275 of which were patterns. Although there are also four rule and 56 profile entries in PROSITE, we used only the patterns, in order to simplify the search and structure comparison procedure. Ten patterns annotated in PROSITE as describing sites commonly found in the majority of known protein sequences (for example the *N*-glycosylation site) were excluded. Pattern matches were sought in the sequence library for each of the remaining 1265 patterns. The results for the 20 patterns having the largest numbers of matches are shown in Table 1.

True and false identification

Each pattern match was examined whether it was a true or false positive according to the PROSITE database. PROSITE contains information of sequence pattern matches found in the SWISS-PROT database, annotated as true or false positives. All the annotations in PROSITE were used ''as is'', and no correction or further identification was made in this study. Although some entries in PROSITE have direct links to PDB entries, they are rather sparse and only for exemplification of true hits. The information about true and false positives is described in general only by the relevant SWISS-PROT codes in PROSITE. Therefore, to make use of the information in PROSITE, it was necessary to know the SWISS-PROT codes of the proteins in which sequence matches were found. The codes were obtained from SWISS-PROT release 35 (Bairoch & Apweiler, 1997) and the PDB. SWISS-PROT links only to PDB entry codes, and not to individual chains. For some proteins, there are no links between its SWISS-PROT and PDB entries. Therefore, some links were obtained by searching the PDB sequences against SWISS-PROT and using the highest identity matches from SWISS-PROT (Martin, A. C. R., private communication).

We used all possible combinations of links from these sources. The matches that could not be identified by following links in the databases were inspected separately. The inspection was done only based on the PROSITE descriptions for the pattern. Matches were identified as true if they were from sequences apparently belonging to the family of the pattern. Some matches from mutant sequences could be identified as false, in the case that their native sequences do not match the pattern. The other hits were left unidentified.

For some patterns, including the ATP_GTP_A pattern, which gave the maximum number of matches in the sequence search, matches were poorly identified because there were no links to SWISS-PROT in their PROSITE entries. Although some of the matches for the ATP_GTP_A pattern could be identified by referring to description of the pattern in PROSITE, many were left unidentified, because all proteins which bind to ATP or GTP do not necessarily have a P-loop, and there was no straightforward means to find out for certain whether each protein has a P-loop structure or not without checking their structures. Those unidentified matches were also included in the further analyses.

For the other patterns in Table 1, we inspected each of the 146 unidentified hits, and every match could be identified as true or false. In the inspected hits, only one sequence was identified as false: a pattern match for the CYTOCHROME_C pattern from a mutant sequence of lysozyme (1LMT; Yamada *et al*.,1995). All the others were identified as true.

Structure comparison

The structure comparisons were performed for all the sequence matches from the patterns in Table 1, including false positives. Structural similarities were evaluated by the rmsd of the main-chain C^{α} atoms. In the calculation, multiple conformations in a crystal structure were taken into account with equal weight. For NMR ensembles, only the minimised average structure, or, if not available, the first structure was used.

Most of the sequence patterns in Table 1 contain ''wildcard'' positions that can match any amino acid. Some of the patterns have more than one wildcard position consecutively, forming a ''spacer''. If a spacer has a very short, fixed length, it can be expected that its three-dimensional structure should be conserved, whereas for a long spacer the structure is more likely to be variable.

Figure 6. The main-chain atoms of the representative structure for each of the 20 patterns. In each case, the representative structure shown is the one having the smallest rmsd value compared with the average structure. Only the true hit structures, including any outlying structures, were considered. The grey atoms correspond to the residues used in the structure comparison and the average structure calculation. (a) ATP_GTP_A: residue range D10-17 of 6Q21 (Milburn *et al*., 1990). (b) IG_MHC: E171-177 of 1DLH (Stern *et al*., 1994). (c) TRYPSIN_HIS: 53-58 of 3EST (Meyer *et al*., 1988). (d) ASP_PROTEASE: B22-33 of 1AID (Rutenber *et al*., 1993). (e) TRYPSIN_SER: E189-200 of 1UCY (Martin *et al*., 1996). (f) EF_HAND: 54-66 of 4ICB (Svensson *et al*., 1992). (g) LACTALBUMIN_LYSOZYME: 76-94 of 1LZ3 (Harata, 1993). (h) CYTOCHROME_C: 12-17 of 1COR (Cai *et al*., 1992). (i) INTRADIOL_DIOXYGENAS: A51-79 of 3PCL (Orville *et al*., 1997b). (j) LECTIN_LEGUME_BETA: A119-125 of 1FAT (Hamelryck *et al*., 1996). (k) XYLOSE_ISOMERASE_2: A52-61 of 1XYC (Lavie *et al*., 1994). (l) LECTIN_LEGUME_ALPHA: B85-94 of 1TEI (Moothoo & Naismith, 1998). (m) XYLOSE_ISOMERASE_1: C180-187 of 2XIN (Jenkins *et al*., 1992). (n) ANNEXIN: 104-156 of 1HVE (Burger *et al*., 1994). (o) AA_TRNA_LIGASE_II_2: A306-315 of 1HTT (Arnez *et al*., 1995). (p) DNA_POLYMERASE_X: A179-198 of 1ZQO (Pelletier & Sawaya, 1996). (q) EUK_CO2_ANHYDRASE: 105-121 of 1RZB (Hakansson *et al*., 1994). (r) PEROXIDASE_1: 167-177 of 1CPE (Miller *et al*., 1994). (s) COPPER_BLUE: A105-121 of 1AZB (Shepard *et al*., 1993). (t) INSULIN: C6-20 of 1BEN (Smith *et al*., 1996). The Figure was generated using the MOLSCRIPT program (Kraulis, 1991).

In the calculation we included only spacers of fixed length shorter than or equal to three residues. For example, in the eight-residue pattern ATP_GTP_A, the C^{α} atoms of only four residues were used to calculate the rmsd, eliminating the four-residue spacer. If a spacer has a flexible length and the number of allowed matches is represented by a range in the pattern, residues that matched to the spacer were excluded from the rmsd calculation without regard to their actual lengths.

In most cases, a pattern match could be determined in only one way, but in a few cases, a pattern match could have more than one correspondence. For example, a short pattern ''A.{2} A'' (i.e. Ala-Any-Any-Ala) matches the sequence AAGAA in two ways, one from the first residue, and the other from the second residue. In practice, such an ambiguous match occurred often in the middle of a pattern, immediately after a variable length region. In the patterns having the largest numbers of matches (Table 1), COPPER_BLUE had such ambiguous matches. In these cases, the structures of all possible correspondences were examined and the one that had the most similar three-dimensional structure to the other matches was used.

The rmsd was calculated for each pair of structures by the least squares method using the McLachlan algorithm (McLachlan, 1982) implemented in the program ProFit (Martin, A. C. R., http://www.biochem.ucl.ac.uk/~martin/programs/#profit). The rmsd for a group of structures was calculated as the root mean square differences of all pairs of corresponding C^{α} atoms within the group.

There was one ATP_GTP_A structure (a mutant of $p21^{H\text{-}ras}$ protein, 1PLL; Scheidig *et al*., 1994) of which the coordinates of one of the C^{α} atoms (the C^{α} of Lys16) was not found in the PDB, and the structure was excluded from the comparison.

The rmsd in the structures were analysed using multidimensional scaling (Cox & Cox, 1994), projecting the results onto the two principal axes. The plots were projected with the first principal component (PC1) on the horizontal axis and the second (PC2) on the vertical.

Based on the results of multidimensional scaling analysis, the structures were clustered into groups by the single linkage method (Everitt, 1993). The structures were hierarchically clustered until the groups recognised by the multidimensional scaling analysis were obtained. The results of the cluster analyses were almost consistent with the multidimensional scaling, except for one outlying structure (2HVP; Navia *et al*., 1989) for the ASP_PROTEASE pattern, which was only found by the cluster analysis. In the analysis, the distance between two clusters was defined as the distance between their closest members, one from each group (nearest-neighbour distance). In order to represent the difference between the true and false clusters, the distance D_{TF}, defined as the distance or the minimum of the distances between the true and false clusters, was used to represent the difference between their clusters.

Acknowledgements

We thank Dr R. A. Laskowski for helpful discussion, constructive comments and corrections to the manuscript; Dr T. K. Attwood for comments on the sequence motif databases; Dr A. C. R. Martin for computer program ProFit and the cross-reference table between the PDB and SWISS-PROT. This work and one of the authors (A. K.) was supported by Sankyo Co., Ltd., Tokyo, Japan. We also acknowledge financial support from the U. K. BBSRC grant no. 31/JRI07365.

References

Ambler, R. P. (1991). Sequence variability in bacterial cytochromes-*c*. *Biochim. Biophys. Acta,* **1058**, 42-47.

Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8**, 675-683.

Arnez, J. G., Harris, D. C., Mitschler, A., Rees, B., Francklyn, C. S. & Moras, D. (1995). Crystal structure of histidyl-tRNA synthetase from *Escherichia coli* complexed with histidyl-adenylate. *EMBO J.* **14**, 4143-4155.

Attwood, T. K., Beck, M. E., Bleasby, A. J. & Parry-Smith, D. J. (1994). PRINTS - a database of protein motif fingerprints. *Nucl. Acids Res.* **22**, 3590-3596.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* **26**, 304-308.

Bairoch, A. (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucl. Acids Res.* **21**, 3097-3103.

Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucl. Acids Res.* **25**, 31-36.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Banci, L., Bertini, I., Bruschi, M., Sompornpisut, P. & Turano, P. (1996). NMR characterization and solution structure determination of the oxidized cytochrome c_7 from *Desulfuromonas acetoxidans*. *Proc. Natl Acad. Sci. USA,* **93**, 14396-14400.

Benz, J., Bergner, A., Hofmann, A., Demange, P., Göttig, P., Liemann, S., Huber, R. & Voges, D. (1996). The structure of recombinant human annexin VI in crystals and membrane-bound. *J. Mol. Biol.* **260**, 638-643.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **122**, 535-542.

Bode, W., Schwager, P. & Huber, R. (1978). The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. *J. Mol. Biol.* **188**, 99-112.

Bode, W., Walter, J., Huber, R., Wenzel, H. R. & Tschesche, H. (1984). The refined 2.2 Å (0.22-nm) X-ray crystal structure of the ternary complex formed by bovine trypsinogen, valine-valine and the Arg^{15} analogue of bovine pancreatic trypsin inhibitor. *Eur. J. Biochem.* **144**, 185-190.

Bompard-Gilles, C., Rousseau, P., Rougé, P. & Payan, F. (1996). Substrate mimicry in the active center of a mammalian α-amylase: structural analysis of an enzyme-inhibitor complex. *Structure,* **4**, 1441-1452.

Bork, P. & Koonin, E. V. (1996). Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366-376.

Bork, P., Ouzounis, C. & McEntyre, J. (1995). Ready for a motif submission? A proposed checklist. *Trends Biochem. Sci.* **20**, 104.

Burger, A., Voges, D., Demange, P., Perez, C. R., Huber, R. & Berendes, R. (1994). Structural and electro-

physiological analysis of annexin V mutants. mutagenesis of human annexin V, an *in vitro* voltage-gated calcium channel, provides information about the structural features of the ion pathway, the voltage sensor and the ion selectivity filter. *J. Mol. Biol.* **237**, 479-499.

Cai, M., Bradford, E. G. & Timkovich, R. (1992). Investigation of the solution conformation of cytochrome *c-551* from *Pseudomonas stutzeri*. *Biochemistry,* **31**, 8603-8612.

Cooke, R. M., Harvey, T. S. & Campbell, I. D. (1991). Solution structure of human insulin-like growth factor 1: a nuclear magnetic resonance and restrained molecular dynamics study. *Biochemistry,* **30**, 5484-5491.

Cox, T. F. & Cox, M. A. A. (1994). *Multidimensional Scaling*, Chapman & Hall, New York.

Cronet, P., Bellsolell, L., Sander, C., Coll, M. & Serrano, L. (1995). Investigating the structural determinants of the p21-like triphosphate and Mg^{2+} binding site. *J. Mol. Biol.* **249**, 654-664.

Cusack, S., Härtlein, M. & Leberman, R. (1991). Sequence, structural and evolutionary relationships between class 2 aminoacyl-rRNA synthetases. *Nucl. Acids Res.* **19**, 3489-3498.

Egloff, M.-P., Marguet, F., Buono, G., Verger, R., Cambillau, C. & van Tilbeurgh, H. (1995). The 2.46 Å resolution structure of the pancreatic lipase-colipase complex inhibited by a C11 alkyl phosphonate. *Biochemistry,* **34**, 2751-2762.

Everitt, B. S. (1993). *Cluster Analysis*, 3rd edit., Edward Arnold, London.

Fenna, R., Zeng, J. & Davey, C. (1995). Structure of the green heme in myeloperoxidase. *Arch. Biochem. Biophys.* **316**, 653-656.

Fiedler, K. & Simons, K. (1995). Annexin homologues in *Giardia lamblia*. *Trends Biochem. Sci.* **20**, 177-178.

Hakansson, K., Wehnert, A. & Liljas, A. (1994). X-ray analysis of metal-substituted human carbonic anhydrase II derivatives. *Acta Crystallog. sect. D,* **50**, 93-100.

Hamelryck, T. W., Dao-Thi, M.-H., Poortmans, F., Chrispeels, M. J., Wyns, L. & Loris, R. (1996). The crystallographic structure of phytohemagglutinin-L. *J. Biol. Chem.* **271**, 20479-20485.

Harata, K. (1993). X-ray structure of monoclinic turkey egg lysozyme at 1.3 Å resolution. *Acta Crystallog. sect. D,* **49**, 497-504.

Hecht, H. J., Szardenings, M., Collins, J. & Schomburg, D. (1991). Three-dimensional structure of the complexes between bovine chymotrypsinogen A and two recombinant variants of human pancreatic secretory trypsin inhibitor (Kazal-type). *J. Mol. Biol.* **220**, 711-722.

Henikoff, S. & Henikoff, J. G. (1994). Protein family classification based on searching a database of blocks. *Genomics,* **19**, 97-107.

Henikoff, S., Pietrokovski, S. & Henikoff, J. G. (1998). Superior performance in protein homology detection with the Blocks Database servers. *Nucl. Acids Res.* **26**, 309-312.

Henrissat, B., Saloheimo, M., Lavaitte, S. & Knowles, J. K. C. (1990). Structural homology among the peroxidase enzyme family revealed by hydrophobic cluster analysis. *Proteins: Struct. Funct. Genet.* **8**, 251-257.

Hermoso, J., Pignol, D., Kerfelec, B., Crenon, I., Chapus, C. & Fontecilla-camps, J. C. (1996). Lipase activation by nonionic detergents - the crystal-structure of the porcine lipase-colipase-tetraethylene glycol monooctyl ether complex. *J. Biol. Chem.* **271**, 18007-18016.

Hua, Q.-X., Gozani, S. N., Chance, R. E., Hoffmann, J. A., Frank, B. H. & Weiss, M. A. (1995). Structure of a protein in a kinetic trap. *Nature Struct. Biol.* **2**, 129-138.

Huber, R. & Bode, W. (1978). Structural basis of the activation and action of trypsin. *Acc. Chem. Res.* **11**, 114-122.

Jenkins, J., Janin, J., Rey, F., Chiadmi, M., van Tilbeurgh, H., Lasters, I., De Maeyer, M., Van Belle, D., Wodak, S. J., Lauwereys, M., Stanssens, P., Mrabet, N. T., Snauwaert, J., Matthyssens, G. & Lambeir, A.-M. (1992). Protein engineering of xylose (glucose) isomerase from *Actinoplanes missouriensis*. 1. Crystallography and site-directed mutagenesis of metal binding sites. *Biochemistry,* **31**, 5449-5458.

Kawasaki, H., Avilasakar, A., Creutz, C. E. & Kretsinger, R. H. (1996). The crystal-structure of annexin-VI indicates relative rotation of the 2 lobes upon membrane-binding. *Biochim. Biophys. Acta,* **1313**, 277-282.

Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Lavie, A., Allen, K. N., Petsko, G. A. & Ringe, D. (1994). X-ray crystallographic structures of D-xylose isomerase-substrate complexes position the substrate and provide evidence for metal movement during catalysis. *Biochemistry,* **33**, 5469-5480.

Lévêque, F., Plateau, P., Dessen, P. & Blanquet, S. (1990). Homology of *lysS* and *lysU*, the two *Escherichia coli* genes encoding distinct lysyl-tRNA synthetase species. *Nucl. Acids Res.* **18**, 305-312.

Liemann, S. & Huber, R. (1997). Three-dimensional structure of annexins. *Cell. Mol. Life Sci.* **53**, 516-521.

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallog. sect. A,* **38**, 871-873.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallog. sect. B,* **39**, 480-490.

Marsden, B. J., Shaw, G. S. & Sykes, B. D. (1990). Calcium binding proteins. Elucidating the contributions to calcium affinity from an analysis of species variants and peptide fragments. *Biochem. Cell Biol.* **68**, 587-601.

Martin, P. D., Malkowski, M. G., DiMaio, J., Konishi, Y., Ni, F. & Edwards, B. F. (1996). Bovine thrombin complexed with an uncleavable analog of residues 7-19 of fibrinogen Aα: geometry of the catalytic triad and interactions of the P1′, P2′, and P3′ substrate residues. *Biochemistry,* **35**, 13030-13039.

Mathews, F. S. (1985). The structure, function and evolution of cytochromes. *Prog. Biophys. Mol. Biol.* **45**, 1-56.

Meyer, E., Cole, G., Radhakrishnan, R. & Epp, O. (1988). Structure of native porcine pancreatic elastase at 1.65 Å resolution. *Acta Crystallog. sect. B,* **44**, 26-38.

Milburn, M. V., Tong, L., deVos, A. M., Brünger, A., Yamaizumi, Z., Nishimura, S. & Kim, S.-H. (1990). Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic *ras* proteins. *Science,* **247**, 939-945.

Miller, M. A., Han, G. W. & Kraut, J. (1994). A cation binding motif stabilizes the compound I radical of cytochrome *c* peroxidase. *Proc. Natl Acad. Sci. USA,* **91**, 11118-11122.

Moothoo, D. N. & Naismith, J. H. (1998). Concanavalin A distorts the β-GlcNAc-(1 → 2)-Man linkage of β-GlcNAc-(1 → 2)-α-Man(1 → 3)-[β-GlcNAc-(1 → 2)-α-Man-(1 → 6)]-Man upon binding. *Glycobiology*, **8**, 173-181.

Muraki, M., Harata, K. & Jigami, Y. (1992). Dessection of the functional role of structural elements of tyrosine-63 in the catalytic action of human lysozyme. *Biochemistry*, **31**, 9212-9219.

Nakayama, S. & Kretsinger, R. H. (1994). Evolution of the EF-hand family of proteins. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 473-507.

Navia, M. A., Fitzgerald, P. M. D., McKeever, B. M., Leu, C.-T., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L. & Springer, J. P. (1989). Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature*, **337**, 615-620.

Ohlendorf, D. H., Orville, A. M. & Lipscomb, J. D. (1994). Structure of protocatechuate 3,4-dioxygenase from *Pseutomonas aeruginosa* at 2.15 Å resolution. *J. Mol. Biol.* **244**, 586-608.

Orville, A. M., Elango, N., Lipscomb, J. D. & Ohlendorf, D. H. (1997a). Structures of competitive inhibitor complexes of protocatechuate 3,4-dioxygenase: multiple exogenous ligand binding orientations within the active site. *Biochemistry*, **36**, 10039-10051.

Orville, A. M., Lipscomb, J. D. & Ohlendorf, D. H. (1997b). Crystal structures of substrate and substrate analog complexes of protocatechuate 3,4-dioxygenase: endogenous Fe^{3+} ligand displacement in response to substrate binding. *Biochemistry*, **36**, 10052-10066.

Pelletier, H. & Sawaya, M. R. (1996). Characterization of the metal-ion binding helix-hairpin-helix motifs in human DNA polymerase-β by X-ray structural-analysis. *Biochemistry*, **35**, 12778-12787.

Perona, J. J. & Craik, C. S. (1995). Structural basis of substrate specificity in the serine proteases. *Protein Sci.* **4**, 337-360.

Rao, J. K. M., Erickson, J. W. & Wlodawer, A. (1991). Structural and evolutionary relationships between retroviral and eucaryotic aspartic proteinases. *Biochemistry*, **30**, 4663-4671.

Rastan, S. & Beeley, L. J. (1997). Functionional genomics: going forward from the databases. *Curr. Opin. Genet. Dev.* **7**, 777-783.

Rawlings, N. D. & Barrett, A. J. (1994). Families of serine peptidases. *Methods Enzymol.* **244**, 19-61.

Rawlings, N. D. & Barrett, A. J. (1995). Families of aspartic peptidases, and those of unknown catalytic mechanism. *Methods Enzymol.* **248**, 105-120.

Rutenber, E., Fauman, E. B., Keenan, R. J., Fong, S., Furth, P. S., Ortiz de Montellano, P. R., Meng, E., Kuntz, I. D., DeCamp, D. L., Salto, R., Rose, J. R., Craik, C. S. & Stroud, R. M. (1993). Structure of a non-peptide inhibitor complexed with HIV-1 protease. Developing a cycle of structure-based drug design. *J. Biol. Chem.* **268**, 15343-15346.

Rydén, L. G. & Hunt, L. T. (1993). Evolution of protein complexity: the blue copper-containing oxidases and related proteins. *J. Mol. Evol.* **36**, 41-66.

Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990). The P-loop-a common motif in ATP- and GTP- binding proteins. *Trends Biochem. Sci.* **15**, 430-434.

Scheidig, A., Sanchez-Llorente, A., Lautwein, A., Pai, E. F., Corrie, J. E. T., Reid, G. P., Wittinghofer, A. & Goody, R. S. (1994). Crystallographic studies on p21H-ras using the synchrotron Laue method: improvement of crystal quality and monitoring of the GTPase reaction at different time points. *Acta Crystallog. sect. D*, **50**, 512-520.

Shepard, W. E. B., Kingston, R. L., Anderson, B. F. & Baker, E. N. (1993). Structure of apo-azurin from *Alcaligenes denitrificans* at 1.8 Å resolution. *Acta Crystallog. sect. D*, **49**, 331-343.

Smith, G. D., Ciszak, E. & Pangborn, W. (1996). A novel complex of a phenolic derivative with insulin: Structural features related to the T → R transition. *Protein Sci.* **5**, 1502-1511.

Smith, L. J., Sutcliffe, M. J., Redfield, C. & Dobson, C. M. (1993). Structure of hen lysozyme in solution. *J. Mol. Biol.* **229**, 930-944.

Stern, L. J., Brown, J. H., Jardetzky, T. J., Gorga, J. C., Urban, R. G., Strominger, J. L. & Wiley, D. C. (1994). Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*, **368**, 215-221.

Strynadka, N. C. J. & James, M. N. G. (1989). Crystal structures of the helix-loop-helix calcium binding proteins. *Annu. Rev. Biochem.* **58**, 951-998.

Svensson, L. A., Thulin, E. & Forsen, S. (1992). Proline *cis-trans* isomers in calbindin $D_{9\,K}$ observed by X-ray crystallography. *J. Mol. Biol.* **223**, 601-606.

van Tilbeurgh, H., Egloff, M.-P., Martinez, C., Rugani, N. R., Verger, R. & Cambillau, C. (1993). Interfacial activation of the lipase-procolipase complex by mixed micelles revealed by X-ray crystallography. *Nature*, **362**, 814-870.

Walker, J. E., Saraste, M., Runswick, M. J. & Gray, N. J. (1982). Distantly-related sequences in the α- and β-subunits of ATP-synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945-951.

Walter, J., Steigemann, W., Singh, T. P., Bartunik, H., Bode, W. & Huber, R. (1982). On the disordered activation domain in trypsinogen: chemical labelling and low-temperature crystallography. *Acta Crystallog. sect. B*, **38**, 1462-1472.

Withers-Martinez, C., Carriere, F., Verger, R., Bourgeois, D. & Cambillau, C. (1996). A pancreatic lipase with a phospholipase A1 activity: crystal structure of a chimeric pancreatic lipase-related protein 2 from guinea pig. *Structure*, **4**, 1363-1374.

Yamada, T., Song, H., Inaka, K., Shimada, Y., Kikuchi, M. & Matsushima, M. (1995). Structure of a conformationally constrained Arg-Gly-Asp sequence inserted into human lysozyme. *J. Biol. Chem.* **270**, 5687-5690.

Zeng, J. & Fenna, R. E. (1992). X-ray crystal structure of canine myeloperoxidase at 3 Å resolution. *J. Mol. Biol.* **226**, 185-207.

Edited by F. E. Cohen

(Received 3 July 1998; received in revised form 20 January 1999; accepted 26 January 1999)