Quarterly Reviews of Biophysics **29**, 2 (1996), pp. 119–167
Printed in Great Britain

Structure and distribution of modules in extracellular proteins

PEER BORK[1], A. KRISTINA DOWNING[2], BRUNO KIEFFER[3] AND IAIN D. CAMPBELL[2]

[1] *Max-Delbrück-Center for Molecular Medicine, 13125 Berlin-Buch, Germany and European Molecular Biology Laboratory, 69117 Heidelberg, Germany*
[2] *Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK*
[3] *Groupe de Cancerogenese, IBMC de CNRS, 75084 Strasbourg, France*

1. INTRODUCTION

It has become standard practice to compare new amino-acid and nucleotide sequences with existing ones in the rapidly growing sequence databases. This has led to the recurring identification of certain sequence patterns, usually corresponding to less than 300 amino-acids in length. Many of these identifiable sequence regions have been shown to fold up to form a 'domain' structure; they are often called protein 'modules' (see definitions below). Proteins that contain such modules are widely distributed in biology, but they are particularly common in extracellular proteins.

Until a few years ago, structural studies of intact proteins with extracellular parts had proved difficult, partly because these proteins are often glycosylated, membrane spanning, large and flexible. The 3D structures of many of their modular components are, however, now known. This advance has come about, not only because of improvements in crystallography and NMR methods, but also because recombinant methods now facilitate the production of portions of the protein that contain identifiable modules. This 'dissection' approach is leading to a very rapid increase in knowledge of module 3D structure. This is thus an appropriate time to analyse recurring modules in terms of their sequences and 3D structures. Some aspects of this subject have already been covered extensively (Bork & Bairoch, 1995; Baron *et al.* 1991; Doolittle, 1995; Campbell & Downing, 1994). The aim of this review is to summarise information about the distribution of modules, to consider detectable relationships between sequences and structures and to look for patterns in the ways that modules can be fitted together in an intact protein. Because of space limitations, this review is limited in scope. The main restrictions are a concentration on modules found in extracellular proteins with known structure. Nevertheless, these alone cover a substantial fraction of known modules, since, for more than 50% of all described extracellular module families, there is at least one member with known three-dimensional structure. Furthermore, we estimate that about 40% of mammalian protein sequences are either completely extracellular or have an extracellular part. Many of them contain modules. The intra-/extra-cellular division for protein modules is not always clear since a few modules occur both outside and inside the cell but, in general, this distinction is useful and convenient. This review will show, largely by illustrations,

the distribution, structure and consensus sequences of extracellular modules whose coordinates are in the public domain.

2. SOME DEFINITIONS

Domain A domain is probably best defined as a spatially distinct structural unit that usually folds independently. In this definition, the sequence need not be contiguous. (In the absence of structural knowledge, the term domain is less clear although it is often used in current literature to describe functional units and sequence segments with characteristic features.)

Module Modules are a subset of domains that are contiguous in sequence and that are repeatedly used as 'building blocks' in functionally diverse proteins. They have identifiable amino-acid patterns that can be described by a 'consensus' sequence. The spread of a module in biological systems is likely to have been the result of a genetic shuffling mechanism that is not merely gene duplication and gene fusion. The presence of an identified sequence region in another, otherwise unrelated, protein and its location between other known modules, are strong indicators for a module. While the existence of exons with compatible phases at the exon/intron boundaries is good evidence that a module has been spread by 'exon shuffling' (Patthy, 1991), other means of module spreading could exist; thus the term 'module' may not only apply to sequence regions with compatible phases at exon boundaries. In the case of extracellular modules compatible boundary phases are in fact very often observed.

Repeat Modules frequently occur in tandem arrays. Nevertheless, the definition of module given above still applies and we will refer to 'repeats' as units that cannot occur as single copies but form super-structures. An example is the leucine rich repeat (Kobe & Deisenhofer, 1994, 1995).

Mosaic Proteins that are composed of several different module types are often referred to as 'mosaic' proteins. The biological role of modules may, or may not, vary in different settings.

3. EVOLUTION AND CELLULAR LOCATION OF MODULES

The major advantage of building proteins from modules is probably to facilitate the creation of new proteins during evolution. Phylogenetically 'old' proteins such as metabolic enzymes are usually only composed of one or two domains and the creation of a new enzyme during evolution required a gene duplication and numerous subsequent point mutations to acquire a new function for a given fold. Although modules are not found in phylogenetically old enzymes, they can already be observed in prokaryotes. However, some of the best-known bacterial modular proteins, such as the extracellular glycohydrolases, may have appeared relatively late in evolution as they are only found in rather specialised bacteria. In eukaryotes, intracellular modules mostly occur in the cytoskeleton and signal transduction pathways that do not seem to have equivalent counterparts in bacteria. Many domains have been known for a long time in nuclear proteins;

more and more of these turn out to fit the definition of a module given above. The best known intracellular modules, including several DNA binding domains, and the SH2, SH3 and PH (see e.g. Pawson, 1995) modules involved in signal transduction, are probably only a small sub-set of those that exist. To date, more than 30 cytoplasmic modules, have been described and many more will no doubt be found in the near future.

The largest fraction of modular proteins is, however, extracellular and seems to have evolved with the radiation of invertebrates. Mammals appear to have the largest fraction of extracellular proteins and most of them contain modules. Since many of the extracellular modules contain disulphide bridges, and thus cannot be located in the nucleus or cytoplasm, and since only a few other modules occur both outside and inside the cell, the distinction made in this review is useful.

4. EXTRACELLULAR MODULES AND THEIR BIOLOGICAL ROLE

In current sequence databases, about 60 of the abundant modules found in extracellular proteins seem to fit our definition of a module. Each of them can be separately identified by sequence analysis and by their detection in other, unrelated, extracellular proteins; this is a first indication that a protein has a modular architecture. The lengths of these modules vary from about 30 amino acids to over 300. The larger ones may have evolved by incorporation of smaller domains. Fig. 1 illustrates the nomenclature used for modules here. It also summarises the module size, the number of disulphide bonds and the number of times the module is found in the sequence database (Bork & Bairoch, 1995).

The abbreviations used are: first column – a 3 to 5 letter variant for unique identification in databases; second column – a two letter variant) the two letter code and the defined colour codes will be used in the various cartoons of modular proteins. Some other symbols used are explained in subsequent figures. The topology of modules with known three-dimensional structure (discussed in detail below) is indicated by the approximate arrangement of their secondary structure elements: a/b (α/β class), b (all-β), a (all-α) b+a ($\beta+\alpha$). b?a indicates a predominantly β class module but where an α helix has been observed in some members of the family. The size is a rough estimate of the module length given in amino acids (rounded up to factors of 10). Single members of a module family may differ considerably. The number of cysteines (Nb Cys) may vary within a module family. 4/6 means 4 or 6 cysteines have been observed; 4–6 means that 4 to 6 cysteines may occur. The number of module occurrences (Occ column) in databases (excluding species redundancies) is mainly based on queries from 1993. The actual number might now be much higher for some of the modules.

The various modules discussed here occur in functionally diverse proteins. The modules interact, singly or in concert with others, with a wide variety of ligands, including proteins, peptides and carbohydrates. While there often seems to be a relatively unique role for a given module in intracellular proteins, e.g. SH2 domains bind phosphotyrosine peptides (Pawson, 1995), this kind of direct correspondence is not always recognisable in the modules illustrated in Fig. 1.

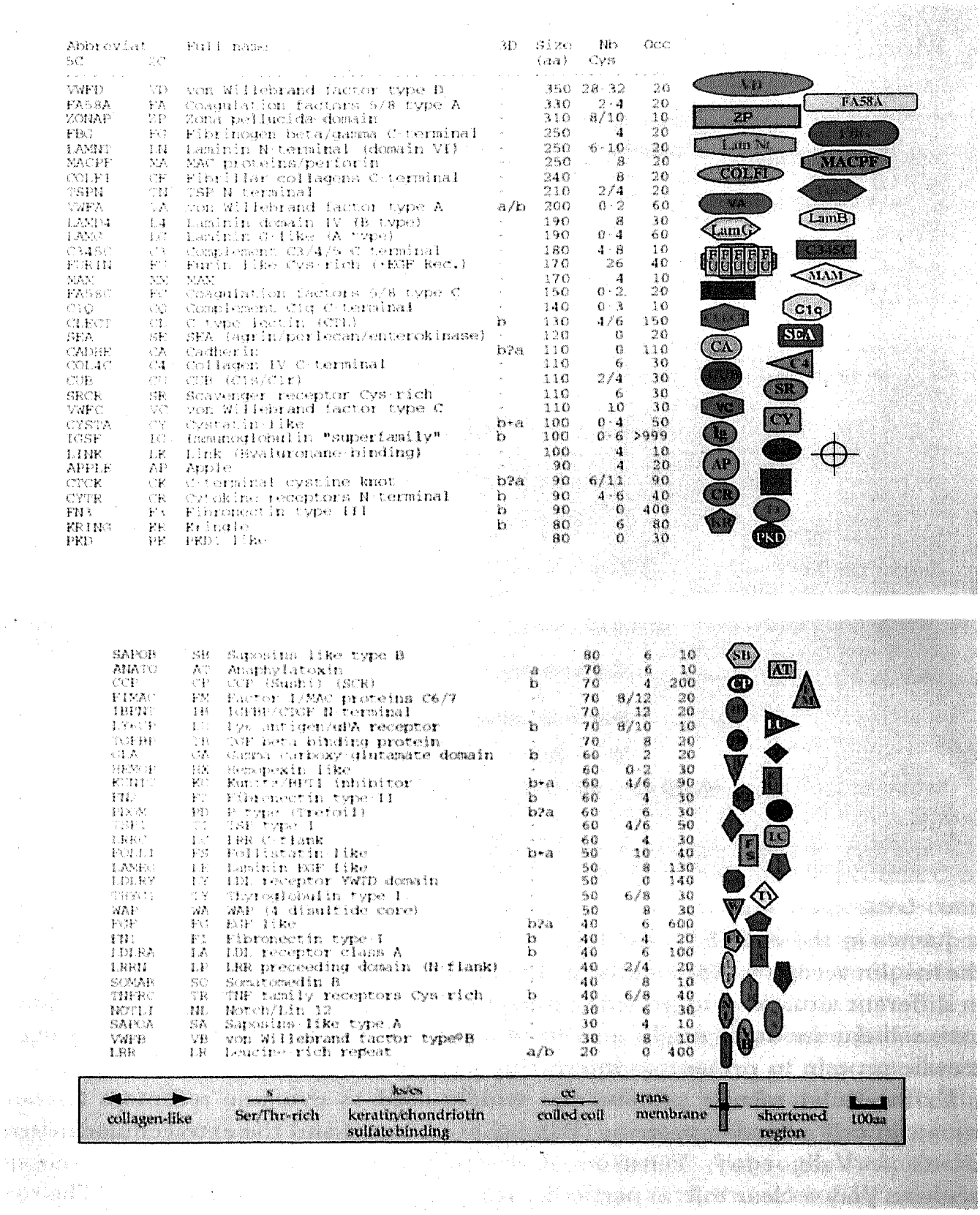

Fig. 1. Overview of extracellular modules and their abbreviations (for further details see Bork & Bairoch, 1995).

Only in a few cases has a unique function been reported e.g. the carbohydrate-binding of CLECT or the membrane-binding of GLA. For others it becomes clear that their function varies in different proteins, e.g. cell binding via an RGD

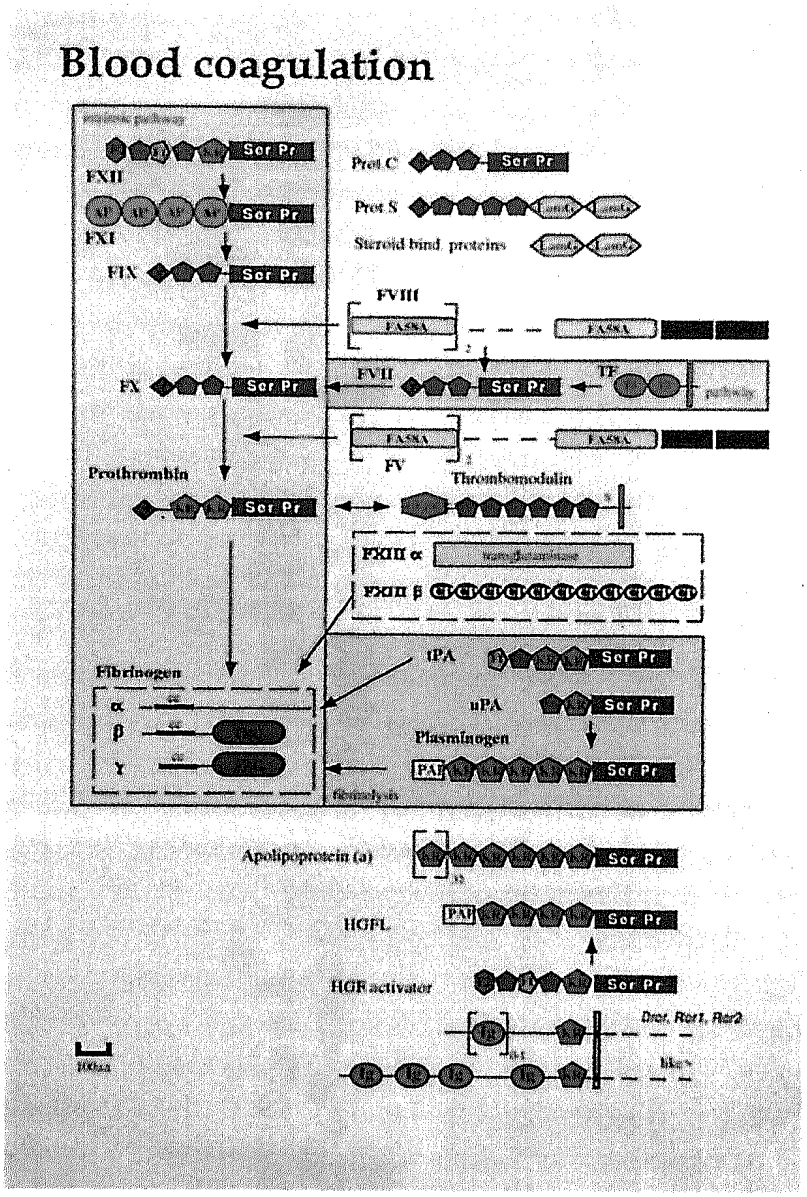

Fig. 2. Cartoon of the mosaic proteins involved in blood coagulation and fibrinolysis with individual pathways highlighted by blocks. Some regulatory proteins and proteins with a similar modular architecture to proteins within the cascades are also displayed. For nomenclature of modules see Fig. 1.

sequence in the 10th FN3 module of fibronectin versus dimerisation of FN3s in the insulin receptor. Apparently, different parts of the module surface can be used in different situations to provide interaction sites, often for other proteins. Some extracellular modules might also have a purely structural role, thus allowing a mosaic protein to present an interacting surface in an appropriate position.

Extracellular mosaic proteins are widely used as cytokine receptors (Bazan, 1993), in cell adhesion proteins (Barclay *et al.* 1993) and the extracellular matrix (Kreis & Vale, 1993; Venstrom & Reichardt, 1993). In some cases, mosaic proteins play a clear role in particular extracellular biological pathways. The two best-studied cases are the blood coagulation/fibrinolysis and the complement systems. The blood coagulation cascade (Fig. 2) is a host defence system that is initiated after blood vessel injury (Patthy, 1993). It comprises alternative pathways in which certain plasma proenzymes are successively activated by cleaving the protease domain from the modular N-terminal regulatory chains. It involves the formation of complexes with non-proteolytic plasma proteins and membrane-associated cofactors that, in the final step, leads to the formation of a blood clot.

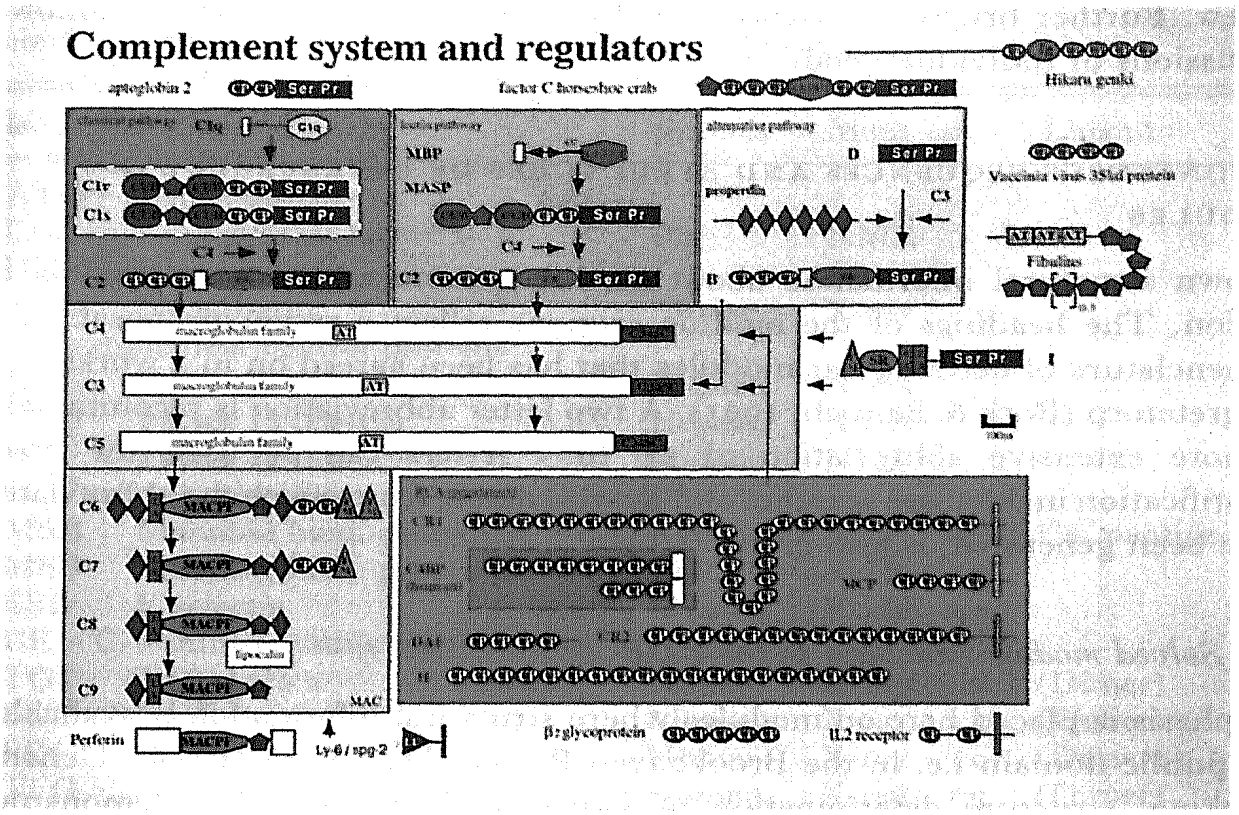

Fig. 3. Cartoon of the complement cascade and some of its regulators. Individual pathways are highlighted by blocks. Proteins with a similar modular architecture to proteins within the cascades are also displayed (modified from Bork & Bairoch, 1995). For nomenclature see Fig. 1.

During wound healing, the clot is dissolved by proteases of the fibrinolytic system (Fig. 2). Clotting and fibrinolysis are under multiple control. In addition to protease inhibitors, one anticoagulant pathway involves activation of protein C by thrombomodulin. Activated protein C then forms a membrane-associated complex with plasma protein S that inactivates factor VII. The level of protein S is regulated by C4 binding protein, a regulator that also inhibits the complement system.

The complement cascade (Fig. 3) is a defence system against infectious agents and plays an important role in inflammation. Via formation of the membrane attack complex (MAC) it finally leads to lysis of the membrane of infecting organisms. The classical pathway is triggered by immunoglobulins that recognise the foreign organisms; other pathways exists (Fig. 3). Although the proteins involved are biochemically well-characterised (Figs. 2, 3), the role of all modules in the participating proteins of these cascades is not yet completely understood in both systems. Only approximate regulatory functions such as Ca-dependent membrane-binding can be assigned to specific modules.

A specific biological function cannot be directly associated with all modules, partly due to lack of information and partly because, in some cases, modules do not function autonomously. Nonetheless, structural information is providing considerable insight into module function and hence forms the focus of this

review. Further brief comments on module function are given in the following discussions of individual modules.

5. CONSENSUS SEQUENCES AND STRUCTURES OF EXTRACELLULAR MODULES

Known structural information about module structures is summarised in this section. The headings of the module sections reflect a recent proposal on the nomenclature of extracellular modules that has been agreed on in a workshop in Magretetorp (Bork & Bairoch, 1995). A two letter abbreviation is recommended; a more extensive abbreviation of up to 5 letters can be used for unique identification in database searches and the full name, from which the abbreviations have been generated, is also given.

5.1 *Solved module structures in the PDB*

Emphasis is placed here on modules where structural information is available in the public domain i.e. in the Brookhaven Protein Data Bank (PDB). Generally speaking, structure determination of sequence homologues has demonstrated considerable similarity in a core structure, with insertions and deletions usually being found only in loop regions, although additional secondary structure elements can also be added (discussed further below). In the illustrations here, one structure has been selected as representative of each type of module. Where available, PDB accession codes for structures of all known homologues are shown (Table 1). In the figures, the secondary structure of each module, derived from its observed pattern of backbone hydrogen bonding, is shown schematically with a MOLSCRIPT figure (Kraulis, 1991). Atomic resolution illustrations for each module produced using the program INSIGHT (Biosym, Inc.) are also included which highlight conserved amino acids. Consensus sequences, derived using multiple sequence alignments, are illustrated for each module type using the following notation: h: hydrophobic; t: turn; a: aromatic; +/− single positively/negatively charged residue; o: Ser/Thr; s: small; -: single insertion; capital letter: one letter amino acid code; c: Cys not always involved in a disulphide bond. The apparent structural role of some consensus residues is discussed briefly in the text for each module, with emphasis placed on residues with roles in fold stabilisation.

While the number of known module structures is growing rapidly, it is still hard to make very rigid classifications among different structural types. Some, however, have clearly similar overall topology; many have a β-sandwich structure with N- and C-termini at opposite ends of one long edge of an ellipsoid-shape (e.g. Ig, FN3, CYTR, CAD and CCP); others are small and cystine-rich with no very obvious hydrophobic core (e.g. EGF, FN1 and TNFR); others have mixed secondary structure and N- and C-termini that are close together in space (e.g. kringle, Ly-6 and C-type). In the following list of modules, some attempt has been made to classify them on the basis of size, secondary structure and the proximity

Table 1. *Available 3D structures of extracellular modules*

Code	Protein (domain no)	Species	Meth	Å	Ligand
ANATO – anaphylatoxins					
1C5a	Complement factor V	Pig	NMR		
CADHE – cadherins					
1NCG	N-cadherin	Mouse	X-ray	1·9	
CCP – complement control proteins					
1HCC	Factor H (16)	Human	NMR		
1HFH	Factor H (15–16)	Human	NMR		
1HFI	Factor H (15)	Human	NMR		
CLECT – C-type lectins					
2MSB	Mannose bind. protein A	Rat	X-ray	1·7	Ca, peptide
1MSB	Mannose bind. protein A	Rat	X-ray	2·3	
1ESL	E-selectin	Human	X-ray	2·0	
CTCK – C-terminal cystine knots					
2TGI	TGF beta 2	Human	X-ray	1·8	(Dimer)
1BET	NGF beta	Mouse	X-ray	2·3	(Dimer)
1PDG	PDGF beta b	Human	X-ray	3·0	(Dimer)
1TFG	TGF beta 2	Human	X-ray	1·9	(Dimer)
CYSTA – cystatins					
1CEW	Cystatin	Chicken	X-ray	2·0	
1STF	Steffin B	Papaya	X-ray	2·4	Papain
CYTR – cytokine receptor N-terminal					
3HHR	Growth hormone receptor	Human	X-ray	2·8	Growth hormone
EGF					
1APO	Coagulation factor X	Bovine	NMR		
1CCF	Coagulation factor X	Bovine	NMR		Ca
1EGF	EGF	Mouse	NMR		
1EPG	EGF (ph2)	Mouse	NMR		
1EPH	EGF (ph2)	Mouse	NMR		
1EPI	EGF (ph6·8)	Mouse	NMR		
1EPJ	EGF (ph6·8)	Mouse	NMR		
1IXA	Coagulation factor XI	Human	NMR		
1HRE	Heregulin alpha	Human	NMR		
1HRF	Heregulin alpha	Human	NMR		
2TGF	TGF alpha	Human	NMR		
3TGF	TGF alpha	Human	NMR		
3EGF	EGF	Mouse	NMR		
4TGF	TGF alpha (mutated)	Human	NMR		
1ESL	EGF (E-selectin)	Human	X-ray	2·0	
FN1 – fibronectin type I repeat					
1TPM	tPA (finger domain)	Human	NMR		
FN2 – fibronectin type II repeat					
1PDC	Sem. fluid prot. PDC109 (2)				
FN3 – fibronectin type III repeat					
1TTG	Fibronectin (10)	Human	NMR		
1CFB	Neuroglian (1, 2)	Fly	X-ray	2·0	Sugar, sulfate
1FNA	Fibronectin (10)	Human	X-ray	1·8	
1TTF	Fibronectin (10)	Human	NMR		

Table 1 (*cont.*)

Code	Protein (domain no)	Species	Meth	Å	Ligand
1TEN	Tenascin (3)	Human	X-ray	1·8	
3HHR	Growth hormone receptor	Human	X-ray	2·8	Growth hormone
FOLLI – Follistatin-like (Kazal type protease inhibitors)*					
2OVO	Ovomucoid (3)	Pheasant	X-ray	1·5	
1OVO	Ovomucoid (3)	Quail	X-ray	1·9	
1BUS	Protease inhibitor IIA	Bovine	NMR		
2BUS	Protease inhibitor IIA	Bull	NMR		
1CHO	Ovomucoid (3)	Bovine	X-ray	1·8	Chymotrypsin
3OVO	Ovomucoid (3, cleaved)	Quail	X-ray	1·5	
4OVO	Ovomucoid (3, cleaved)	Pheasant	X-ray	2·5	
3SGB	Ovomucoid (3)	Turkey	X-ray	1·8	Protease
1PPF	Ovomucoid (3)	Turkey	X-ray	1·8	Elastase Hnec
1CGJ	Trypsin pancr. secr. inhib.	Human	X-ray	2·3	Chymotrypsin
1CGJ	Trypsin pancr. secr. inhib.	Human	X-ray	2·3	Chymotrypsin
1PCE	Pec-60	Pig	NMR		
1HPT	Trypsin pancr. secr. inhib.	Human	X-ray	2·3	
1TGS	Trypsin pancr. secr. inhib.	Pig	X-ray	1·8	Trypsinogen
1TUR	Ovomucoid (3)	Turkey	NMR		
1TUS	Ovomucoid (3)	Turkey	NMR		
GLA					
2PF2	Prothrombin fragment 1	Bovine	X-ray	2·2	Ca
2PF1	Prothrombin fragment 1	Bovine	X-ray	2·2	Ca
2SPT	Prothrombin fragment 1	Bovine	X-ray	2·5	Ca
1CFH	Factor IX	Human	NMR		
Kringle					
2PF2	Prothrombin fragment 1	Bovine	X-ray	2·2	
2PF1	Prothrombin fragment 1	Bovine	X-ray	2·2	
2SPT	Prothrombin fragment 1	Bovine	X-ray	2·5	
1PK4	Plasminogen kringle 4	Human	X-ray	1·9	
1PML	t-PA kringle 2	Human	X-ray	2·4	
2PK4	Plasminogen kringle 4	Human	X-ray	2·3	ε-NH_3-cap.ac.
1KDU	Urokinase	Human	NMR		
1PK2	t-PA kringle 2	Human	NMR		6-NH_3-hexanoic ac.
1PKR	Plasminogen kringle 1	Human	X-ray	2·5	
1TPK	t-PA kringle 2	Human	X-ray	2·4	
1URK	Urokinase	Human	NMR		
2HPQ	Thrombin	Human	X-ray	3·3	Keton
2HPP	Thrombin	Human	X-ray	3·3	Keton
KUNIT – Kunitz type protease inhibitors					
1SHP	Trypsin inhibitor	Sea anem.	NMR		
2PTC	BPTI	Bovine	X-ray	1·9	Trypsin
4PTI	BPTI	Bovine	X-ray	1·9	
5PTI	BPTI	Bovine	X-ray	1·0	
6PTI	BPTI	Bovine	X-ray	1·7	
7PTI	BPTI (mutated)	Bovine	X-ray	1·6	
8PTI	BPTI (mutated)	Bovine	X-ray	1·8	

Table 1 (*cont.*)

Code	Protein (domain no)	Species	Meth	Å	Ligand
9PTI	BPTI	Bovine	X-ray	1·2	
2TGP	BPTI	Bovine	X-ray	1·9	Trypsin
1TPA	BPTI	Bovine	X-ray	1·9	Trypsin
2TPI	BPTI	Bovine	X-ray	2·1	Trypsin, IL
3TPI	BPTI	Bovine	X-ray	1·9	Trypsin, IL
4TPI	BPTI (mutated)	Bovine	X-ray	2·2	Trypsin, VV
2KAI	BPTI	Bovine	X-ray	2·5	Kallekrein
1BPT	BPTI (mutated)	Bovine	X-ray	2·0	
¥1BTI	BPTI (mutated)	Bovine	X-ray	2·2	
1AAP	Alzheimer amyloid precursor	Human	X-ray	1·5	
1KNT	collagen type VI	Human	X-ray	1·6	
1FAN	BPTI (mutated)	Bovine	X-ray	2·0	
1AAL	BPTI (mutated)	Bovine	X-ray	1·6	
1NAG	BPTI (mutated)	Bovine	X-ray	1·9	
1BPI	BPTI form II	Bovine	X-ray	1·1	
1TIE	Erythrina Trypsin Inhibitor	Erythrina	X-ray	2·5	
1DTK	dendrotoxin K	Mamba	NMR		
1DTX	Alpha dendrotoxin	Mamba	X-ray	2·2	SO_4
LY6UP – Ly6/CD59 family					
1ERG	CD59	Human	NMR		
1ERH	CD59	Human	NMR		
1CDQ	CD59	Human	NMR		
PDOM – P-domain (trefoil)					
1PSP	spasmolytic polypeptide	Pig	X-ray	2·5	
1PCP	spasmolytic polypeptide	Pig	NMR		
1POS	spasmolytic polypeptide	Pig	X-ray	2·6	
TNFR – TNF/NGF receptors					
1TNR	TNF receptor p55	Human	X-ray	2·8	TNF

* Folli has not been modelled due to differences in the N-terminus. Nevertheless, the folds are probably very similar.

of the N-and C-termini. Clearly the position of the N- and C-termini has implications for the way modules are linked together.

IG/IGSF/Immunoglobulin 'superfamily'

Out of the about 60 modules characterised so far, the most frequently occurring ones seem to belong to the immunoglobulin type. About 40% of the surface proteins of leukocytes have been estimated to belong to the Ig family, based on the presence of a consensus sequence for immunoglobulins (Barclay *et al.* 1993). It is becoming clear that several other modules with only weak or no detectable sequence similarity to Igs, have the same, or at least a very similar β-sandwich topology; this percentage might thus be considerably higher. Sequence classification of proteins with sequence similarity to immunoglobulins (Williams

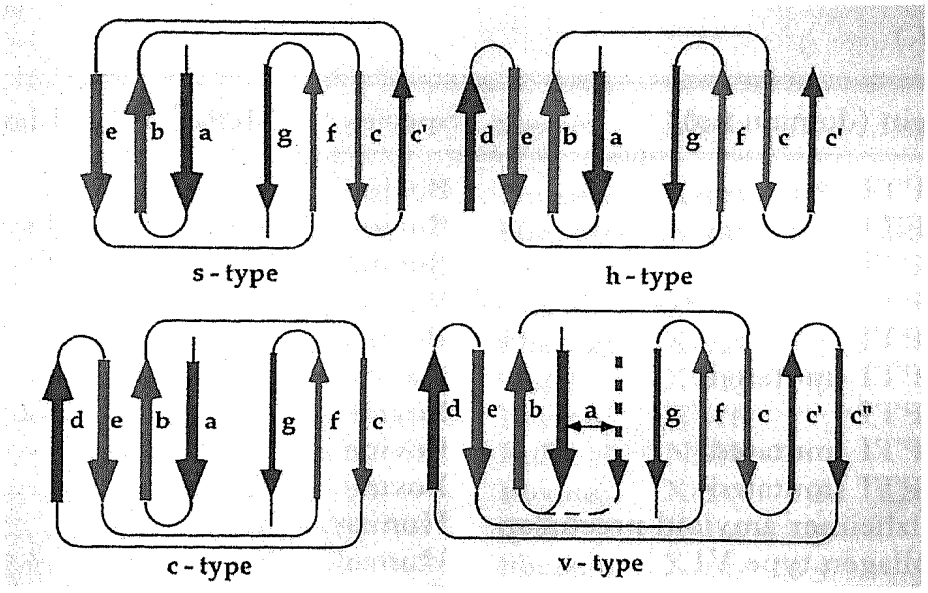

Fig. 4. Topology diagram of hydrogen bonding patterns in sandwich-like Ig-folds (reproduced from Bork *et al.* 1994). The 7–9 β strands (a, b, c, c', c'', d, e, f, g) form a sandwich of two β-sheets (front sheet, thick, back sheet, thin). The loop lengths are not drawn to scale. The common core observed in all Ig domains when superimposed (Bork et al. 1994) is shown in red. Ig constant domains (bottom left, c-type for constant) have seven strands, whereas the Ig variable domains (bottom right, v-type for variable) have an additional hairpin (c'–c'') between strand c and d, with a total of, usually, 9 strands.

& Barclay, 1988; Harpaz & Chothia, 1994) as well as structural comparisons of domains with Ig-like topology (Bork *et al.* 1994; Jones, 1993) have been extensively reviewed so the Igs will not be discussed here in much detail.

Domains that are structurally related to Igs are not confined to extracellular mammalian proteins; they seem to used in a variety of different contexts. Prokaryotic chaperonins (PapD), intra- and extracellular enzymes acting on carbohydrates (e.g. β-galactosidase, galactose oxidase, cyclodextrin glycosyltransferase and thermostable cellulase celD), haem-binding cytochrome f (Martinez *et al.* 1994) and Ca-binding synaptogamins (Sutton *et al.* 1995; L. Holm, personal communication) contain variants of the fold which contains between 7 and 9 β-strands arranged in two sheets (Bork *et al.* 1994). There is, however, considerable diversity in Ig-like folds that range from β-sandwiches (Fig. 4) to barrel-like structures; the hydrogen bonding pattern of the β-strands also varies considerably. For example, strand a has two alternative locations in v-type domains, being part of either the front sheet (parallel pairing with strand g) or the back sheet (antiparallel pairing with strand b). A similar behaviour for strand g, in yet to be solved structures, might be expected. Other Ig-like domains with 7 strands are distinguished because the 4th strand appears to have switched sheets (s-type for switched); the name of the 4th strand changes from d to c' to reflect the sheet switch. The last type represents an 8-stranded hybrid between c- and s-type that has a short c' that continues after a kink as d, so that both sheets have 4 strands (h-type, for hybrid). For many Ig-like domains it is not clear whether they have a common ancestor or whether they have evolved independently towards a stable fold. There are also technical difficulties related to 'thresholds'

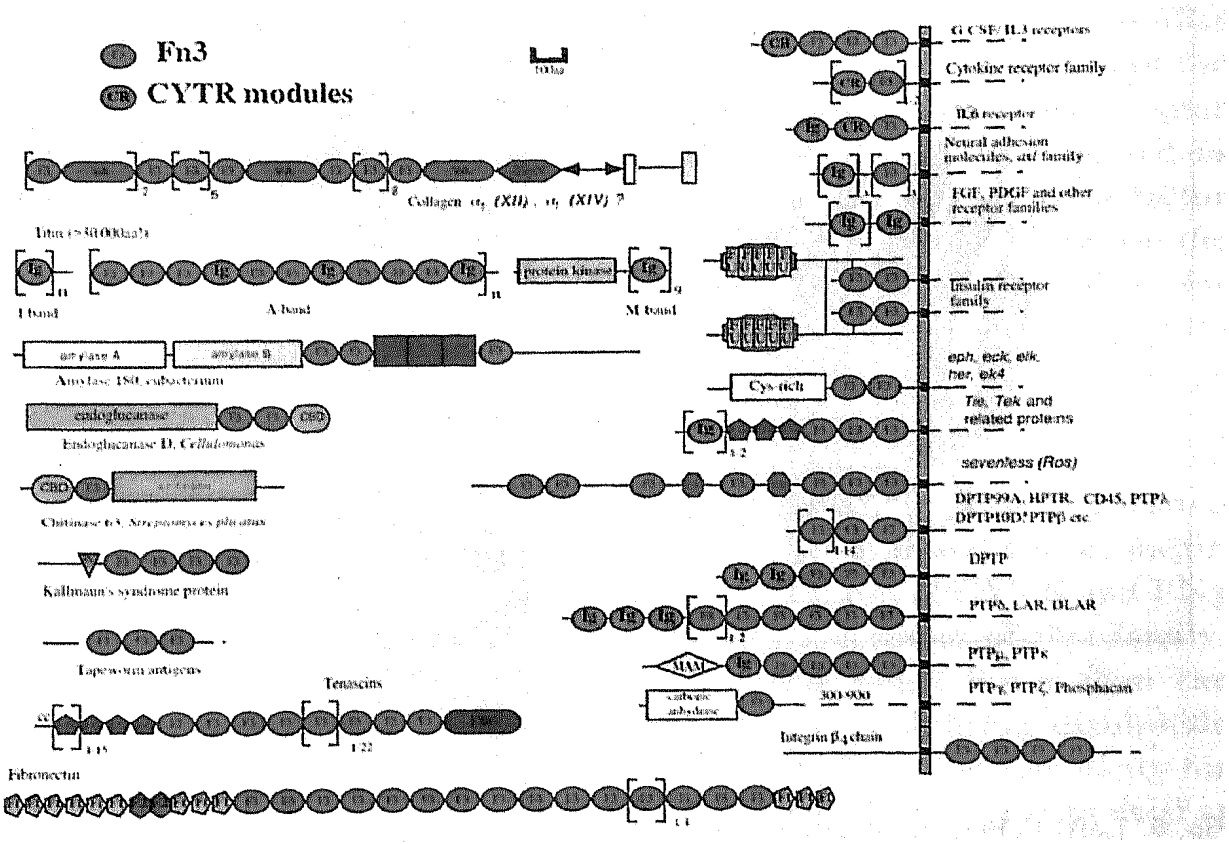

Fig. 5. Selected proteins that contain FN3 and CYTR modules. Note that additional mosaic proteins containing FN3s can be found in other figures of this review. For nomenclature of modules see Fig. 1.

in the distinction of modules with similar three-dimensional structure but rather limited sequence similarity. To illustrate this point, we discuss below several other extracellular modules with distinct sequence features that also have the β-sandwich topology.

F3/FN3/Fibronectin type-III

The FN3 domain is another extremely widespread module (Fig. 5) that was originally identified as the third type of repeat in the matrix protein fibronectin. Database searches have led to estimates that at least 2% of all animal proteins contain the fibronectin type III (FN3) unit (Bork & Doolittle, 1992); the continuing reports of proteins containing this module appear to support these estimates. The module topology, shown schematically in Fig. 6, left hand side differs mainly in a strand-switch of a single β-strand from classical Ig-like folds, although it is very similar to some Ig folds identified by sequence homology such as the second domain of CD2 (Bork *et al.* 1994). The two β-sheets have a right-handed twist, and they stack on top of each other. The domain does not contain any conserved disulphide bridges, hence a high number of consensus hydrophobic and aromatic residues are required to stabilise the packing of the two sheets (Fig. 6, right hand side). One of these residues is a highly conserved leucine which tethers the long loop connecting strands 5 and 6 to the core. The domain usually

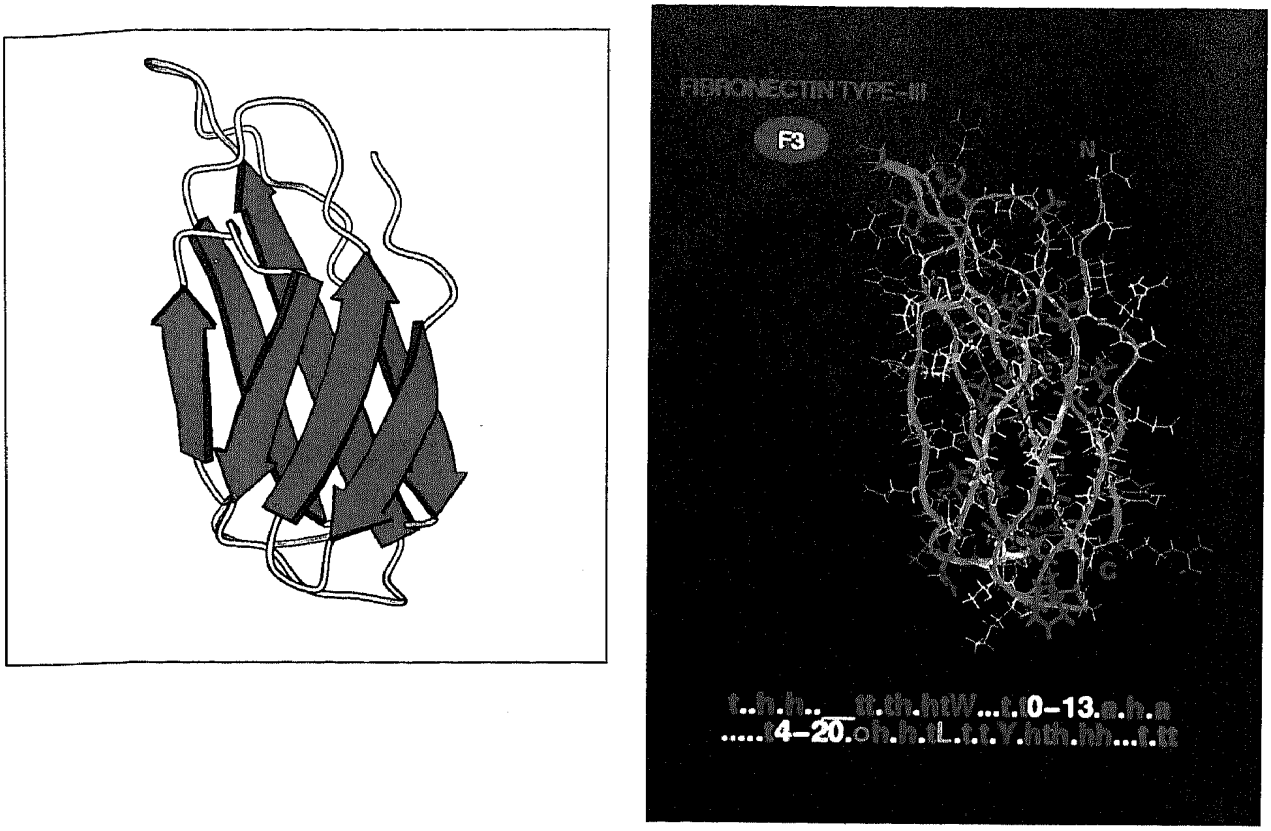

Fig. 6. Left) Schematic illustration of the fibronectin type-III domain fold. The structure selected as representative for this family is the tenth FN3 module from human fibronectin (Main *et al.* 1992). Right) Atomic resolution illustration of the fibronectin type-III fold highlighting conserved residues.

does not contain fold-stabilising disulphide bridges. The well-defined consensus enables pattern and profile searches to identify the numerous occurrences in proteins at the sequence level.

Proteins containing FN3s are mainly extracellular matrix proteins, although the majority of domains (measured in total amounts) are found in a few giant intracellular muscle molecules such as titin and twitchin. Titin, the largest protein sequenced so far, with more than 30000 amino acids, is almost exclusively composed of Igs and FN3 modules; they have a length of about 100 and 90 residues, respectively (Labeit and Kolmerer, 1995). At least one more protein seems to contain intracellular FN3s: integrin $\beta 4$. As FN3 does not require disulphide bridges to stabilise the fold, incorporation into intracellular proteins should not be difficult and additional occurrences in intracellular proteins can be expected.

While most of the extracellular FN3 modules have been found only in animals, FN3s have, surprisingly, also been identified in numerous extracellular glycohydrolases from soil bacteria (Bork & Doolittle, 1992). Quantitative sequence analysis suggests a rather recent horizontal gene transfer from the animals to the prokaryotes (Bork & Doolittle, 1992; Little *et al.* 1994).

The widespread occurrence of FN3 is accompanied by a variety of different functions. It often coexists with other modules in numerous receptors proximal to the membrane. This location, supported by some known 3D structures, e.g. the receptors for growth hormone and prolactin (de Vos *et al.* 1992; Somers *et al.*

1994), suggests that it might often play a role in dimerisation i.e. binding to similar FN3 modules. This model might have to be adjusted, however, in the light of the recent structure of interferon-γ (Walter *et al.* 1995) complexed with two receptor chains. Although difficult to detect at the sequence level and often treated as separate modules, the interferon-γ receptor, like the sequence-related tissue factor (Harlos *et al.* 1994; Muller *et al.* 1994), contains FN3-like domains. More specific binding features have been reported for FN3s in several matrix proteins; for example, the 10th FN3 repeat in fibronectin binds to cell surfaces via its RGD motif; neighbouring FN3s also seem to be involved in heparin binding (Potts & Campbell, 1994).

Structurally related to and sometimes grouped together with FN3s are the CYTR domains which usually precede FN3 domains within cytokine and related receptors, including growth hormone receptor (de Vos et al. 1992) and prolactin receptor (Somers *et al.* 1994). Sometimes tandem duplications of CYTR and FN3 domains or several FN3 units can be observed within receptors of this family. CYTR domain sequences appear to change or evolve much faster than the associated FN3 domains. This might be due to the 2–3 stabilising disulphide bridges, or perhaps it is a reflection of the development of binding specificity for respective substrates. {Receptors containing CYTRs and FN3s appear to exist at least as dimers and several CYTRs as well as FN3s seem to be involved in the process of ligand binding.}

CA/CADHE/Cadherin

Until recently, cadherin domains had only been found in repeats of four or five copies within the cadherin multi-gene family. With the identification of the *fat* tumour suppresser from *Drosophila*, containing as many as 34 CADHE copies as well as EGF and LamG modules, the widespread occurrence of CADHE domains has became clear (Fig. 7). At present, up to 8 copies of CADHE in cadherins have been reported; the receptor tyrosine kinase *ret* and related proteins contain a single CADHE module in their extracellular part (Fig. 7). The expanding family and the presence of CADHE in different functional contexts suggests that it acts as a structural scaffold for distinct functions, although the Ca-binding features seem to be conserved.

The cadherin domain is approximately the same size as the FN3 module, and it is also topologically similar to the immunoglobulin fold (Overduin *et al.* 1995; Shapiro *et al.* 1995). Unlike other Ig-like structures, however, its β-sheets are hydrogen bonded to form a cylindrical β-barrel rather than a sandwich of two β-sheets (Fig. 8, left hand side). Also, the cadherin domain contains two short helical regions, one of which is involved in calcium binding. Residues associated with calcium binding have been identified by NMR studies that monitor chemical shift changes upon addition of calcium (Overduin *et al.* 1995). Not all of these are conserved in the consensus sequence shown in Fig. 8, right hand side. Those which are potential ligands include the conserved D.E and D.N- regions which map, in 3-dimensional space, to the C-terminal region of the domain. The other

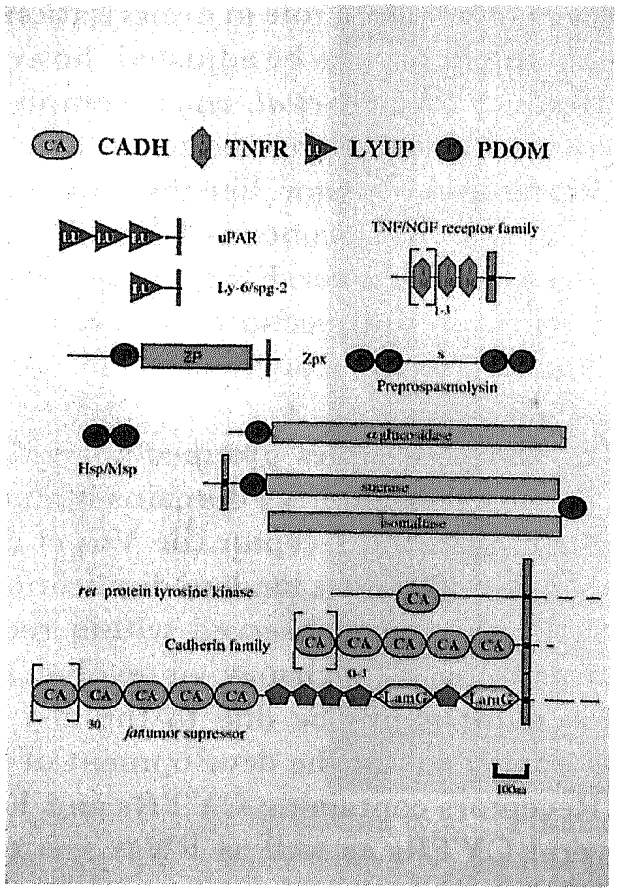

Fig. 7. Some proteins that contain CADHE, TNFR, LYUP and PDOM modules.

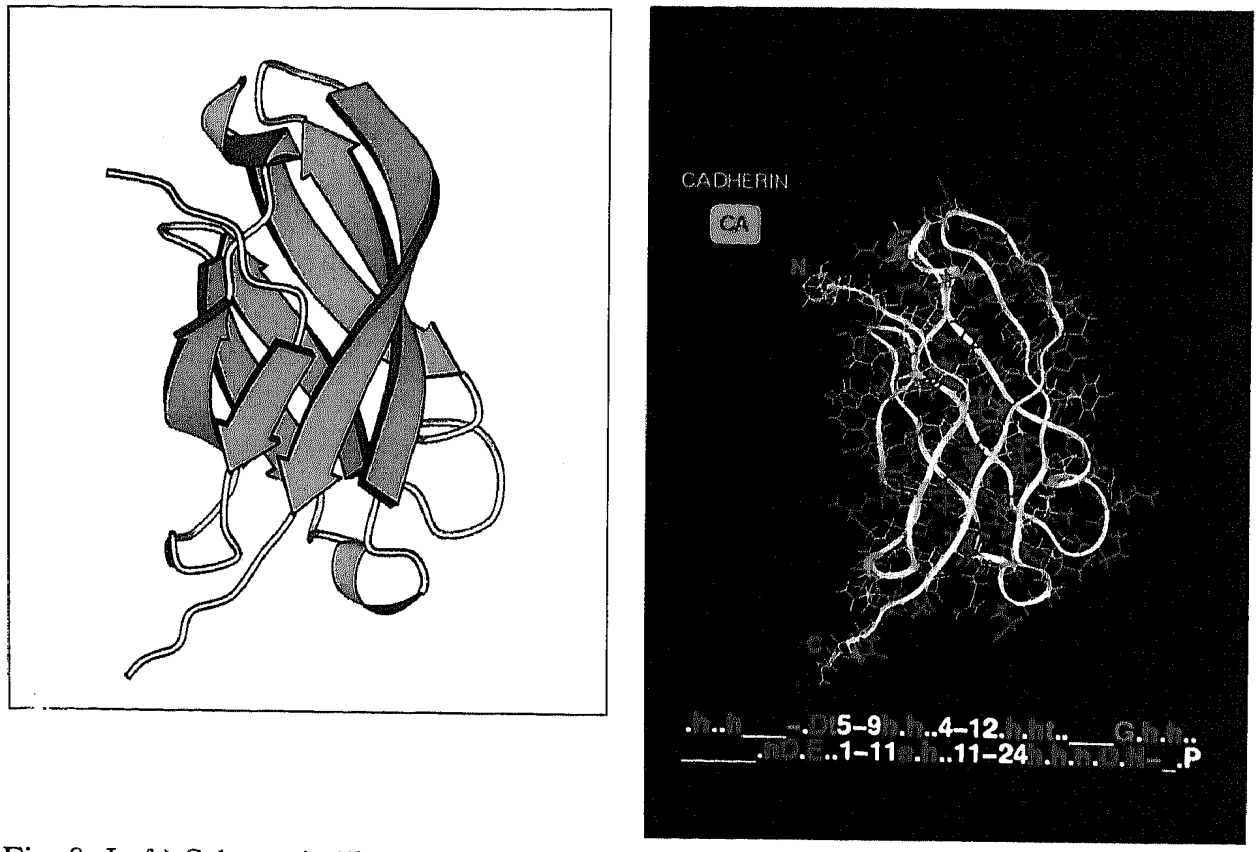

Fig. 8. Left) Schematic illustration of the cadherin domain fold. The structure selected as representative for this family is the epithelial cadherin domain responsible for selective cell-adhesion (Overduin *et al.* 1995). Right) Atomic resolution illustration of the cadherin domain fold highlighting conserved residues.

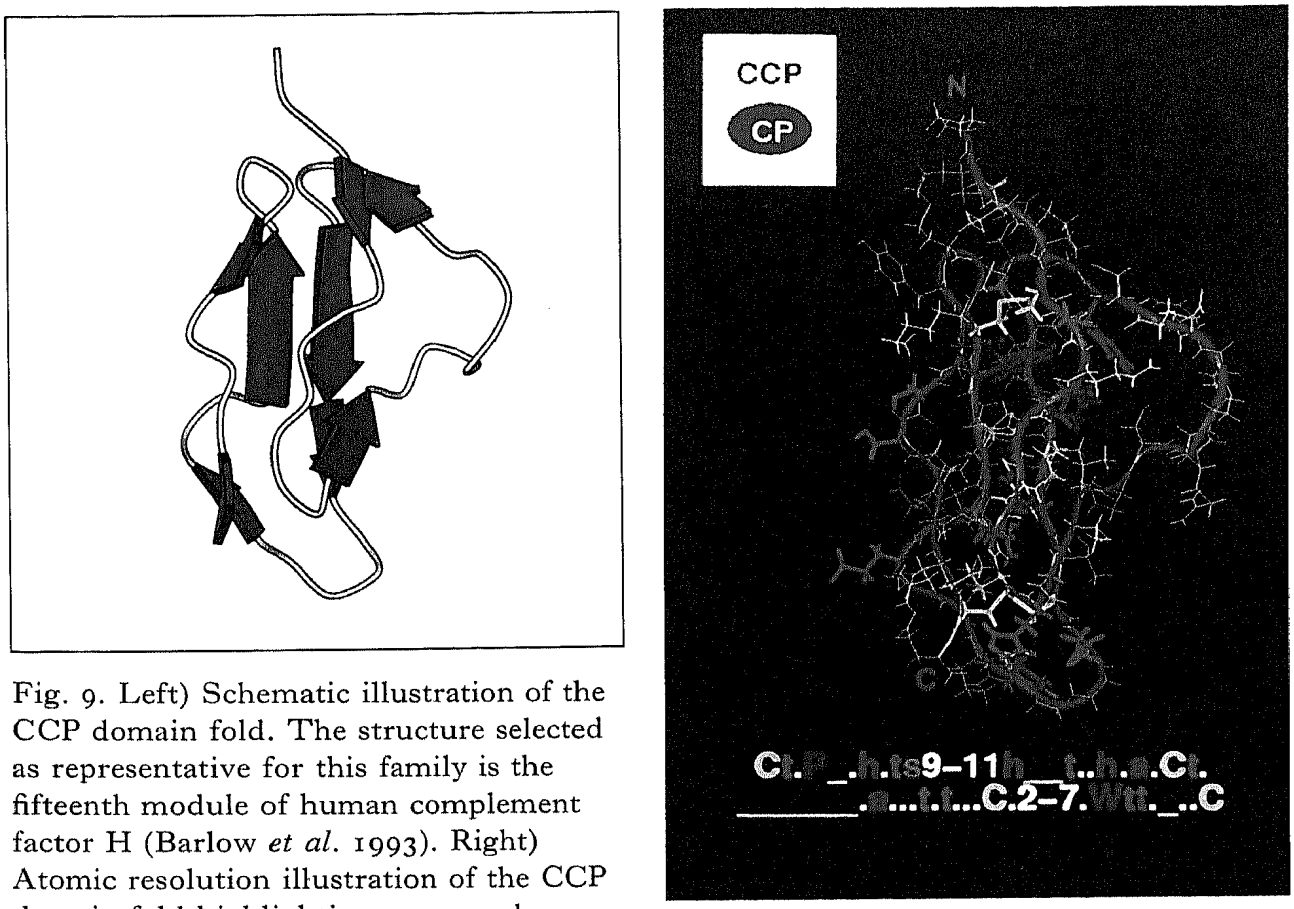

Fig. 9. Left) Schematic illustration of the CCP domain fold. The structure selected as representative for this family is the fifteenth module of human complement factor H (Barlow *et al.* 1993). Right) Atomic resolution illustration of the CCP domain fold highlighting conserved residues.

conserved carboxylate residues near the N-terminus appear to play a role in orienting the first helix with respect to the following β-strand by side chain-backbone hydrogen bonding interactions.

CP/CCP/CCP, Sushi, SCR

Like Ig, FN3, CYTR and CADHE, the CCP module has a β-sandwich structure with N- and C-termini at the opposite ends of one long edge of an ellipsoid shape. Fig. 9, left hand side illustrates the secondary structure of this module which is comprised of several short segments of β-strand held together by two disulphide bridges in a 1–3, 2–4 pattern.

This module is abundant in complement control proteins and also in the complement system itself (Fig. 3). It has also been found in viruses, the enzyme thyroxide peroxidase and in protein families such as selectins and mucins. Its function is still unclear and may well be different in the various proteins where it has been found so far. CCPs occur most frequently in tandem arrays but also a few single copies have been identified. An example is the complement receptor CR2, that has 16 CCP modules in its extra-membrane region. It is the receptor for proteins of the complement system and the Epstein Barr virus. It has been shown that the N-terminal two CCP modules are necessary and sufficient for binding both the virus and the natural ligand (Lowell *et al.* 1989).

The consensus sequence (Fig. 9, right hand side) includes a number of conserved hydrophobic and aromatic residues, most of which lie buried in the core of the module. One exception is the aromatic residue between the second and third cysteines which maps to the structure at the C-terminal end of the molecule. Interestingly this side chain ring appears to pack against the turn which is conserved between the first two β-strands, thereby stabilising its conformation.

A conserved Pro near the N-terminus terminates the first β-strand, directing the backbone chain towards the central core of the molecule. This change in direction, elicited by the proline leaves the 'hypervariable' loop extending in solution on the right-hand side of the molecule. This loop, which comprises the most disordered region of the solution structure of the module (Barlow *et al.* 1993), might be important for the protein-protein interactions.

EG/EGF/EGF-like

Like Ig and FN3 modules, EGF seems to be extremely widespread (Fig. 10). Tandem arrays of more than 70 EGFs have been identified in *C. elegans* proteins and the total number of distinct occurrences in databases probably already exceeds 600 (Campbell & Bork, 1993). All EGF-like domains contain a major double-stranded β-sheet, and some, such as the one shown in Fig. 11, top left, also include additional short segments of β-sheet or α-helix. The EGF-like domain may be described as relatively small and ellipsoid with its N- and C-termini located at opposite ends of its major axis. Unusually, no conserved hydrophobic residues are required to maintain the EGF-like domain fold; it is stabilised primarily by three disulphide bonds which form in a 1–3, 2–4, 5–6 pattern (Fig. 11, bottom left). While there are few residues which are conserved for the entire family of EGF-like domains, it is noted that amongst sub-groups, conservation patterns are detectable (Campbell & Bork, 1993).

EGFs are frequently found in both receptors and in matrix proteins (Fig. 10) and have distinct functions in different settings. Some EGFs bind specifically to receptors using large parts of their surface, others mediate interactions via Ca-binding. Recently an X-ray structure of the N-terminal calcium loaded EGF-like domain from factor IX has been reported (Rao *et al.* 1995) which has provided a very detailed picture of the coordination of the calcium atom as shown in Fig. 11, top right. Comparison of the structure of the calcium binding EGF-like domain in Fig. 11, bottom right to the non-calcium binding domain in Fig. 11, bottom left shows that the two folds are very similar. It has been suggested that Ca-binding may be used for formation of rod-like superstructures involving tandem arrays of EGFs (Rao *et al.* 1995).

EGFs coexist with certain domain types in quite different settings. The shuffling of blocks containing several modules is one possible explanation as is suggested, for example, by the arrangement of YWTD modules and EGF modules in various receptors similar to EGF precursor and LDL receptor, as well as in nidogen-like proteins. More puzzling is the frequent coexistence of EGF and the LAMB module e.g. in agrin, crumbs, the neurexins, protein S, perlecan and

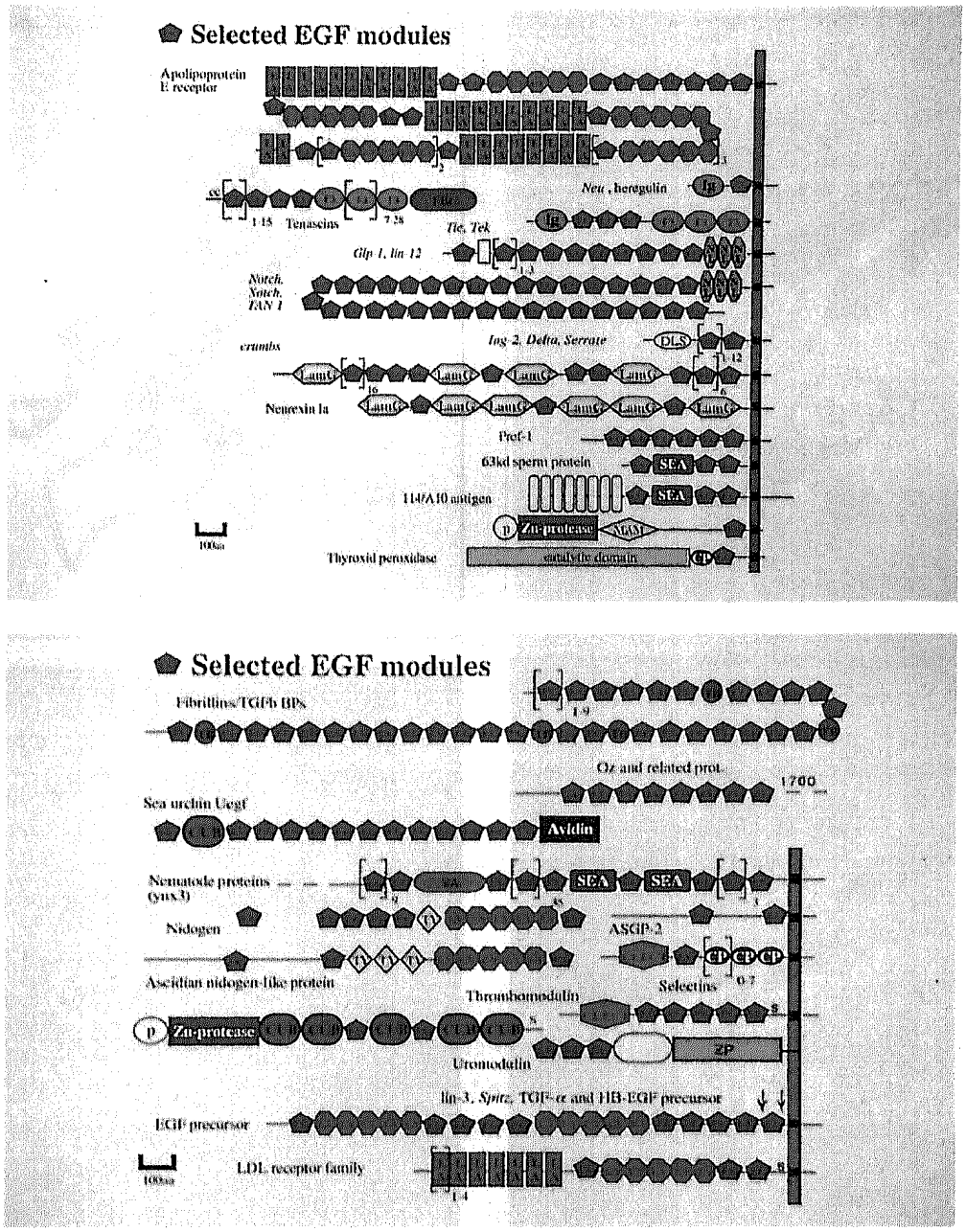

Fig. 10. Selected proteins that contain EGF modules. Note that additional mosaic proteins containing EGFs can be found in other figures of this review.

the fat tumour suppresser. EGFs are also found attached to enzymes, not only in blood coagulation and complement but also in thyroid peroxidase and prostaglandin synthase, the latter being located in the endoplasmic reticulum. Various viruses encode EGF-containing proteins and they might even exist in several parasitic single cell eukaryotes.

F1/FN1/Fibronectin type-I

A rather rare module is the fibronectin type 1 repeat (FN1), first identified in the matrix protein fibronectin but later also found in tPA and in coagulation factor XII. Hepatocyte growth factor activator (HGFA) with a modular architecture similar to factor XII also contains one copy of this module (Fig. 2). FN1 modules

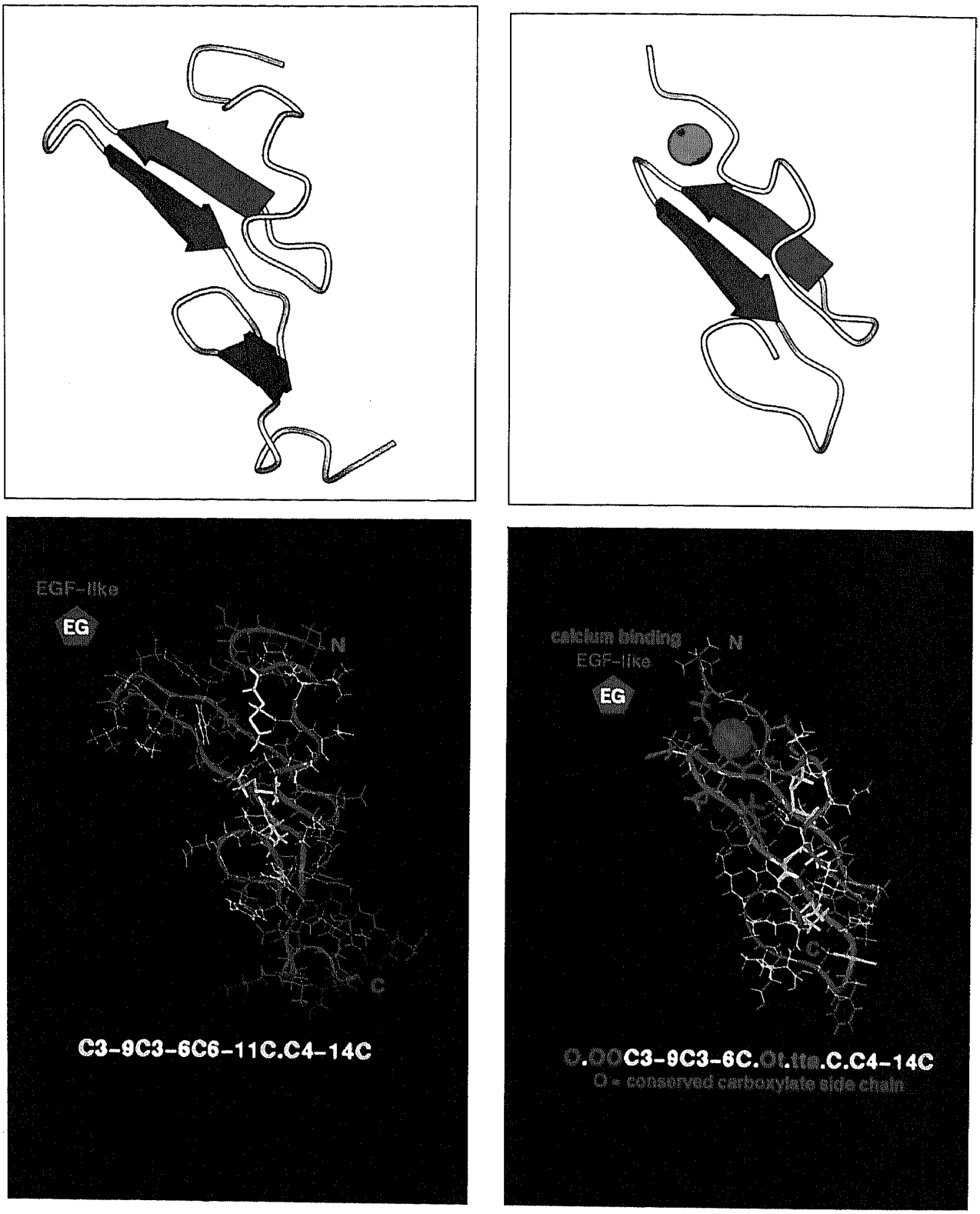

Fig. 11. Top left) Schematic illustration of the EGF-like domain fold. The structure selected as representative for this family is that of the human epidermal growth factor peptide hormone (Hommel *et al.* 1992). Bottom left) Atomic resolution illustration of the EGF-like domain fold highlighting conserved residues. Top right) Schematic illustration of the calcium binding EGF-like domain fold. The structure shown is the N-terminal calcium loaded epidermal growth factor-like domain from factor IX (Rao *et al.* 1995). Bottom right) Atomic resolution illustration of the calcium loaded EGF-like domain highlighting conserved residues. Conserved carboxylate residues (indicated by an O in the consensus sequence) all participate in defining the geometry of the calcium binding site.

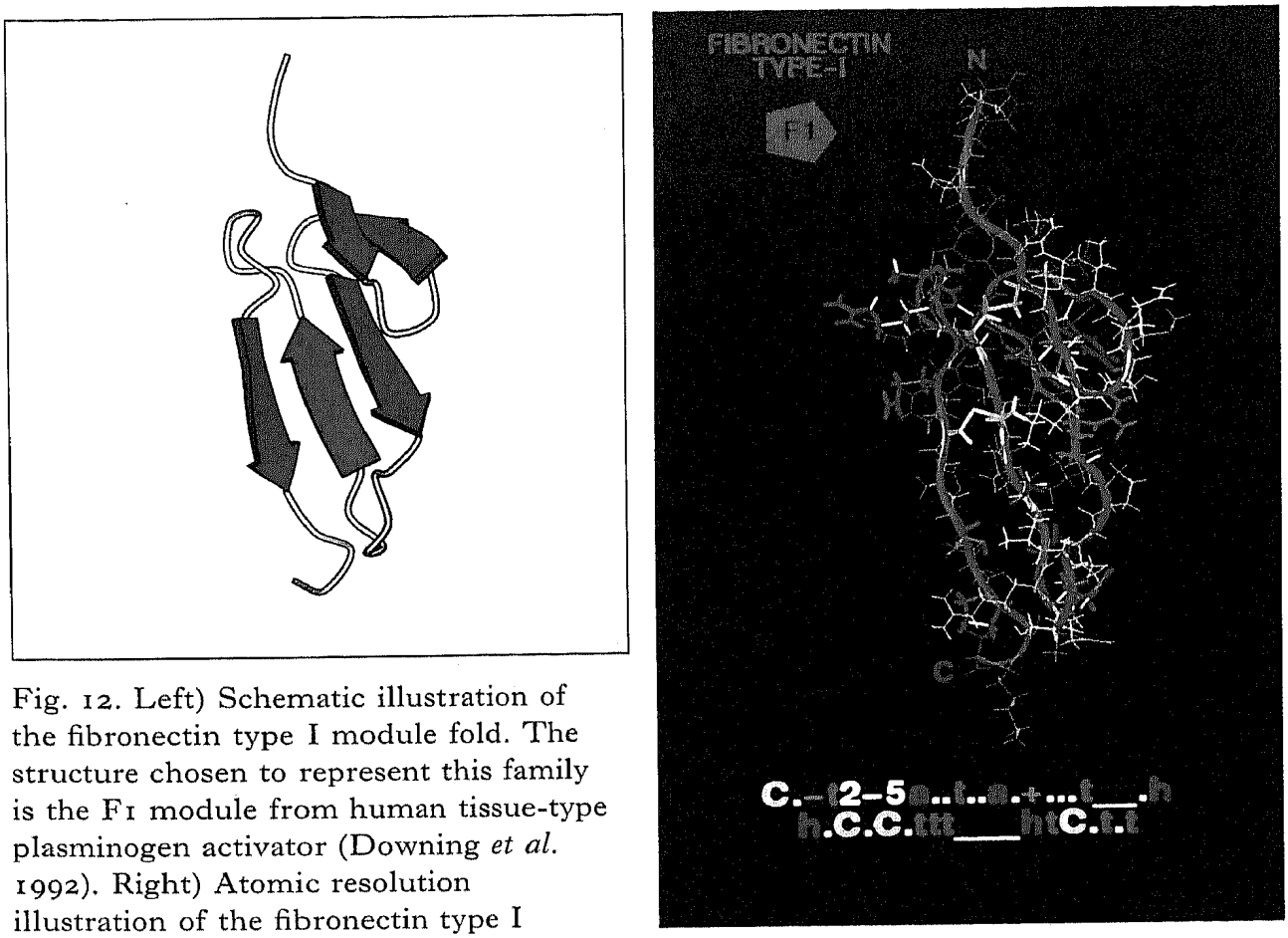

Fig. 12. Left) Schematic illustration of the fibronectin type I module fold. The structure chosen to represent this family is the F1 module from human tissue-type plasminogen activator (Downing *et al.* 1992). Right) Atomic resolution illustration of the fibronectin type I module highlighting conserved residues.

are found in the fibrin and collagen binding region of fibronectin and they also seem to be involved in fibrin binding in other proteins such as tissue plasminogen activator (Potts and Campbell, 1994; Smith *et al.* 1995).

The topology of the F1 module is dominated by two regions of antiparallel β-sheet that are held together by two disulphide bridges that link conserved cysteines in a 1–3, 2–4 pattern (Fig. 12, left hand side). The double-stranded sheet is further anchored to the triple-stranded one by a salt bridge formed between the conserved +/− pair of amino acids (Fig. 12, right hand side). The majority of the conserved aromatic and hydrophobic residues are found in the domain core. One exception is noted at the back of the triple-stranded β-sheet. Interdomain contacts in F1-containing module pairs suggests that this exposed hydrophobic side chain is conserved for pairwise domain packing (Williams *et al.* 1994; Smith *et al.* 1995). For example, in the tissue-type plasminogen activator FN1-EGF pair this residue packs between the two domains and helps to define their relative orientation.

TR/TNFRC/TNF family receptors Cys-rich

The TNF/NGF receptor family is rapidly growing and includes receptor-like proteins that all seem to bind dimeric or multimeric cytokines; even several distinct ligands can be bound (reviewed in Mallet & Barclay, 1991; Bazan, 1993; Smith *et al.* 1994). The family includes (at least) two pox virus proteins which might block several defence mechanisms of the hosts. Typically, the proteins of

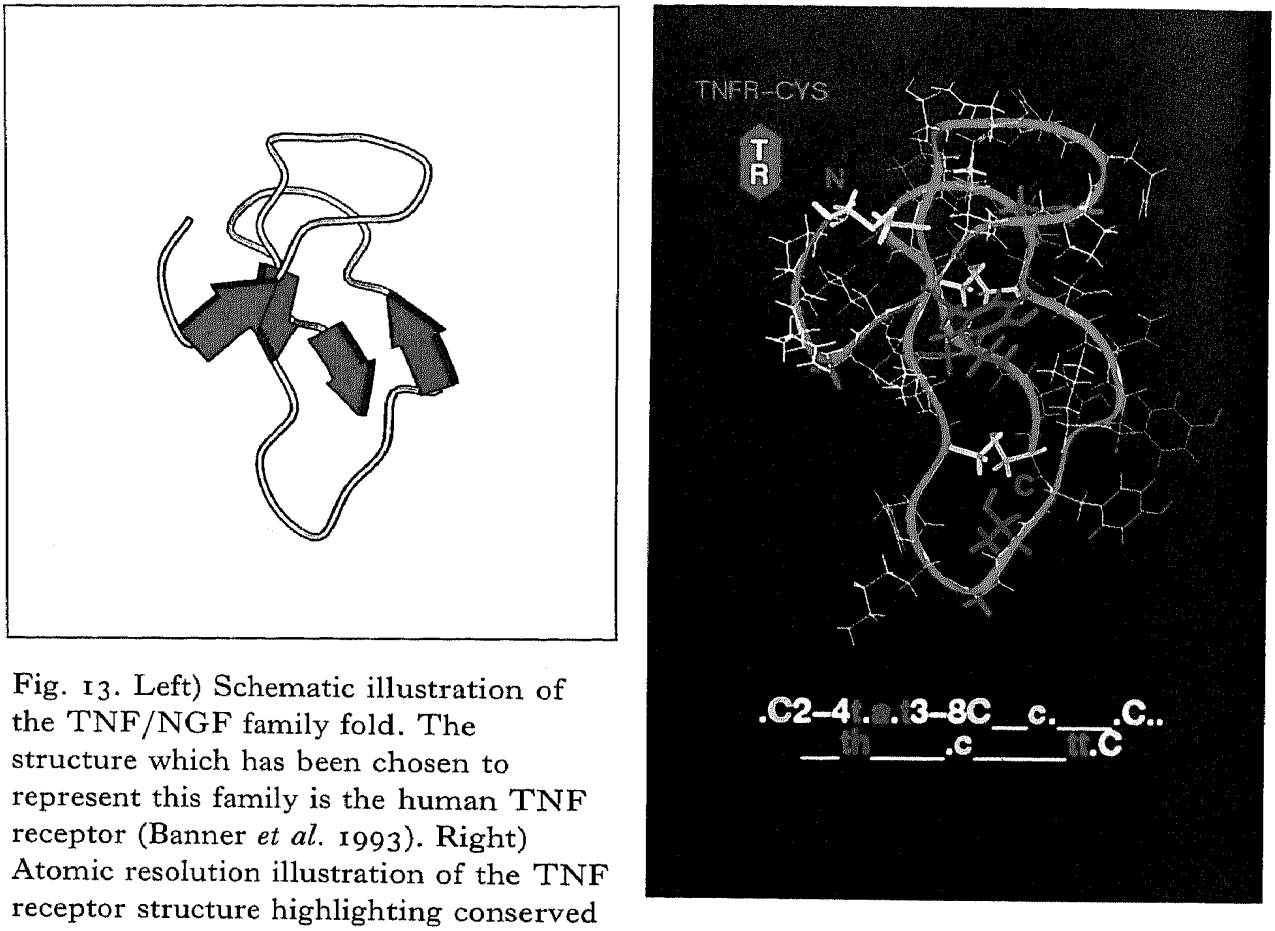

Fig. 13. Left) Schematic illustration of the TNF/NGF family fold. The structure which has been chosen to represent this family is the human TNF receptor (Banner *et al.* 1993). Right) Atomic resolution illustration of the TNF receptor structure highlighting conserved residues.

the TNF receptor family contain 3–5 copies of the TNFRC module, shown schematically in Fig. 7. Although TNFRC cannot be classified as a module in a strict sense as it has not yet been found to be associated with other modules, we have included it because of the diversity of the TNF/NGF family. Recently the presence of this module in the EGF and insulin receptors has also been proposed (Ward *et al.* 1995).

The secondary structure of the small, globular TNFRC module includes two double-stranded β-sheets (Fig. 13, left hand side). Its fold is constrained by three disulphide bonds connecting cysteines 1–2, 3–5 and 4–6, although the 3–5 bond does not form in all members of the TNFRC family (see below). These bridges tether the N-terminus, the two β-sheets and the C-terminus respectively, as highlighted in Fig. 13, right hand side. Few other residues besides the cysteines are conserved. The aromatic ring which extends from the first β-strand packs against the 3–5 disulphide bond as shown, and similarly the hydrophobic residue near the C-terminus packs against the 4–6 S-S bond.

KR/KRING/Kringle

The kringle domain was one of the first modules identified in the blood coagulation system (Fig. 2). The name stems from the Danish pastry that shows some similarities to the 2D representation of the sequence with its S-S bond constraints. Until recently, its spread was restricted to proteins of the coagulation

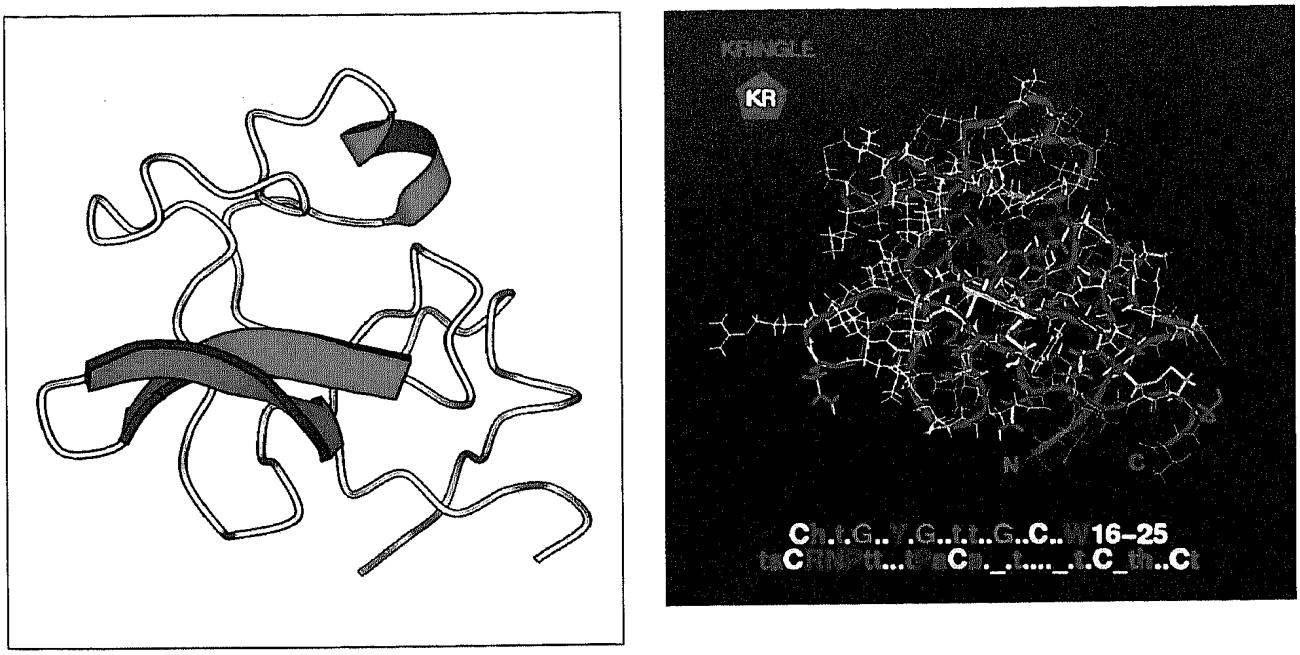

Fig. 14. Left) Schematic illustration of the kringle domain. The structure of the second kringle from tissue-type plasminogen activator (de Vos *et al.* 1992) has been selected to represent this family. Right) Atomic resolution illustration of the kringle domain highlighting the positions of conserved residues.

and fibrinolysis pathways. Much later, the discovery of apolipoprotein A, with as many as 37 kringle modules all very similar to those in plasminogen, has suggested functions outside the coagulation system. Hepatocyte growth factor and related proteins with modular architectures very similar to clotting/anticlotting proteases reveal similar cascades in other functional contexts (Fig. 2). The identification of kringles in *ror*- like and *trk*- like receptor tyrosine protein kinases revealed the presence of a completely new class of proteins in which kringles exist independent of serine proteases in one molecule (Fig. 2). The kringles within the coagulation system are well-studied and have apparently adapted different ligand-binding features that contribute to the fine-tuning of this complex cascade. For example, of the two kringle modules in tissue-type plasminogen activator, only one contains an active lysine binding site (van Zonneveld *et al.* 1986).

Unusually, the kringle domain contains a large number of β-turns and relatively little α or β secondary structure (Fig. 14, left hand side). Nonetheless, the module has a well-defined conformation due to the large number of conserved residues, all of which play important structural roles. In the consensus sequence shown in Fig. 14, right hand side, all five aromatic and the two hydrophobic side chains cluster and form a stable hydrophobic core. Proline residues that restrict backbone flexibility are especially important for maintaining turn conformations. The two conserved prolines perform this function and also bury their imino side chain rings in the core. Numerous conserved turn and glycine residues dot the consensus sequence, further favouring the kringle domain fold. In addition, each side chain of the conserved Arg-Asn pair near the centre of the sequence makes multiple hydrogen bonds to the protein backbone. Of course, the three disulphide bridges connecting cysteines 1–6, 2–4 and 3–5 covalently enhance the kringle

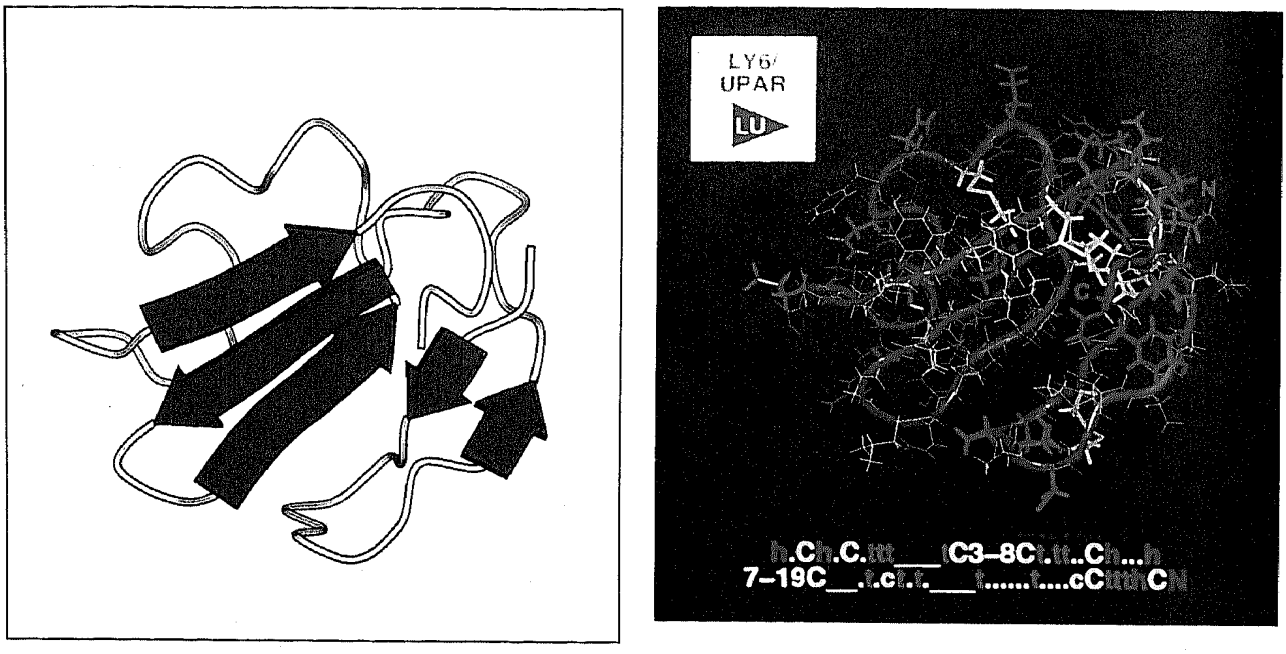

Fig. 15. Left) Schematic illustration of the Ly6 antigen/uPA receptor family fold. The structure of the extracellular region of the complement regulatory protein CD59 (Kieffer *et al.* 1994) has been chosen to represent this family. Right) Atomic resolution illustration of the Ly6/uPA module highlighting conserved residues.

domain structure; in particular the 1–6 S-S bond maintains the close proximity of the kringle N- and C-termini.

LU/LY6UP/Ly6 antigen/uPA receptor

This module has been found so far in members of the Ly-6/CD59/spg-2/ThB family as a single copy and in uPA receptor in a tandem array of three copies (Fig. 7). The precise functions of these proteins are unknown, but the different reported activities of CD59 such as complement inhibition and T-cell activation suggest distinct functions for the modules in the different settings. Structural similarity and some common sequence features between LY6UP and a group of snake venom neurotoxins exist (Fletcher *et al.* 1994; Kieffer *et al.* 1994). However, no common functional features have been predicted so far.

The LY6UP domain secondary structure is dominated by five antiparallel β-strands. The long turn on the left side of the molecule, as shown in Fig. 15, left hand side, is helix-like but lacks the regular backbone hydrogen bonds normally used to define this conformation. The pattern of disulphide bonds adopted by the 10 conserved cysteine residues was first elucidated by this structure of the extracellular region of the complement regulatory protein CD59 (Kieffer *et al.* 1994; Fletcher *et al.* 1994). Covalent S-S bond linkages are observed between cysteines 1–5, 2–3, 4–6, 7–8 and 9–10, although the 7–8 disulphide bond does not form in all Ly-6 homologues. The shape of the domain is approximately spherical, but flattened along the z-axis as shown in Fig. 15, right hand side. Most of the conserved hydrophobic residues cluster on the surface of the molecule near the N- and C-termini rather than in the core. This arrangement appears to stabilise the

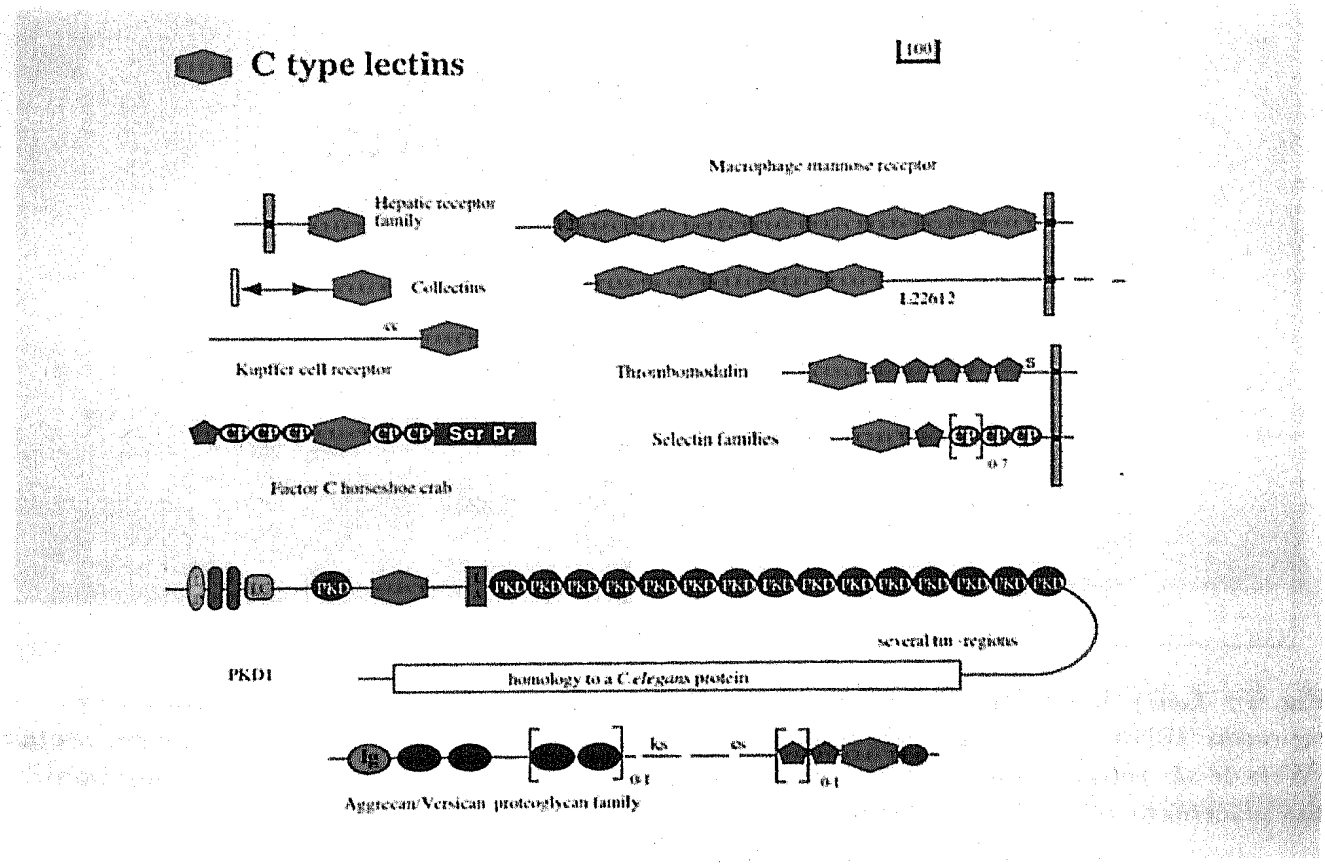

Fig. 16. Some proteins that contain CLECT modules. Note that C-type lectins are not homologous to other lectin types. For nomenclature of modules see Fig. 1.

relative orientation of the termini with a zipper-like mechanism. The C-terminal hydrophobic side chain is neatly sandwiched between the two N-terminal ones. Precise displacement of the termini may be important for attachment of glycophosphatidylinositol (GPI) which serves as an anchor to the membrane for this domain. The site of attachment has been speculated to be at the C-terminal Asn in Fig. 15, right hand side, based on the otherwise inexplicable amino-acid identity at this position.

CL/CLECT/C-type lectin

C-type lectins are one of the few modules for which an overall function has been proposed: carbohydrate-binding. Nevertheless, only a few of the numerous C-type lectins in modular proteins have been subjected to binding studies to confirm their sugar-binding specificities. As many extracellular proteins are glycosylated, no precise predictions for their specific binding functions can be made. A classification of the different C-type lectins has been proposed and many features have already been reviewed (Drickamer, 1992), but more divergent types of lectins also exist (Fig. 16), as became recently obvious by the identification of a divergent C-type lectin in the gene product of PKD1 (Glucksmannkuis *et al.* 1995; Fig. 16).

The C-type lectin domain, which is depicted in Fig. 17, left hand side, is roughly spherical in shape with an overall diameter of $\sim$ 30–35 Å. Its core contains a network of antiparallel β-strands some of which appear distorted,

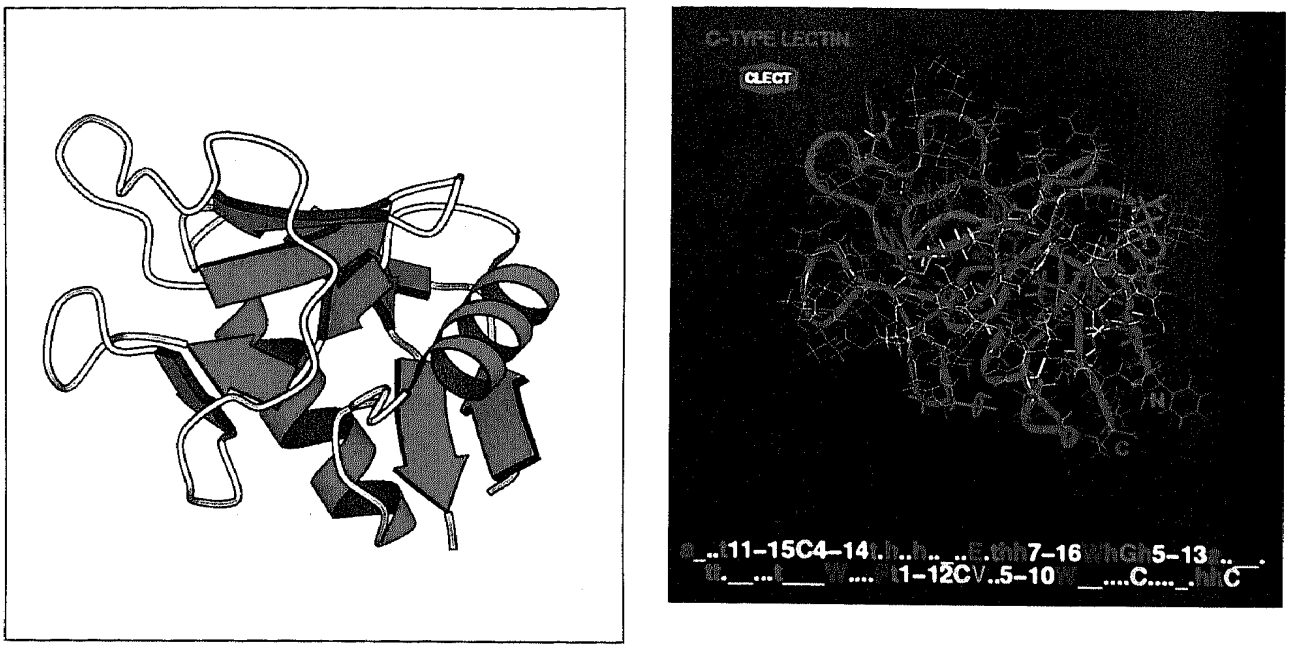

Fig. 17. Left) Schematic illustration of the C-type lectin fold. The structure chosen to represent this family is the carbohydrate recognition domain of rat mannose-binding protein (Weis *et al.* 1992). Right) Atomic resolution illustration of the CLECT domain highlighting the placement of conserved residues.

possibly due to packing interactions. Two amphipathic helices flank the β-secondary structure. All of the hydrophobic and aromatic residues which are highlighted in Fig. 17, right hand side lie in the core of the molecule. The conserved valine appears at the centre of the molecule as it is shown, and lies underneath a hydrophobic canopy created by three aromatic rings. Two conserved disulphide bonds also contribute to the definition of the fold. In particular, the 1–4 S-S bond fixes the orientation of the first helix and the C-terminus. Finally, a conserved Glu, shown at the right of the figure, stabilises the first turn of the second helix via (i, i+3) and (i, i−4) side chain to backbone hydrogen bonds.

GA/GLA/Gamma-carboxy-glutamate domain

The GLA domain is found at the N-terminus of several clotting factors (Fig. 2). The name was derived because of the presence of numerous γ-carboxyglutamate residues. As illustrated in Fig. 18, left hand side, its secondary structure is composed of three short α-helices. GLA is a module with a well-defined function: Ca-dependent membrane-binding. On the right-hand side of the molecule as shown in Fig. 18, right hand side, two internal carboxylate surfaces are covered with γ-carboxyglutamate side chains which can bind multiple calcium atoms. Details of the mechanism are still far from being completely understood (Valcarce *et al.* 1994), although some insight has been provided by structural comparison of the calcium loaded and calcium free forms of the GLA module (Sunnerhagen *et al.* 1995). The restriction of GLA to clotting factors might be due to its post-translational modifications that require specific recognition mechanisms.

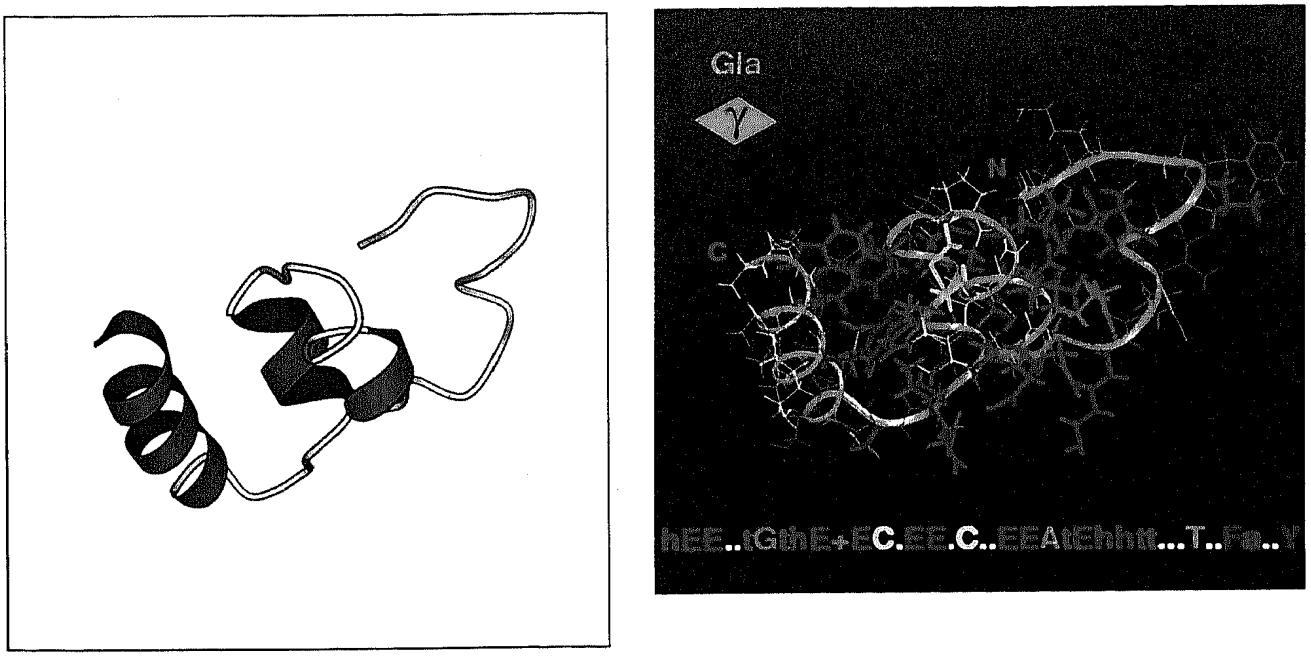

Fig. 18. Left) Schematic illustration of the gamma-carboxyglutamate domain. The GLA domain of Ca-prothrombin fragment 1 (Soriano-Garcia *et al.* 1992) has been chosen as representative of this family. Right) Atomic resolution illustration of the GLA domain highlighting conserved residues. In this figure γ-carboxyglutamate residues are denoted by 'E'

The main sources of stability for the GLA fold are conserved hydrophobic and aromatic side chains concentrated in the region between the three helices and a single disulphide bond. Two other conserved positions, one positively charged between GLA residues 3 and 4, and a Thr near the C-terminus, also help maintain the backbone conformation via side chain to carbonyl oxygen hydrogen bonds.

AT/ANATO/Anaphylatoxin

Anaphylatoxins are fragments of the central complement proteins C3, C4 and C5 which are released after proteolytic degradation during complement activation (Fig. 3). They seem to be involved in numerous biological activities including induction of smooth muscle contraction, histamine release from mast cells and basophilic leukocytes, increase of vascular permeability, antibody production and in chemotactic and chemokinetic processes. C5a is one of the major inflammatory proteins that has a range of activities (Hugli, 1981).

In spite of harbouring the anaphylatoxins and several other active degradation products, the related complement components C3, C4 and C5 were not originally considered as modular proteins because nearly their entire sequences are homologous to macroglobulins. However, their modular architecture began to be unravelled with the identification of the C-terminal domain in the netrin family and the detection of anaphylatoxin-like domains in the fibulin family (Fig. 3).

The structure of a complement anaphylatoxin module is a four α-helix bundle with down-up down-up topology when viewed in the orientation shown in Fig. 19, left hand side. However, fibulin anaphylatoxin modules only include approximately two-thirds of the anaphylatoxin homology region defined using

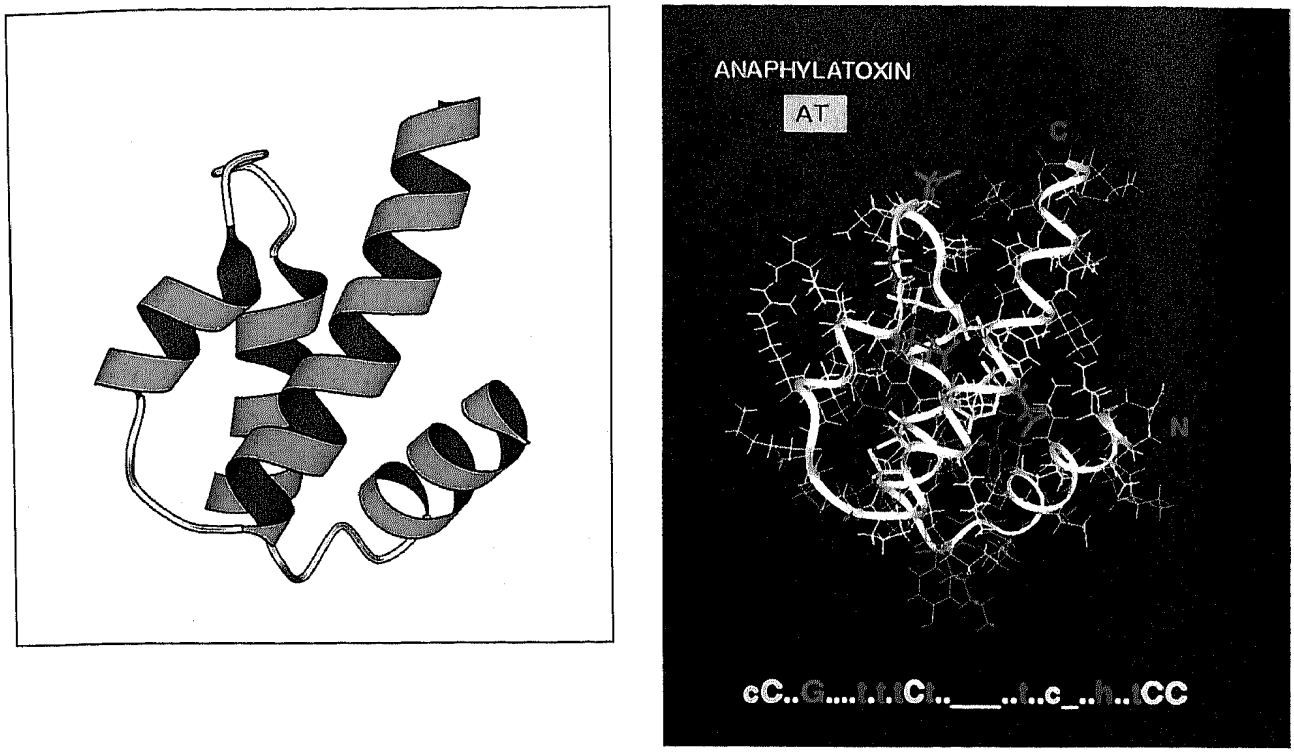

Fig. 19. Left) Schematic illustration of the anaphylatoxin fold. The structure of porcine C5a (Williamson & Madison, 1990) represents this family. Right) Atomic resolution illustration of the ANATO structure highlighting the positions of conserved residues.

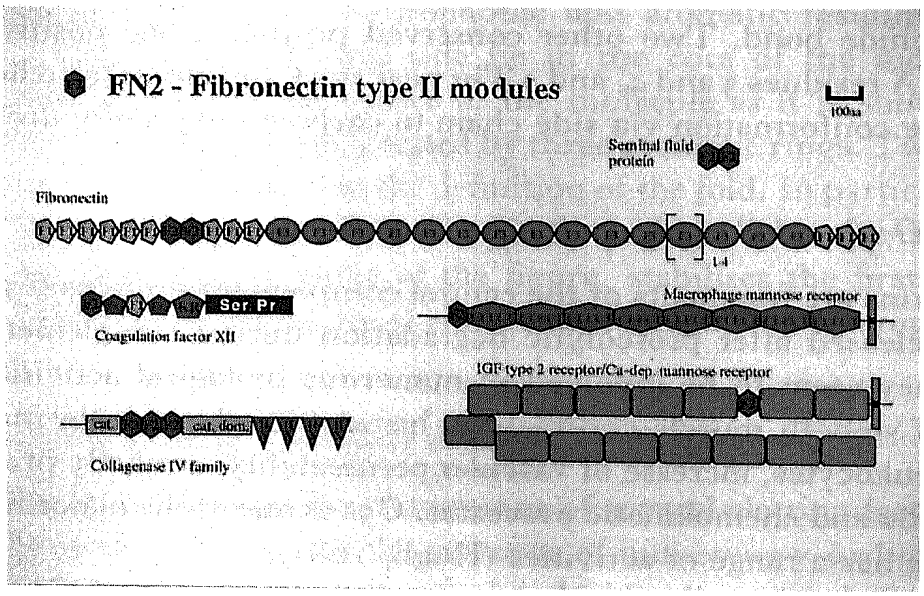

Fig. 20. Some proteins that contain FN2 modules; see also Fig. 2.

complement protein sequences. Specifically, if each helix shown is divided in two and labelled A'-A'', ... D'-D'', then fibulin anaphylatoxin repeats include the region defined from B'' to D'; the first of the three disulphide bonds which join cysteines 1–4, 2–5 and 3–6, does not form in all members of the family. Fibulins contain three copies of the domain arranged in a tandem array, followed by several copies of EGFs. Fibulin-2 also contains a preceding N-terminal domain. The function of ANATO in fibulins remains unclear.

Due to the great diversity of complement and fibulin anaphylatoxin sequences, most residues in the consensus sequence play structural roles (Fig. 19, right hand side). A hydrophobic core is not conserved; only one non polar residue consistently packs in the interior of the module. Interestingly, this particular residue is located on the interior face of the D' helix and its side chain packs against the B'' helix. The high degree of conservation of this residue suggests that the fibulin AT repeat might fold independently.

F2/FN2/Fibronectin type-II

Although the second repeat of fibronectin usually coexists with FN1, not only in fibronectin itself but also in factor XII and in HGFA, several independent occurrences of FN2s are known (Fig. 20). Two FN2s form seminal fluid and related proteins, one, or several, are inserted in various collagenases and improve ligand binding affinity. FN2 have also been found in mannose receptor and related proteins (Fig. 20).

The F2 domain is small and contains two short segments of double-stranded antiparallel β-sheet (Fig. 21, left hand side). Its structure is similar to the kringle domain in that it contains little secondary structure and consequently a high proportion of residues are conserved to stabilise its fold. Also like the kringle domain, disulphide bonds hold the N- and C-termini close together in a crossed conformation as shown in Fig. 21, right hand side. The core of the domain contains mainly of aromatic residues which form a partially exposed hydrophobic surface. NMR monitored binding studies of the domain suggests that these residues form a putative binding site for collagen (Constantine *et al.* 1992). The orientation of the side chain of the conserved threonine with its hydroxyl group on the surface probably enhances solubility of the domain.

PD/PDOM/P-type (Trefoil)

The P domain, illustrated in Fig. 22, left hand side, is small like the GLA and FN2 domains. It is stabilised by three disulphide bonds which are joined in a 1–5, 2–4, 3–6 pattern, although the first does not form in all members of the PDOM family defined by sequence homology. Up to six copies of this module have been found in numerous proteins that are mainly associated with the gastrointestinal tract (Fig. 7) and are thought to be growth factors that also contribute to mucosal defence mechanisms.

The particular function of PDOM in this environment has been related to its resistance to proteolytic degradation. Nevertheless, reports of P domains in the lysosomal α-glucosidase and in the lumenal sucrase isomaltase as well as in the major rabbit *zona pellucida* protein (Fig. 7) suggest a role in carbohydrate-binding, as all proteins containing PDOM exist in heavily glycosylated environments or, as in the case of the two enzymes, directly interact with sugars. This suggestion is well supported by the pattern of conservation shown in Fig. 22, right hand side. Three aromatic residues line a cleft which has been suggested

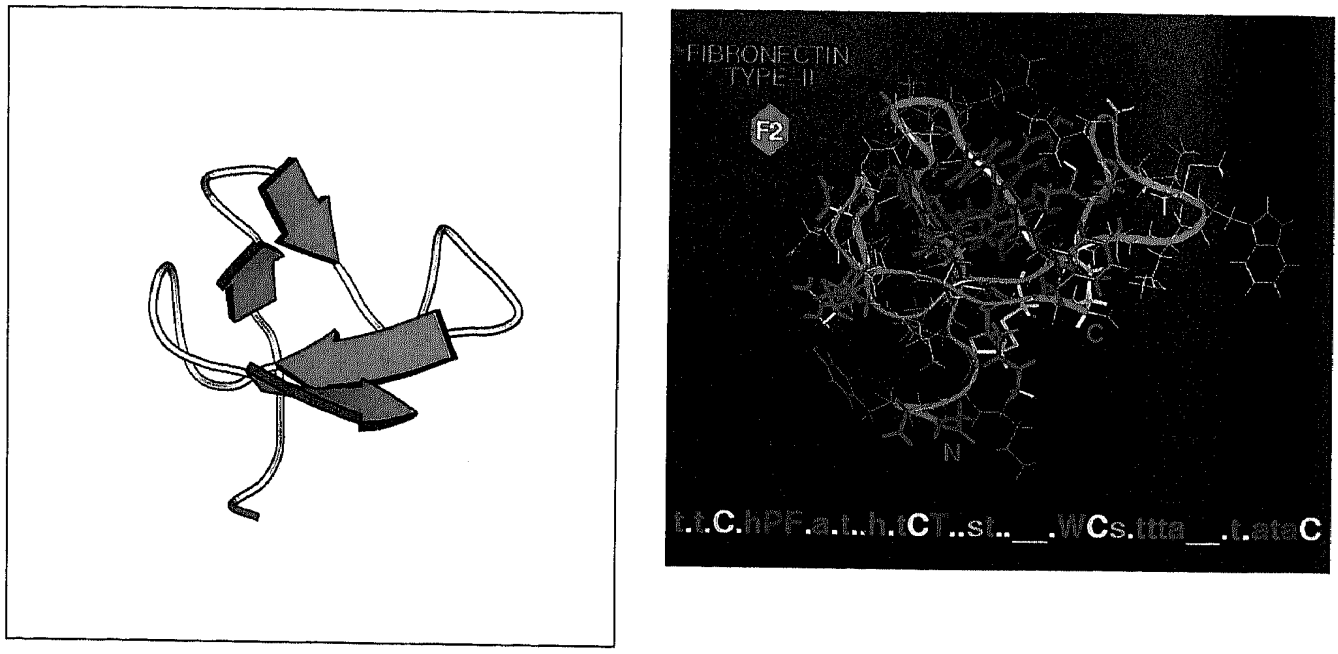

Fig. 21. Left) Schematic illustration of the fibronectin type-II domain fold. The structure of PDC-109 domain B (Constantine *et al.* 1992) has been used to represent this family. Right) Atomic resolution illustration of the FN2 domain highlighting the positions of conserved residues.

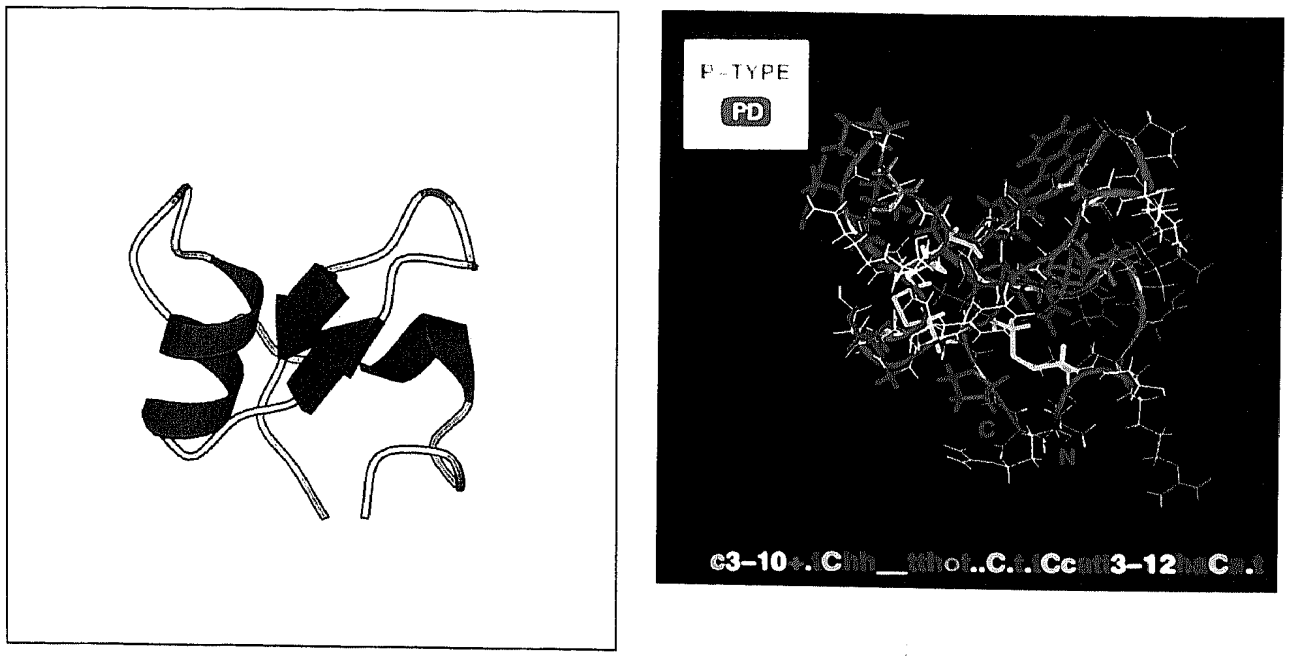

Fig. 22. Left) Schematic illustration of the P-domain fold. The structure of porcine pancreatic spasmolytic polypeptide (Gajhede *et al.* 1993) has been used to represent this family. Right) Atomic resolution illustration of the PDOM module highlighting conserved residues.

to bind mucin glycoproteins (Gajhede *et al.* 1993). These aromatic rings no doubt also play a role in stabilisation of the fold of the domain, as do the three hydrophobic residues clustered in the upper left-hand loop when the molecule is viewed in the orientation shown in Fig. 22, right hand side. This loop is further rigidified by multiple hydrogen bonds from the conserved hydroxyl (o) side chain

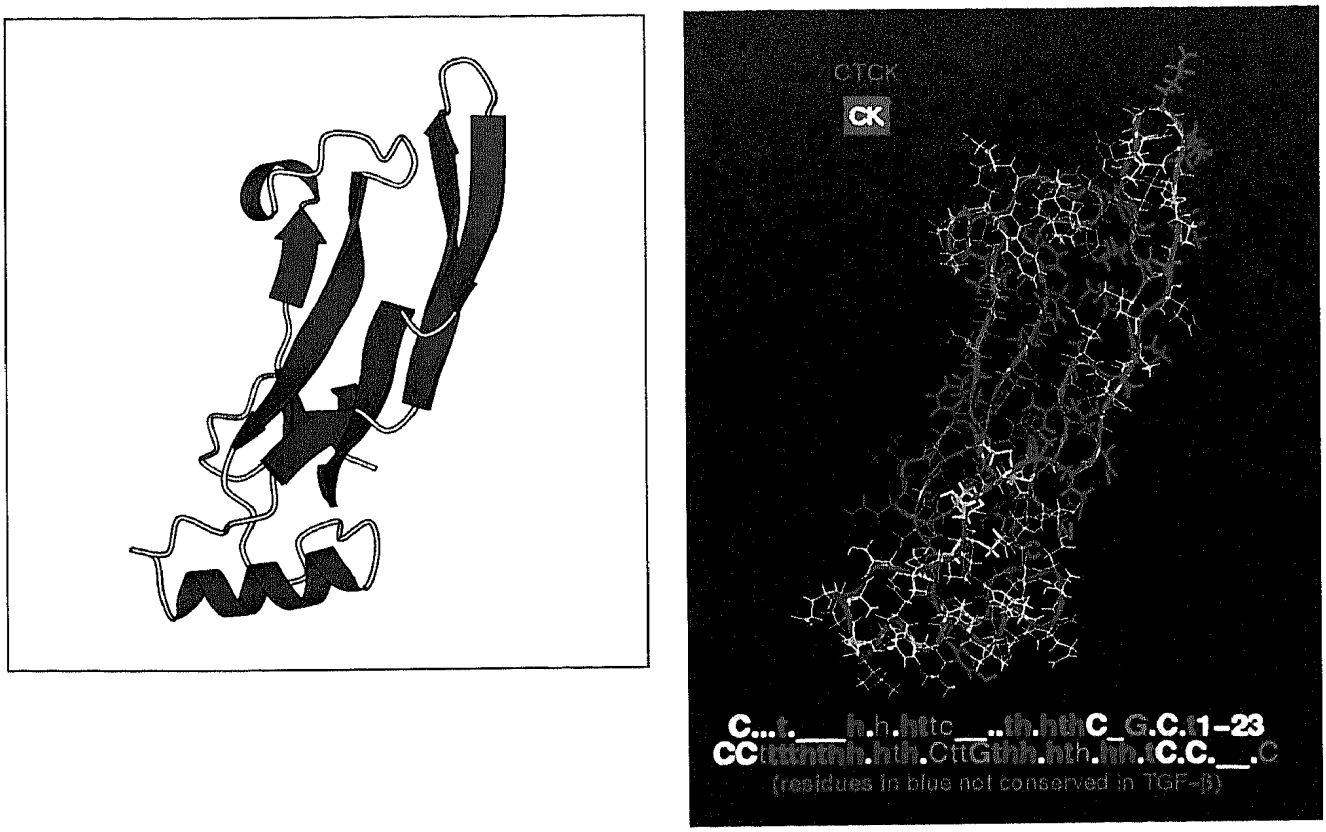

Fig. 23. Left) Schematic illustration of the cystine knot fold. The structure of transforming growth factor β-2 (Daopin *et al.* 1994) has been used to illustrate this family. Right) Atomic resolution illustration of the CTCK (TGF-β) domain. Only those residues which are conserved both in TGF-β and the cystine knot modules have been highlighted on the structure, the positions of amino acids conserved in the cystine knot family (but not in TGF-β) are noted in blue in the consensus sequence.

which initiates the longer, left-hand helix. Finally, the conserved positive side chain is unexpectedly buried in the core, and forms a salt bridge joining the first β-strand to the end of the short helix at the back of the structure in the orientation shown.

CK/CTCK/C-terminal cystine knot (TGF-β)

The three-dimensional structures of transforming growth factor-beta (TGFβ) (Daopin *et al.* 1994), nerve growth factor (NGF) (McDonald *et al.* 1991), platelet-derived growth factor (PDGF) (Oefner *et al.* 1992) and gonadotropin (Lapthorn *et al.* 1994; Wu *et al.* 1994) all form two highly twisted antiparallel pairs of β-strands and contain three disulphide bonds, of which two form a cystine ring through which passes the third cystine juncture. The topology, shown in Fig. 23, left hand side, has thus been called cystine-knot (McDonald *et al.* 1993; Isaacs, 1995). The CTCK module is notably non-globular; it has an extended, crescent shape with dimensions of $\sim$ 60 Å by 20 Å by 15 Å. The network of disulphide bonds formed between eight cysteines in a 1–2, 3–6, 4–7, 5–8 pattern is notable to the left of the N- and C-termini in Fig. 23, right hand side. There are few amino acids conserved in the region of the disulphide 'knot' apart from the cysteines,

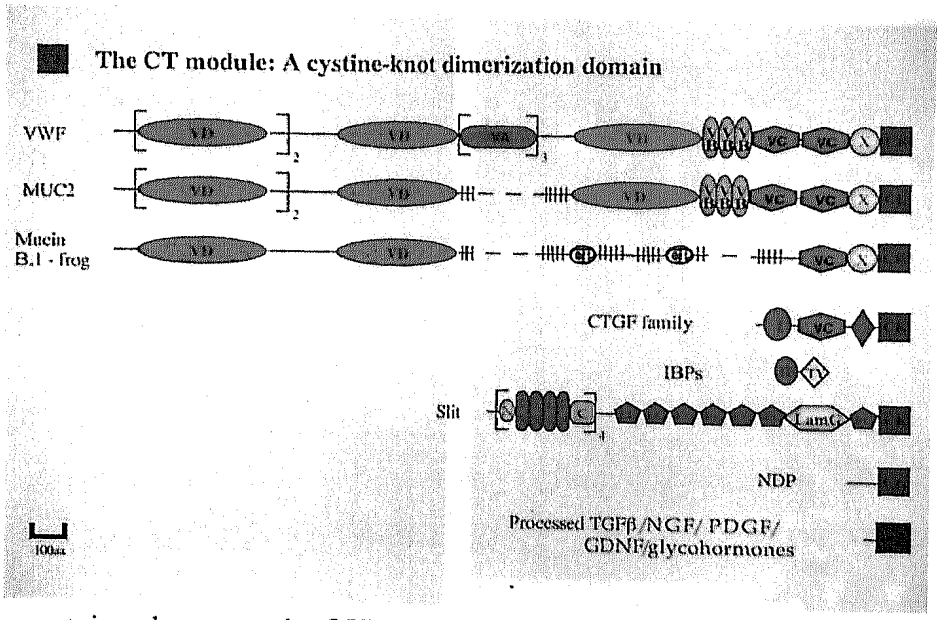

Fig. 24. Some proteins that contain CT modules at their C-terminus (modified from Meitinger *et al.* 1993). Note that the grouping of the CT sequence family into the cystine knot superfamily was only done after inclusion of structural information (model building).

while at the top of the molecule as shown, a number of hydrophobic residues are conserved, particularly between the two double-stranded β-sheets.

While TGFβ, NGF, PDGF and gonadotropin have not been reported to form parts of larger modular proteins, a distinct module that has been shown to be structurally related to this superfamily is found at the C-terminus of numerous modular proteins (Meitinger *et al.*, 1993). These include von Willebrand factor, slit, growth factors of the CTGF family, several mucins and norrie disease protein (Fig. 24). As all proteins of the cystine knot superfamily appear to form homodimers stabilised by a particular intermolecular disulphide bridge and complimentary hydrophobic interactions (Daopin *et al.* 1994), this feature is also predicted for the CTCK module in the modular proteins mentioned above (Meitinger *et al.* 1993). The exposure of some of the conserved hydrophobic residues shown in Fig. 23, right hand side is consistent with this suggestion. Also, dimerisation via a disulphide bridge in the C-terminal part has been reported in von Willebrand factor. Another common functional feature is likely to be the receptor or ligand binding, but different specificities can be expected due to sequence diversity.

CY/CYSTA/Cystatin-like

The cystatins are a family of cysteine protease inhibitors that mainly occur as single domain proteins and are also present in plants. As some extracellular animal proteins contain several CYSTA domains (Fig. 25), they can be considered as modules. Two cystatin domains are found in proteins like fetuin and a histidine-rich glycoprotein. The successive CYSTA modules are fused with distinct C-terminal parts in these proteins. In kininogens, three successive CYSTA modules form about half of the molecule, the C-terminal part harbours bradykinin (Fig. 25). In the latter well-studied protein with important regulatory functions, the

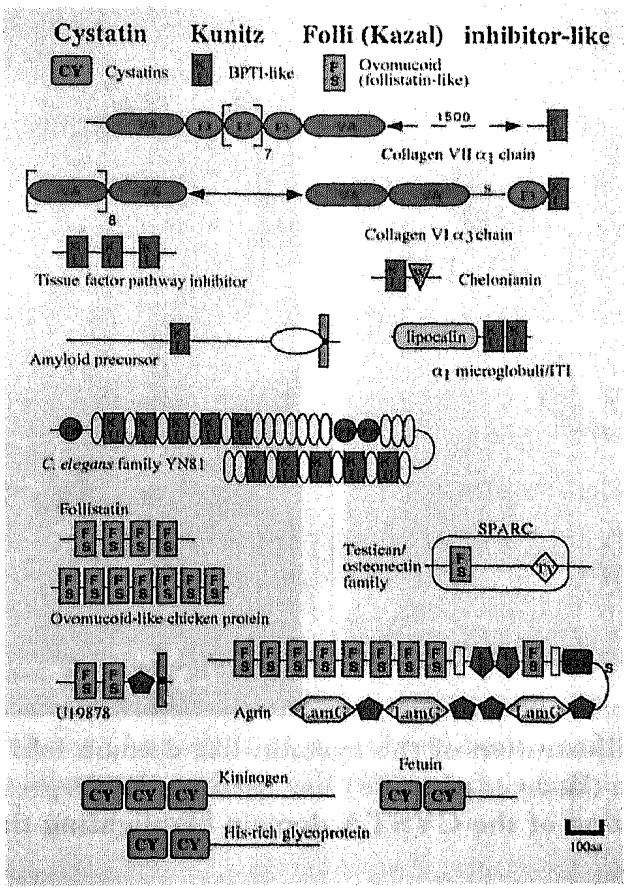

Fig. 25. Some proteins that contain modules of the protease inhibitor families CY, KU and FS (Kazal-type).

first domain seems to have lost its inhibitory activity (see PROSITE database documentation PDOC00259).

The cystatin molecule is approximately described as a prolate ellipsoid, with major and minor axes of ~ 45 and 30 Å, respectively. The fold is comprised of a large, twisted, antiparallel β-sheet which wraps around a long, straight α-helix. A shorter region of α-helix is also found packed against the base of the backside of the β-sheet as shown in Fig. 26, left hand side. All of the residues which are highly conserved in cystatin-like modules are hydrophobic or in turns. As one would expect, the majority of the conserved non polar side chains are buried in the interface between the major α-helix and β-sheet. Fig. 26, right hand side illustrates that the positions of turn residues mainly coincide with major changes in chain direction, e.g., at the ends of the long helix and at the elbow of the longest β-strand.

KU/KUNIT/Kunitz/BPTI inhibitor

BPTI is a very well-studied proteinase inhibitor. Many sequence relatives from various species have been known for a long time (Creighton & Charles, 1987). It has also been noted that single Kunitz-type like (BPTI-related) protease inhibitors (KUNIT) are inserted in the amyloid precursor protein and in collagens VI and VII or are fused with other single modules as in chelonianin. Two successive

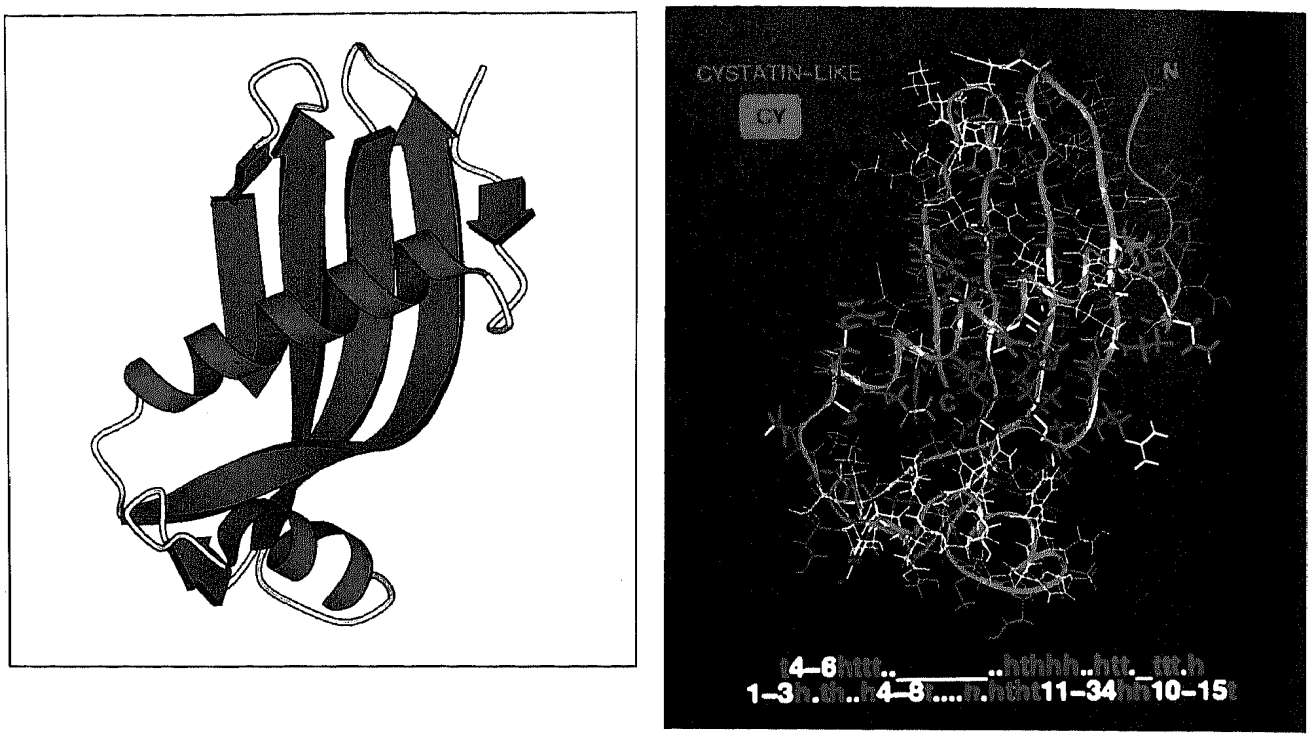

Fig. 26. Left) Schematic illustration of the cystatin-like domain fold. The structure of chicken egg white cystatin (Bode *et al.* 1988) has been used to represent this family. Right) Atomic resolution illustration of the CYSTA domain highlighting the positions of conserved residues.

domains exist in the precursor of inter-α-trypsin inhibitor, three in tissue factor pathway inhibitor, and the *C. elegans* genome project revealed huge proteins with at least 12 copies of KUNIT and several other modules (Fig. 25). It remains to be seen whether the KUNIT module acts as a protease inhibitor in all modular proteins or whether only the protease recognition features are used.

The KUNIT fold includes a short region of double-stranded β-sheet, which is flanked on either side by long loops, and a C-terminal α-helix (Fig. 27, left hand side). Of the three disulphides only the ones joining cysteines 1–6 and 3–5 form consistently. These link the N- and C-termini and the helix and sheet, respectively. At the top of the figure the other S-S bond joins the two long loops, probably reducing their relative motion. No hydrophobic residues are conserved in the Kunitz/BPTI sequence, only aromatic rings. Of these, the two phenylalanines which are strictly conserved are buried to a greater extent than the other three aromatic rings, and they appear to be critical to the fold of the domain. The conserved hydroxyl side chain (o) hydrogen bonds to the backbone at position i+3 and contributes to initiation of the first turn of the α-helix.

FS/FOLLI/Follistatin-like (Kazal-type protease inhibitors)

A third type of protease inhibitor that has been found in modular proteins are the Kazal-type (Fig. 25). However, only a strongly modified version of this inhibitor, first identified in follistatin as a repetitive domain, has apparently been subjected to shuffling events. FOLLI has doubtless sequence similarity to Kazal-type protease inhibitors (Fig. 28) but several disulphide bridges in the N-terminus

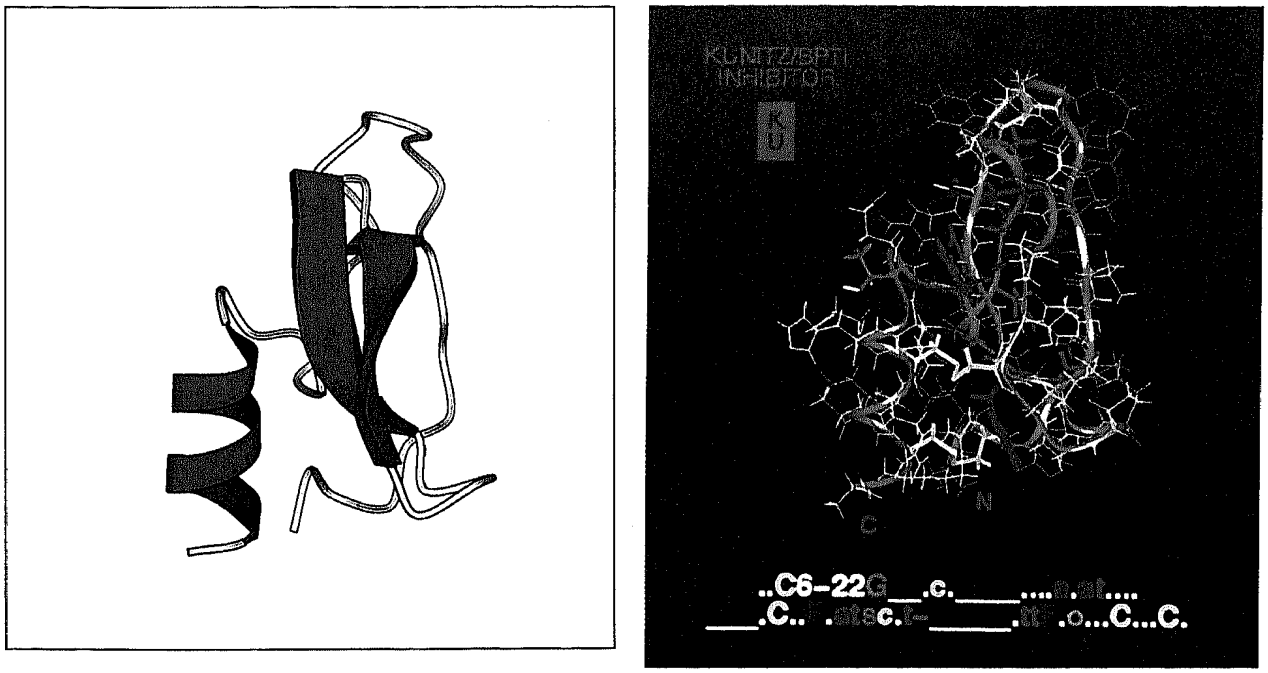

Fig. 27. Left) Schematic illustration of the Kunitz/BPTI fold. The structure of the sea anemone kunitz-type inhibitor (Antuch *et al.* 1993) has been used to illustrate this family. Right) Atomic resolution illustration of the KUNIT module highlighting the positions of conserved residues.

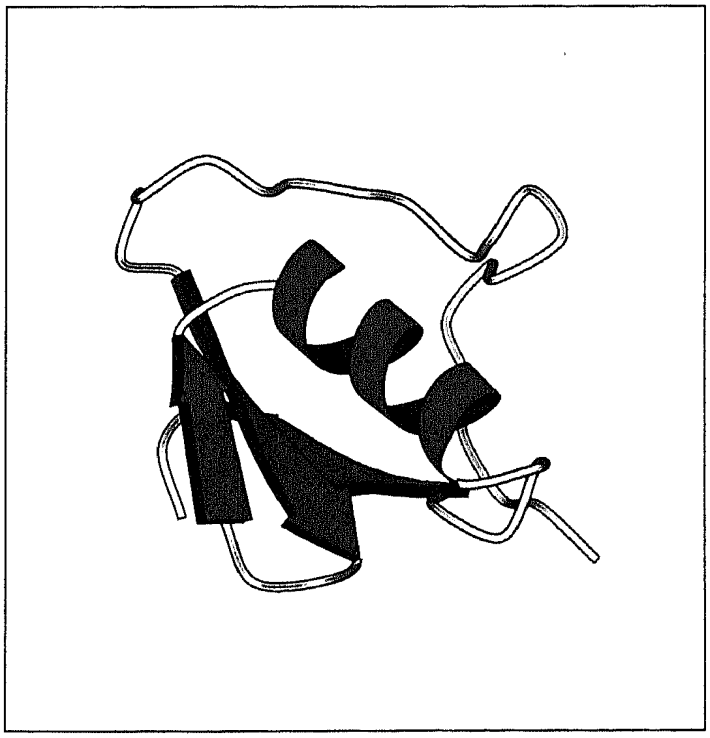

Fig. 28. Schematic illustration of the follistatin-like fold. The third domain of silver pheasant ovomucoid (Bode *et al.* 1985) has been used to represent this family.

seem to be rearranged and straight model-building by homology is difficult. FOLLI modules are found as part of the so-called SPARC unit of proteoglycans of the testican/osteonectin family; as many as 9 copies have been found in the

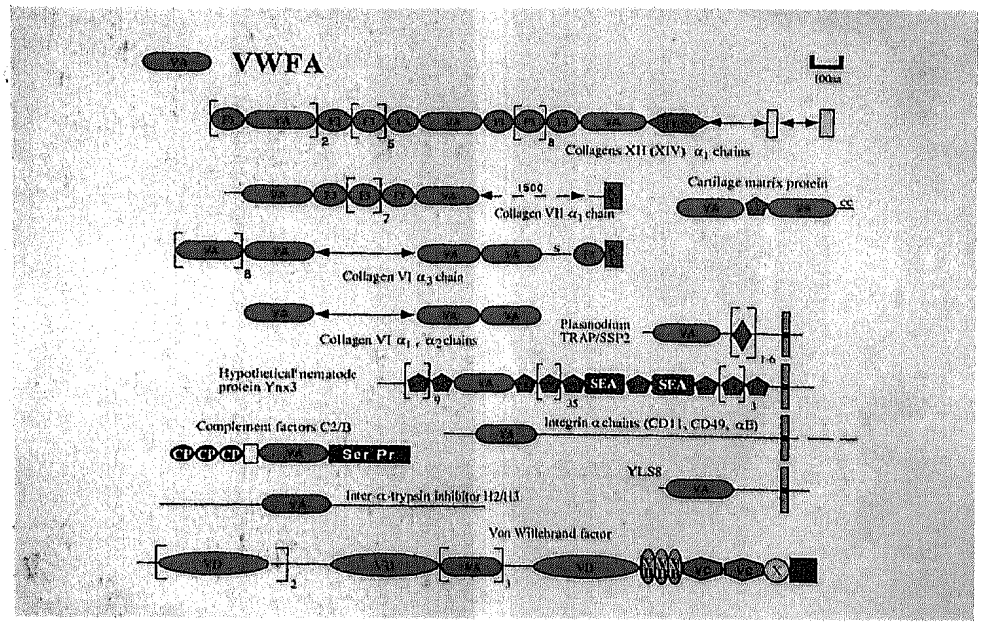

Fig. 29. Some proteins that have been reported to contain the VWA domain.

proteoglycan agrin (Fig. 25). Whether FOLLI acts as protease inhibitor or is associated with glycosylation remains to be studied.

5.2 *Known module structures not yet in the PDB*

New modules structures are appearing continuously. Some that have been published, and where the coordinates are not yet publicly available are described briefly here.

VA/VWFA/von Willebrand factor type A

Von Willebrand factor is a large multifunctional extracellular protein with several distinct module types. The most abundant one is the so-called 'type A repeat', a large domain occurring three times in von Willebrand factor itself, but also in numerous other extracellular proteins including collagens (Fig. 29). It does not contain disulphide bridges and is a rather large module with more than 200 amino-acids in length. The three-dimensional structure (Lee *et al.* 1995) of one VWFA in an integrin revealed a structural similarity to ras P21 and similar α/β-proteins. This suggests that a classical enzyme fold can be used as a module and that VWFA must have been subjected to diverse shuffling events. It has apparently lost its catalytic activity, but, as the metal-binding residues are well-conserved; metal-mediated-binding of numerous proteins might be a common functional theme of VWFA modules.

LA/LDLRA : LDL receptor class A

This module was originally identified in the LDL receptor where it occurs at the N-terminus in seven consecutive copies (Sudhof *et al.* 1985; Fig. 10) . Although the majority of LDLAs has been found in various LDR receptor -related proteins,

the module has been observed in many other proteins: for example, in complement proteins, enterokinase, basement membrane proteoglycan and PKD1 protein.

The structure of the first N-terminal module from the human LDL receptor has been determined recently (Daly *et al.* 1995); it is composed of a β-hairpin, followed by a series of β-turns . Many of the conserved acidic residues are located next to each other at one side of the molecule and are probably involved in binding functions

HX/HEMOP/Haemopexin-like

The module was originally found as 8 repetitive units within haemopexin, but 4 units were later identified in several other proteins including vitronectin (7 units), pea albumin PA-2(4) and some collagenases (4).

Structural analysis revealed that the units form a four-propeller β-superstructure (Li *et al.* 1995; Faber *et al.* 1995) similarly arranged to higher order β-propellers in enzymes such as neuraminidase, galactose oxidase or methanol dehydrogenase. Thus, the repetitive units appear to be repeats rather than modules; several of them are apparently required for structural stability and, probably, functional activity. Based on this structural observation one can, for example, predict that an eighth repeat will be present in vitronectin. Perhaps four repeats together form a haemopexin 'module' that seems to be involved in complex binding functions in several distinct proteins.

6. SOME GENERAL OBSERVATIONS ON MODULES

The following section outlines some interesting general features of modules. Without going into the difficult question of the origins of modules and how many kinds exist, some general structure related observations can be made. For example, modules with less regular secondary structure (e.g., KR and FN2) appear to have a higher proportion of conserved residues which stabilise their fold.

6.1 *Differences between members of one module family*

Generally where several structures have been determined for one particular module, the 3D structures have been found to be very similar. Homology modelling of one member from another is usually viable. This has been extensively studied in the case of the immunoglobulin family (Harpaz & Chothia, 1994). The changes between members of one family are usually found in loop lengths between secondary structure elements and sometimes in the addition of extra secondary structure elements. For example, in all the known structures of EGF and FN1 modules the observed structural changes between members of a family occur mainly in the length of the loops between β-strands. In contrast, proteins with Ig-like topology not only have changes in their loop lengths but also have various strands added to a core structure (Fig. 4; Bork *et al.* 1994).

6.2 *Conservation and variability of disulphide bridges*

Most of the extracellular modules contain disulphide bridges, probably to increase the stability of relatively small domains and to protect against proteolysis. The structural analysis above has suggested that a correlation may exist between the size of a module's hydrophobic core and the number of disulphide bridges which stabilise its fold. Almost all conserved cysteine residues in various extracellular modules form disulphide bridges, mostly intra-modular ones. Cysteines have thus been, for a long time, the most important identifier of a particular module by sequence analysis. With the exponential increase of sequence data, more and more exceptions to this rule become obvious. A well-known example is the immunoglobulin family, that contains a subgroup that does not contain the, otherwise conserved, disulphide bridge (Williams & Barclay, 1988). The recent 3D structures of several cell surface proteins containing Ig-like domains clearly demonstrate the switch of disulphide bridges between strands (Jones, 1993; Bork *et al.* 1994; Wagner & Wyss, 1994).

Domains of the FN3-type usually do not contain disulphide bridges, but the structure of neuroglian (Huber *et al.* 1994) shows that they sometimes do. Modules that contain the FN3 consensus have been found in CD45 and the numerous cysteines therein probably also form disulphide bridges. Structurally very similar to FN3, and possibly related in evolution (see below), are the N-terminal domains of various cytokine and related receptors. Whereas in the growth hormone receptor 3 disulphide bridges are formed within this domain, sequence related modules in other receptors only contain 2; the location of the cysteine residues in related sequences also suggests different strand connections.

Some C-type lectins contain an additional S-S bridge, while some members of the LY6UP, ANATO and KUNIT families seem to lack one. Within the TNFR family different Cys-bonding patterns are suggested by the multiple alignment. The cystine knot arrangement of 3 intra-module disulphide bridges is surrounded by differing Cys-bonding patterns in the different members (PDGF, TGFβ, NGF, glycohormone, and CTCK families); most of them also share a Cys bond connecting the dimers. The different arrangement of S-S bonds in the, sequence related, follistatin-like modules and Kazal-type protease inhibitors complicate 3D homology modelling of the follistatin-like modules from the Kazal inhibitor structures.

Disulphide bridges might stabilise certain conformations and are thus important and, usually conserved, structural features. However, loss, addition, or change of position of S-S bridges frequently occurs. These changes might lead to different stable topologies where the constraints on supporting hydrophobic core residues might change, leading to the formation of structurally related modules that are no longer detectable at the sequence level.

6.3 *Common topological features of different modules*

As the module size increases, the fraction of core-stabilising disulphide bridges

decreases and the hydrophobic core becomes obvious. Another striking feature of extracellular modules is that many of them are β-sheet and, in particular, several larger ones have very similar topology. This is especially the case with the Greek key architecture of the Ig, FN3, CYTR and CAD modules. A comparison was made recently (Bork *et al.* 1994) of 23 structures of members of the first three of these module types, that had less than 25 % pairwise residue identity. A structural core of four β-strands (b, c, e and f) was identified in all three types with three or five additional strands (a, c', c'', d and g). The structure of the additional strands is highly variable.

Structure comparisons of the different module types were carried out using a program that maximises a geometrical similarity score. Analysis of the pairwise structural similarity scores revealed three main structural clusters. However, it was concluded that the presence of the conserved structural core did not imply a similar hydrophobicity pattern among the different sequence families – only strand f appeared to retain conserved hydrophobic features. Compare, for example, the consensus sequences for FN3 and CADHE (see Figs 6, 8, right hand sides) and extend this comparison to other Ig-like domains (Bork *et al.* 1994). It can be observed that there are often conserved aromatic positions in one or more strands of the common core (Fig. 4) although they do not correspond to equivalent positions in the topology. Loop lengths vary in all positions between and even within sequence families of the Ig-type. This kind of analysis suggests that a common topology is achieved by fundamentally different sequences. It remains to be studied whether the different sequences have evolved from each other or whether the fold was invented several times independently during evolution.

Interestingly, the three functionally related, but structural distinct protease inhibitors (cystatin, Kunitz and Kazal) share some common topological features – they all contain a small antiparallel β-sheet with a surrounding helix.

7. MODULE ASSEMBLY

The database for individual module structures is extensive and knowledge about the way some modules fit together is growing. Biology seems to have used a limited set of protein structures. Are there also a limited number of ways in which modules fit together? Can we discern general rules for module assembly? Here we present a novel analysis of the geometry of different extracellular modular proteins, composed of at least two homologous modules of known structure. In this treatment, individual modules are treated as ellipsoid shapes and the geometry of the double module is described by the rotational parameters linking one module to the other. A rigid frame (**O1**, **x1**, **y1**, **z1**) is first assigned to the N-terminal module, with origin **O**, the calculated centre of mass, and unitary vector **z1**, aligned along the major axis of the inertia tensor of the module backbone atoms and pointing towards the C terminus (see Fig. 30a). A second frame (**O2**, **x2**, **y2**, **z2**) is then calculated for the C-terminal module by superimposing the homologous backbone atoms of conserved secondary structural elements. Three rotational and one translational parameter are necessary to describe the

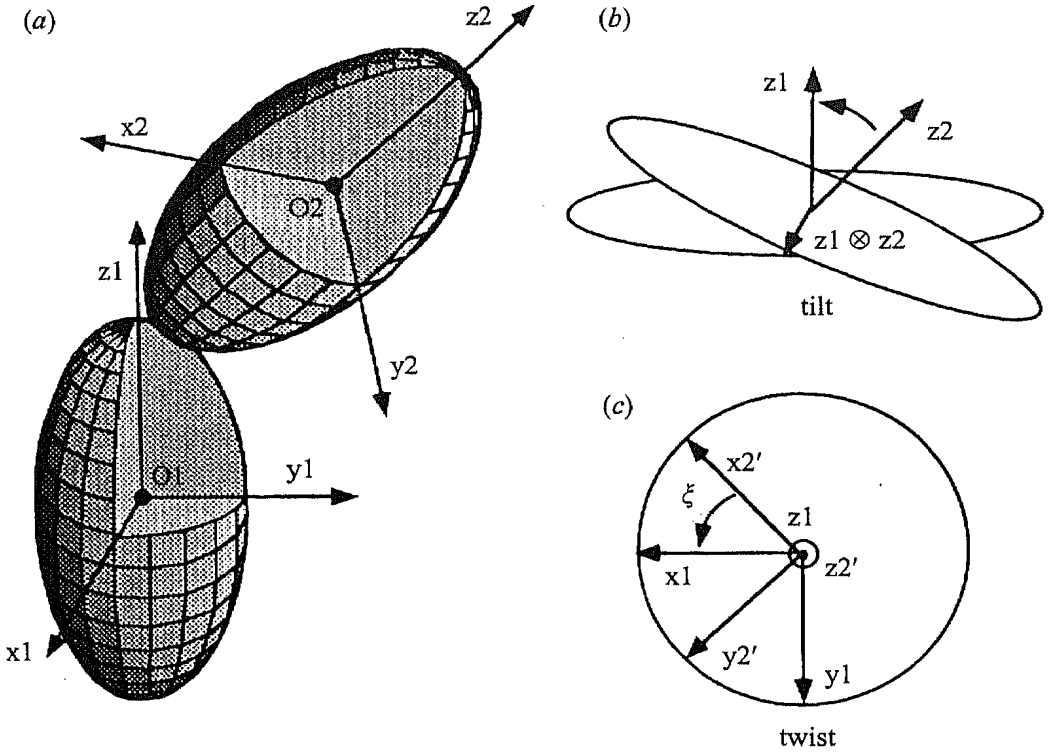

Fig. 30. *a*) Representation of a double module protein by two coordinate frames. *b*) Definition of the 'tilt' angle, ϵ; $Z_1 \otimes Z_2$ is the vector product of the Z axes of the two frames. *c*) Definition of the 'twist' angle, ζ; a top view of the two frames along the z axis is shown.

transformations linking the second frame to the first one. However, most of the modules shown have revolution symmetry around their major axis (two eigenvalues of the inertia tensor are nearly identical), therefore the orientation of the **x1** and **y1** vectors is arbitrary. In the main, the following two rotational parameters were thus used:

the tilt angle ϵ, given by: $\cos(\epsilon) = \mathbf{z1.z2}$ (Fig. 30*b*)

the twist angle ζ, given by: $\cos(\zeta) = \mathbf{x1.x2'}$ (Fig. 30*c*)

x2' is the orientation of the **x2** vector after tilting the second frame around the axis **z1⊗z2** by an angle ϵ (**z1** and **z2** are collinear after this transformation; **z1⊗z2** is the vector product of **z1** and **z2**).

The first step of this analysis requires identification and alignment of the C_α positions of the conserved structural elements. For example for the immunoglobulin modules pairs, the strands B, C, D, E and F were used in this alignment. The observed range of observed tilt (ϵ) and twist (ζ) angles for a number of solved module pairsare illustrated in Fig. 31 and Table 2.

A wide set of Fab structures has been recently made available (see e.g. Wilson & Stanfield 1994). These structures showed a high diversity in the module pairing geometry. We have analysed the geometry of five Fab domains with elbow angles ranging from 127° (1bbd) to 213° (8fab). In our study, the flexibility of Fab molecules translates into a large variation of the tilt and twist angles which are represented as relatively large grey boxes in the figure. The human growth

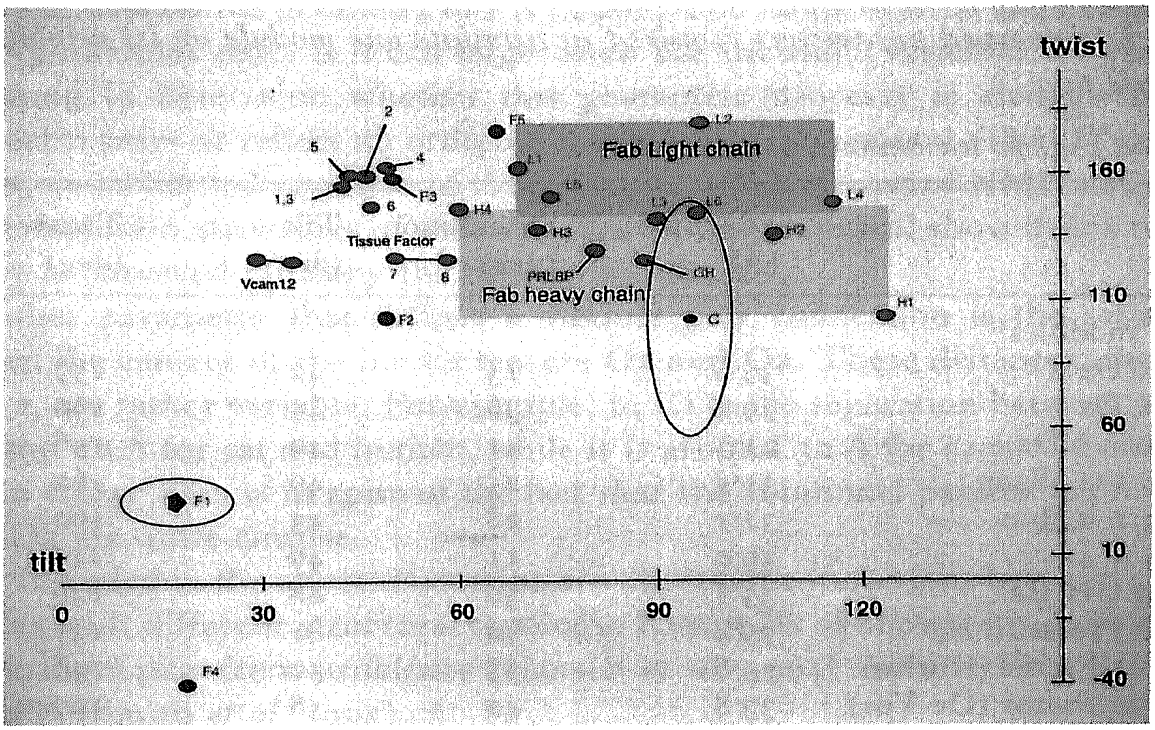

Fig. 31. Tilt and twist angles (see Fig. 30 for definition) for pairs of modules with structural similarity. The points refer to the following structures: 1, 2, rat CD2 (Jones *et al.* 1992); 3, human CD2 (Bodian *et al.* 1994); 4, human CD4 modules 1–2 (Wang *et al.* 1990); 5, human CD4 modules 3–4 (Brady *et al.* 1993); 7, tissue factor (Harlos *et al.* 1994); 8, tissue factor (Muller *et al.* 1994); Vcam12, vascular cell adhesion molecule (Jones *et al.* 1995); F1, fibronectin ^{4}FN1.^{5}FN1 pair (Williams *et al.* 1994); F2, fibronectin ^{7}FN3.^{8}FN3 (Leahy, 1995); F3, fibronectin ^{8}FN3.^{9}FN3 pair (Leahy, 1995); F4, fibronectin ^{9}FN3.^{10}FN3 pair (Leahy *et al.* 1996); F5, FN3 module pair from neuroglian (Huber *et al.* 1994); C, factor H CCP modules pair ^{15}C.^{16}C (Barlow *et al.* 1993); GH, human growth hormone receptor (de Vos *et al.* 1992); L1, H1 light and heavy chain of human rhinovirus neutralizing Fab fragment (1bbd) (Tormo *et al.* 1992); L2, H2, light and heavy chain of the NEW Fab (3fab) (Poljack *et al.* 1974); L3, H3. light and heavy chain of a monoclonal anti-arsonate antibody fab fragment (1fai) (Lascombe *et al.* 1992); L4, H4, light and heavy chain of fab fragment from the human myeloma IGg1 Hil (8fab) (Saul & Poljack 1992); L5, light chain of fab fragment from the influenza virus N9 neuraminidase-NC41 Fab complex (1nca) (Tulip *et al.* 1992), L6, light chain of fab fragment from hen lysosyme HyHEL5 fab complex (2iff) (Chacko *et al.* 1995).

hormone receptor belongs to this 'Fab cluster' as well as the C-module pair. In this later case however, the NMR study showed a large variation of the twist angle (represented by an ellipse in Fig. 31). This variation may be due to a true domain flexibility or a lack of definition at the domain interface.

The cell surface molecules CD2, CD4, tissue factor and VCAM 12 form a second cluster characterised by a smaller value of the tilt angle (around 40°). VCAM 12 has an Ig module pair with the lowest tilt angle yet observed. It is striking to observe that CD2 and the two CD4 pairs have a very similar geometry despite their different linker conformation. In many of the X-ray structures some variation between module pair angles is also apparent. For example, there is a 5° difference in the tilt angle between the two molecules in the crystal asymmetric unit in the VCAM 12 structure. There are also significant differences between the

Table 2. *Observed parameters required to translate one module on its neighbour in a module pair*

(See Fig. 30 for definitions.)

Type	Distance (Å)	Orientation (deg)	Tilt (deg)	Twist (deg)
Ig-like modules				
Rat CD2 a	39·8	13	42	157
Rat CD2 b	39·4	36	45	160
Human CD2	41·3	41	42	156
CD4 1–2 Wang	31·5	25	49	162
CD4 1–2 Ryu	31·1	28	43	160
CD4 3–4	29·0	11	46	149
Hum GH Re	33·1	50	88	126
Hum Prolactin Re	35·4	70	81	131
Tissue Factor (Harlos)	35·6	55	50	128
Tissue Factor (De Vos)	33·8	42	58	127
vcam12 a	37·4	20	34	126
vcam12 b	37·9	20	29	128
Fab domains (PDB names)				
1bbd H	29·9	8	124	104
3fab H	32·4	−10	107	136
1fai H	38·6	0	72	139
8fab H	38·4	5	60	147
1bbd L	39·6	−27	69	163
3fab L	39·4	−29	96	180
1nca L	38·3	21	74	152
2iff L	37·2	−25	96	145
1fai L	36·9	11	90	143
8fab L	31·9	15	116	148
Other modules				
F1 (4–5)	25·5	–	17	32
F3 (7–8)	37·7	–	48·4	104·2
F3 (8–9)	37·8	–	49·2	159·2
F3 (9–10)	37·5	–	19	−39
F3 (neuroglian)	37·3	–	65·8	177·9
Factor H 15–16	28·7	–	94·7	103

Distance	Distance in Å between the centre of mass of the two modules.
Orientation	This angle is only calculated for Ig-like modules since it requires the definition of an absolute frame on the first module.
Tilt	Angle between the major axis of the two modules (ϵ in Fig 31).
Twist	Rotation angle around the major axis of the second module relative to the first one once tilted (ζ in Fig. 31).

module join angles in the same structures from different group e.g. CD4 and tissue factor (see Table 2).

A third class of geometry is formed by the fibronectin modules pairs. The recent structure of a four modules fragment ^{7}FN3–^{10}FN3 (Leahy et al. 1995) provides

additional data on the geometry of FN3 pairs. The range of twist angles observed for FN3 modules pairs is much larger than the tilt angle variation. It would be interesting to determine whether this geometric property is characteristic of fibronectin pairs or reflect an ordering constraint in the crystal. The ^{4}FN1^{5}.FN1 pair showed a limited variation of the tilt angle within the set of NMR structures. Further solution structures of different fibronectin pairs will show if this relative rigidity is observed all along the fibronectin chain.

Another parameter that defines a module pair connection is the separation between the centres of the inertia tensors **O1** and **O2**. These distances, shown in Table 2, are rather variable. For example, in CD2 the separation between centres is around 40 Å for rat and human, while it is around 30 Å for each of the various pairs in CD4. This is in spite of the fact that the rotational parameters for CD2 and CD4 are quite similar.

Not included in the above discussion are observed connections between pairs of modules with different structural topology. Examples of known structure are the EGF-kringle pair from urokinase (Hansen *et al.* 1994) and the FN1-EGF pair from tPA (Smith *et al.* 1995).

In general it seems that rules are hard to discern for the ways that modules fit together. There are trends that can be noted – for example, twist angles of around 130° for Ig domains – but it appears that biology can readily change the ways in which a given module pair fits together. The structures of the individual modules are predictable but the relative orientation of a pair can be changed by changing the length of a linker peptide or a few surface residues. This means that a modular protein structure can provide great variety in the spatial position and type of surface that can be presented. The relative positions of modules in a modular protein can also be readily changed by environmental changes (ligand concentration, pH etc.), thus providing a regulation mechanism. Examples are the changes in angles observed between the receptor FN3 modules of the prolactin receptor and the growth hormone receptor when growth hormone binds – see Table 2 (Somers *et al.* 1994) or SH3, SH2 module rearrangement in intracellular signalling proteins, induced by phosphorylation (Pawson, 1995).

8. CONCLUSIONS

During the writing of this review there has been an astonishing increase in available sequences and structures; no doubt by the time this appears, several new module structures will be in the PDB database. In spite of this increase it is becoming clear that the total number of extracellular module types is probably limited to about 100 and the structures of about many of these are already known. What remains uncertain in many cases is the ways in which mosaic proteins are assembled and how they present their appropriate surfaces for interaction with other proteins. The present level of structural information is, however, sufficient to construct first approximation models for many mosaic proteins. This is allowing a new phase in the determination of biological function since site-directed amino-acid changes and domain deletion and swapping experiments can

now be done in a much more rational way than was previously possible. A new dimension in module research will be added soon with the completion of the genome sequencing programmes for several multicellular organisms. The availability of the complete gene pool of an organism and thus of all extracellular proteins will allow a new kind of comparative analysis. This should give unprecedented information about the phylogenetic questions concerning the spread of modules. More importantly, there is the possibility of unravelling many of the functional networks in which modules are found. Perhaps we are about to understand the 'language' of modules.

9. ACKNOWLEDGEMENTS

IDC is a member of the Oxford Centre for Molecular Sciences that is funded by MRC, BBSRC and EPSRC. IDC and AKD acknowledge financial support from the Wellcome Trust. AKD was funded by St. John's College. We thank Drs Leahy, de Vos, Harlos and Jones for access to coordinates before they were available in the PDB database.

REFERENCES

ANTUCH, W., BERNDT, K. D., CHAVEZ, M. A., DELFIN, J. & WÜTHRICH, K. (1993). The NMR solution structure of a kunitz-type proteinase inhibitor from the sea anemone stichodactyla helianthus. *Eur. J. Biochem.* **212**, 675–684.

BANNER, D. W., D'ARCY, A., JANES, W., GENTZ, R., SCHOENFELD, H.-J., BROGER, C., LOETSCHER, H. & LESSLAUER, W. (1993). Crystal structure of the soluble human 55 Kd TNF receptor-human TNF-β complex: Implications for TNF receptor activation. *Cell* **73**, 431–445.

BARCLAY, A. N., BIRKELAND, M. L., BROWN, M. H., BEYERS, A. D., DAVIS, S. J., SOMOZA, C. & WILLIAMS, A. F. (1993). The Leukocyte Anitgen. Academic Press.

BARLOW, P. N., STEINKASSERER, A., NORMAN, D. G., KIEFFER, B., WILES, A. P., SIM, R. B. & CAMPBELL, I. D. (1993). Solution structure of a pair of complement modules by Nuclear Magnetic Resonance. *J. Mol. Biol.* **232**, 268–284.

BARON, M., NORMAN, D. & CAMPBELL, I. D. (1991). Protein modules. *TIBS* **16**, 13–17.

BAZAN, J. (1993). Emerging families of cytokines and receptors. *Curr. Biol.* **3**, 603–606.

BODE, W., ENGH, R., MUSIL, D., THIELE, U., HUBER, R., KARSHIKOV, A., BRZIN, J., KOS, J. & TURK, V. (1988). The 2·0 Å X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. *EMBO J.* **7**, 2593–2599.

BODE, W., EPP, O., HUBER, R., LASKOWSKI, M. J. & ARDELT, W. (1985). The crystal and molecular structure of the third domain of silver pheasant ovomucoid (OMSVP3). *Eur. J. Biochem.* **147**, 387–395.

BODIAN, D. L., JONES, E. Y., HARLOS, K., STUART, D. I. & DAVIS, S. J. (1994). Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2·5 Å resolution. *Structure* **2**, 755–766.

BORK, P. & BAIROCH, A. (1995). Extracellular protein modules. *TIBS* **02** (Supplement).

BORK, P. & DOOLITTLE, R. F. (1992). Proposed acquisition of an animal domain by bacteria. *Proc. Natl. Acad. Sci. USA* **89**, 8890–8994.

BORK, P., HOLM, L. & SANDER, C. (1994). The Immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* **242**, 309–320.

BORK, P. & MARGOLIS, B. (1995). A phosphotyrosine interaction domain. *Cell* **80**, 693–694.

BRADY, R. L., DODSON, E. J., LANGE, G., DAVIS, S. J., WILLIAMS, A. F. & BARCLAY, A. N. (1993). Crystal structure of domains 3 and 4 of rat CD4: relation to the NH_2-terminal domains. *Science* **260**, 979–983.

CAMPBELL, I. D. & BORK, P. (1993). Epidermal growth factor-like modules. *Curr. Opin. Struct. Biol.* **3**, 385–392.

CAMPBELL, I. D. & DOWNING, A. K. (1994). Building protein structure and function from modular units. *TIBTECH* **12**, 168–172.

CHACKO, S., SILVERTON, E., KAMMORGAN, L., SMITHGILL, S., COHEN, G. & DAVIES, D. (1995). Structure of an antibody lysozyme complex unexpected effect of a conservative mutation. *J. Mol. Biol.* **245**, 261–274.

CONSTANTINE, K. L., MADRID, M., BÁNYAI, L., TREXLER, M., PATTHY, L. & LLINÁS, M. (1992). Refined solution structure and ligand-binding properties of PDC-109 domain b; a collagen-binding type II domain. *J. Mol. Biol.* **223**, 281–298.

CREIGHTON, T. E. & CHARLES, I. G. (1987). Biosynthesis, processing, and evolution of bovine pancreatic trypsin-inhibitor. *Cold Spring Harbor Symposia on Quantitative Biology* **52**, 511–519.

DALY, N., SCANLON, M. J., DJORDJEVIC, J. T., KROON, P. A. & SMITH, R. (1995). 3-dimensional structure of a cysteine-rich repeat from the low-density-lipoprotein receptor. *Proc. Natl. Acad. Sci. USA* **92**, 6334–6338.

DAOPIN, S., PIEZ, K. A., OGAWA, Y. & DAVIES, D. R. (1994). Crystal structure of transforming growth factor-$\beta 2$: An unusual fold for the superfamily. *Science* **257**, 369–373.

DE VOS, A. M., ULTSCH, M. & KOSSIAKOFF, A. A. (1992). Human growth hormone and the extracellular domain of its receptor: Crystal structure of the complex. *Science* **255**, 306–312.

DE VOS, A. M., ULTSCH, M. H., KELLY, R. F., PADMANABHAN, K., TULLINSKY, A., WESTBROOK, M. L. & KOSSIAKOFF, A. A. (1992). Crystal structure of the kringle 2 domain of tissue plasminogen activator at 2·4 Å resolution. *Biochem.* **31**, 270–279.

DOOLITTLE, R. F. (1995). The multiplicity of domains in proteins. *Ann. Rev. Biochem.* **64**, 287–314

DOWNING, A. K., DRISCOLL, P. C., HARVEY, T. S., DUDGEON, T. J., SMITH, B. O., BARON, M. & CAMPBELL, I. D. (1992). The solution structure of the fibrin binding finger domain of tissue-type plasminogen activator determined by ^{1}H NMR. *J. Mol. Biol.* **225**, 821–833.

DRICKAMER, K. (1992). Engineering galactose-binding activity into a C-type mannose-binding protein. *Nature* **360**, 183–186.

FABER, H. R., GROOM, C. R., BAKER, H. M., MORGAN, W. T., SMITH, A. & BAKER, E. N. (1995). 1·8-Å crystal-structure of the C-terminal domain of rabbit serum hemopexin. *Structure* **3**, 551–559.

FLETCHER, C. M., HARRISON, R. A., LACHMANN, P. J. & NEUHAUS, D. (1994). Structure of a soluble, glycosylated form of the human complement regulatory protein CD59. *Structure* **2**, 185–199.

GAJHEDE, M., PETERSEN, T. N., HENRIKSEN, A., PETERSEN, J. F. W., DAUTER, Z., WILSON, K. S. & THIM, L. (1993). Pancreatic spasmolytic polypeptide: first three-dimensional structure of a member of the mammalian trefoil family of peptides. *Structure* **1**, 253–262.

GLUCKSMANNKUIS, M. A., TAYBER, O., *et al.* (1995). Polycystic kidney-disease – the complete structure of the PKD1 gene and its protein. *Cell* **81**, 289–298.

HANSEN, A. P., PETROS, A. M., MEADOWS, R. P., NETTESHEIM, D. G., MAZAR, A. P., OLEJNICZAK, E. T., XU, R. X., PEDERSON, T. M., HENKIN, J. & FESIK, S. W. (1994). Solution structure of the amino-terminal fragment of urokinase-type plasminogen activator. *Biochemistry* **33**, 4847–4864.

HARLOS, K., MARTIN, D. M. A., O'BRIEN, D. P., JONES, E. Y., STUART, D. I., POLIKARPOV, I., MILLER, A., TUDDENHAM, E. G. D. & BOYS, C. W. G. (1994). Crystal structure of the extracellular region of human tissue factor. *Nature* **370**, 662–666.

HARPAZ, Y. & CHOTHIA, C. (1994). Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.* **238**, 528–539.

HOMMEL, U., HARVEY, T. S., DRISCOLL, P. C. & CAMPBELL, I. D. (1992). Human epidermal growth factor. High resolution solution structure and comparison with human TGF-α. *J. Mol. Biol.* **227**, 271–282.

HUBER, A. H., WANG, Y. E., BIEBER, A. J. & BJORKMAN, P. J. (1994). Crystal structure of tandem type III fibronectin domains from Drosophila neuroglian at 2·0 Å. *Neuron* **12**, 717–731.

HUGLI, T. E. (1981). The structural basis for anaphylatoxin and chemotactic functions of C3a, C4a, and C5a. Critical reviews in immunology Ed. M. Z. Atassi. CRC Press. 321–366.

ISAACS, N. W. (1995). Cystine knots. *Curr. Opin. Struct. Biol.* **5**, 391–395.

JONES, E. Y. (1993). The immunoglobulin superfamily. *Curr. Opin. Struct. Biol.* **3**, 846–852.

JONES, E. Y., DAVIS, S. J., WILLIAMS, A. F., HARLOS, K. & STUART, D. I. (1992). Crystal structure at 2·8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature* **360**, 232–239.

JONES, E. Y., HARLOS, K., BOTTOMLEY, M. J., ROBINSON, R. C., DRISCOLL, P. C., EDWARDS, R. M., CLEMENTS, J. M., DUDGEON, T. J. & STUART, D. I. (1995). Crystal structure of an integrin-binding fragment of vascular cell adhesion molecule-1 at 1·8 Å resolution. *Nature* **373**, 539–544.

KIEFFER, B., DRISCOLL, P. C., CAMPBELL, I. D., WILLIS, A. C., VAN DER MERWE, P. A. & DAVIES, S. J. (1994). Three-dimensional solution structure of the extracellular region of the complement regulatory protein CD59, a new cell-surface protein domain related to snake venom neurotoxins. *Biochemistry* **33**, 4471–4482.

KOBE, B. & DEISENHOFER, J. (1994). The leucine-rich repeat: a versatile binding motif. *TIBS* **19**, 415–421.

KOBE, B. & DEISENHOFER, J. (1995). A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature* **374**, 183–186.

KRAULIS, P. J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structure. *J. Appl. Cryst.* **24**, 946–950.

KREIS, T. & VALE, R., EDS. (1993). Guidebook to the extracellular matrix and adhesion proteins. Oxford University Press.

LABEIT, S. & KOLMERER B. (1995). Titins, giant proteins in charge of muscle ultrastructure and elasticity. *Science* **270**, 293–296.

LAPTHORN, A. J., HARRIS, D. C., LITTLEJOHN, A., LUSTBADER, J. W., CANFIELD, R. E., MACHIN, K. J., MORGAN, F. J. & ISSACS, N. W. (1994). Crystal-structure of human chorionic-gonadotropin. *Nature* **369**, 455–461.

LASCOMBE, M.-B., ALZARI, P. M., POLJAK, R. J. & NISONOFF, A. (1992). Three-

dimensional structure of two crystal forms of FabR19.9 from a monoclonal anti-arsonate antibody. *Proc. Natl. Acad. Sci. USA* **89**, 9429–9433.

LEAHY, D. J., AUKHIL, I. & ERICKSON, H. P. (1996). 2·0 Å crystal structure of a four domain segment of human fibronectin encompassing the RGD loop and synergy region *Cell* **84**, 155–164.

LEE, J.-O., RIEU, P., ARNAOUT, A. & LIDDINGTON, R. (1995). Crystal structure of the A domain from the α subunit of integrin CR3 (CD11b/CD18). *Cell* **80**, 631–638.

LI, J., BRICK, P., OHARE, M. C., SKARZYNSKI, T., LLOYD, L. F., CURRY, V. A., CLARK, I. M., BIGG, H. F., HAZLEMAN, B. L., CAWSTON, T. E. & BLOW, D. M. (1995). Structure of full-length porcine synovial collagenase reveals a C-terminal domain-containing a calcium-linked, 4-bladed β-propeller. *Structure* **3**, 541–549.

LIEPINSH, E., BERNDT, K. D., SILLARD, R., MUTT, V. & OTTING, G. (1994). Solution structure and dynamics of PEC-60, a protein of the Kazal type inhibitor family, determined by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.* **239**, 137–153.

LITTLE, E., BORK, P. & DOOLITTLE, R. F. (1994). Tracing the spread of fibronectin type-III domains in bacterial glycohydrolases. *J. Mol. Evol.* **39**, 631–643.

LOWELL, C. A., KLICKSTEIN, L. B., CARTER, R. H., MITCHELL, J. A., FEARON, D. T. & AHEARN, J. M. (1989). Mapping of the Epstein-Barr virus and C3dg binding sites to a common domain on complement receptor type 2. *J. Exp. Med.* **170**, 1931–1946.

MAIN, A. L., BARON, M., HARVEY, T. S., BOYD, J. & CAMPBELL, I. D. (1992). The three dimensional structure of the tenth type III module of fibronectin: an insight into RGD mediated interactions. *Cell* **71**, 671–678.

MALLETT, S. & BARCLAY, A. N. (1991). A new superfamily of cell surface proteins related to the nerve growth factor receptor. *Immunology Today* **12**, 220–227.

MARTINEZ, S. E., HUANG, D., SZCZEPANIAK, A., CRAMER, W. A. & SMITH, J. L. (1994). Crystal structure of chloroplast cytochrome *f* reveals a novel cytochrome fold and unexpected heme ligation. *Structure* **2**, 95–105.

MCDONALD, N. Q. & HENDRICKSON, W. A. (1993). A structural superfamily of growth factors containing a cystine knot motif. *Cell* **73**, 421–424.

MCDONALD, N. Q., LAPATTO, R., MURRAYRUST, J., GUNNING, J., WLODAWER, A. & BLUNDELL, T. L. (1991). New-protein fold revealed by a 2·3-Å resolution crystal-structure of nerve growth-factor. *Nature* **354**, 411–414.

MEITINGER, T., MEINDL, A., BORK, P., ROST, B., SANDER, C., HASSEMANN, M. & MURKEN, J. (1993). Molecular modelling of the norrie disease protein predicts a cystine knot growth-factor tertiary structure. *Nature Genetics* **5**, 376–380.

MULLER, Y. A., ULTSCH, M. H., KELLEY, R. F. & DE VOS, A. M. (1994). Structure of the extracellular domain of human tissue factor: location of the factor VIIa binding site. *Biochemistry* **33**, 10864–10870.

OEFNER, C., DARCY, A., WINKLER, F. K., EGGIMANN, B. & HOSANG, M. (1992). Crystal-structure of human platelet-derived growth factor-BB. *EMBO J.* **11**, 3921–3926.

OVERDUIN, M., HARVER, T. S., BAGBY, S., TONG, K. I., YAU, P., TAKEICHI, M. & IKURA, M. (1995). Solution structure of the epithelial cadherin domain responsible for selective cell-adhesion. *Science* **267**, 386–389.

PATTHY, L. (1991). Exons – original building blocks of proteins. *Bioessays* **13**, 187–192.

PATTHY, L. (1993). Modular design of proteases of coagulation, fibrinolysis, and complement activation: implications for protein engineering and structure-function studies. *Methods. Enzymol.* **222**, 10–21.

PAWSON, T. (1995). Protein modules and signalling networks. *Nature* **373**, 573–580.

POLJAK, R. J., AMZEL, L. M., CHEN, B. L., PHIZACKERLEY, R. P. & SAUL, F. (1974). The

three-dimensional structure of the Fab' fragment of a human myeloma immunoglobulin at 2·0-Å resolution. *Proc. Nat. Acad. Sci. USA* **71**, 3440–3444.

POTTS, J. R. & CAMPBELL, I. D. (1994). Fibronection structure and assembly. *Curr. Opin. Cell Biol.* **6**, 648–655.

RAO, Z., HANDFORD, P., MAYHEW, M., KNOTT, V., BROWNLEE, G. G. & STUART, D. (1995). The structure of a Ca^{2+}-binding epidermal growth factor-like domain: its role in protein-protein interactions. *Cell* **82**, 131–141.

RYU, S. E., KWONG, P. D., TRUNEH, A., PORTER, T. G., ARTHOS, J., ROSENBERG, M., DAI, X., XUONG, N., AXEL, R., SWEET, R. W. & HENDRICKSON, W. A. (1990). Crystal structure of HIV-binding recombinant fragment of human CD4. *Nature* **348**, 419–426.

SAUL, F. A. & POLJAK, R. J. (1992). Crystal structure of human immunoglobulin fragment fab new refined at 2·0 Å resolution. *PROTEINS* **14**, 363–371.

SHAPIRO, L., FANNON, A. M., KWONG, P. D., THOMPSON, A., LEHMANN, M. S., GRÜBEL, G., LEGRAND, J.-F., ALS-NIELSEN, J., COLMAN, D. R. & HENDRICKSON, W. A. (1995). Structural basis of cell-cell adhesion by cadherins. *Nature* **374**, 327–337.

SMITH, B. O., DOWNING, A. K., DRISCOLL, P. C., DUDGEON, T. J. & CAMPBELL, I. D. (1995). The solution structure and backbone dynamics of the fibronectin type 1 and epidermal growth factor-like pair of modules of tissue-type plasminogen activator. *Structure* **3**, 823–833.

SMITH, C. A., FARRAH, T. & GOODWIN, R. G. (1994). The TNF receptor superfamily of cellular and viral proteins: activation, costimulation, and death. *Cell* **76**, 959–962.

SOMERS, W., ULTSCH, M., DE VOS, A. M. & KOSSIAKOFF, A. A. (1994). X-ray structure of a growth hormone-prolactinreceptor complex. *Nature* **372**, 478–481.

SORIANO-GARCIA, M., PADMANABHAN, K., DE VOS, A. M. & TULINSKY, A. (1992). The Ca^{++} ion and membrane binding structure of the Gla domain of Ca-prothrombin fragment 1. *Biochemistry* **31**, 2554–2566.

SUDHOF, T. C., GOLDSTEIN, J. L., BROWN, M. S. & RUSSELL, D. W. (1985). The LDL receptor gene–a mosaic of exons shared with different proteins. *Science* **228**, 815–822.

SUNNERHAGEN, M., FORSEN, S., HOFFREN, A. M., DRAKENBERG, T., TELEMAN, O. & STENFLO, J. (1995). Structure of the Ca^{2+}-free GLA domain sheds light on membrane-binding of blood-coagulation proteins. *Nature Struct. Biol.* **2**, 504–509.

SUTTON, R. B., DAVLETOV, B. A., BERGHUIS, A. M., SÜDHOF, T. C. & SPRANG, S. R. (1995). Structure of the first C_2 domain of synaptotagmin 1: a novel Ca^{2+}/phospholipid-binding fold. *Cell* **80**, 929–938.

TORMO, J., STADLER, E., SKERN, T., AUER, H., KANZLER, O., BETZEL, C., BLAAS, D. & FITA, I. (1992). Three-dimensional structure of the Fab fragment of a neutralizing antibody to human rhinovirus serotype 2. *Protein Science* **1**, 1154–1161.

TULINSKY (1991). The structures of domains of blood proteins. *Thromb. & Haemo.* **66**, 16–31.

TULIP, W. R., VARGHESE, J. N., LAVER, W. G., WEBSTER, R. G. & COLMAN, P. M. (1992). Refined crystal structure of the influenza virus N9 neuraminidase-NC41 fab complex. *J. Mol. Biol.* **227**, 122–148.

VALCARCE, C., HOLMGREN, A. & STENFLO, J. (1994). Calcium-dependent interaction between gamma-carboxyglumtamic acid-containing and N-terminal epidermal growth factor-like modules in factor-X. *J. Biol. Chem.* **269**, 26011–26016.

VAN ZONNEVELD, A. J., VEERMAN, H. & PANNEKOEK, H. (1986). Autonomous function of structural domains on human tissue-type plasminogen activator. *Proc. Natl. Acad. Sci.* **83**, 4670–4674.

VENSTROM, K. A. & REICHARDT, L. F. (1993). Extracellular matrix 2: role of extracellular matrix molecules and their receptors in the nervous system. *FASEB J.* **7**, 997–1003.

WAGNER, G. & WYSS, D. F. (1994). Cell surface adhesion receptors. *Curr. Opin. Struct. Biol.* **4**, 841–851.

WALTER, M. R., WINDSOR, W. T., NAGABHUSHAN, T. L., LUNDELL, D. J., LUNN, C. A., ZAUODNY, P. J. & NARULA, S. K. (1995). Crystal structure of a complex between interferon-γ and its soluble high-affinity receptor. *Nature* **376**, 230–235.

WANG, J., YAN, Y., GARRET, T. P. J., LIU, J., RODGERS, D. W., GARLICK, R. L., TARR, G. E., HUSSAIN, Y., REINHERZ, E. L. & HARRISON, S. C. (1990). Atomic structure of a fragment of CD4 containing two immunoglobulin-like domains. *Nature* **348**, 411–418.

WARD, C. W., HOYNE, P. A. & FLEGG, R. H. (1995). Insulin and epidermal growth factor receptors contain the cysteine repeat motif found in the tumour necrosis factor receptor. *PROTEINS* **22**, 141–153.

WEIS, W. I., DRICKAMER, K. & HENDRICKSON, W. A. (1992). Structure of a C-type mannose-binding protein complexed with an oligosaccharide. *Nature* **360**, 127–134.

WILLIAMS, A. F. & BARCLAY, A. N. (1988). The immunoglobulin superfamily – domains for cell surface recognition. *Ann. Rev. Immunol.* **6**, 381–405.

WILLIAMS, M. J., PHAN, I., HARVEY, T. S., ROSTAGNO, A., GOLD, L. I. & CAMPBELL, I. D. (1994). Solution structure of a pair of fibronectin type I modules with fibrin binding activity. *J. Mol. Biol.* **235**, 1302–1311.

WILLIAMSON, M. P. & MADISON, V. S. (1990). Three-dimensional structure of porcine C5a(desArg) from ^{1}H nuclear magnetic resonance data. *Biochemistry* **29**, 2895–2905.

WILSON, I. A. & STANFIELD, R. L. (1994). Antibody-antigen interactions: new structures and new conformational changes. *Cur. Opin. in Struct. Biol.* **4**, 857–867.

WU, H., LUSTBADER, J. W., LIU, Y., CANFIELD, R. E. & HENDRICKSON, W. A. (1994). Structure of human chorionic-conadotropin at 2·6-angstrom resolution from MAD analysis of the selenomethionyl protein. *Structure* **2**, 545–558.