Development 1994 Supplement, 35-42 (1994)
Printed in Great Britain © The Company of Biologists Limited 1994

Domain organizations of extracellular matrix proteins and their evolution

Jürgen Engel, Vladimir P. Efimov and Patrik Maurer

Department of Biophysical Chemistry, Biozentrum, University of Basel, Switzerland

SUMMARY

The astonishing diversity in structure and function of extracellular matrix (ECM) proteins originates from different combinations of domains. These are defined as autonomously folding units. Many domains are similar in sequence and structure indicating common ancestry. Evolutionarily homologous domains are, however, often functionally very different, which renders function prediction from sequence difficult. Related and different domains are frequently repeated in the same or in different polypeptide chains. Common assembly domains include α-helical coiled-coil domains and collagen triple helices. Other domains have been shown to be involved in assembly to other ECM proteins or in cell binding and cell signalling. The function of most of the domains, however, remains to be elucidated. ECM proteins are rather recent 'inventions', and most occur either in plants or mammals but not in both. Their creation by domain shuffling involved a number of different mechanisms at the DNA level in which introns played an important role.

Key words: extracellular matrix proteins, protein domains, intron

INTRODUCTION

The extracellular matrix (ECM) is not just the glue between cells as believed for a long time. It is instead a highly elaborate association of proteins, proteoglycans and glycosaminoglycans, each of which has a specialized function in fulfilling the manifold purposes that the ECM has. The main purpose is serving the cell as a substrate for growth and providing a stable structure around them. This is a fundamental precondition for the existence of multicellular organisms. The central systems in eukaryotes (neural, circulatory, digestive and fertilization systems) evolved within and along with the ECM.

The ECM has to serve two masters: it must be a pleasant living space for the cell and a suitable scaffold for functional elements of the organism. To fulfil these purposes, a huge set of proteins and proteoglycans of unusual size and shape have evolved, which furnish the tissue with its distinct features and anchor the cell in its surroundings. Electron microscopy gives insight into a strange microcosm of crosses, spiders, strings of pearls, brushes, dumb-bells, rods and other oddities. The astonishing diversity in structure and function of proteins in this bizarre arsenal, however, originates from a building set of a limited number of modules.

Here we review the present knowledge of the fundamentals of domain organization and scrutinize functional assignments derived from experimental data and from sequence homology. The complex multidomain organization of the multifunctional ECM proteins offers a fascinating view on mechanisms of evolution. The apparent redundancy of certain ECM proteins opens questions on selective forces in evolution.

In order to provide a concise overview, only the more recent primary publications and review articles could be cited. References to the original literature can be found in these.

EXTRACELLULAR MATRIX PROTEINS ARE MOSAIC MULTIDOMAIN PROTEINS BUILT OF MODULAR UNITS

Extracellular matrix proteins are typical multidomain proteins (Table 1). Most domains show identity with domains of the same protein or with domains found either in other ECM proteins or in multidomain proteins not normally classified as ECM proteins. These include cell adhesion molecules (CAMs and cadherins), many cellular receptors including integrins, and proteins of the immune and complement system and of the blood clotting cascade. Because of this wide and repeating distribution of domains, these types of proteins have been termed mosaic proteins built of modular units (Doolittle 1985, 1992; Doolittle et al., 1986).

Table 1 summarizes the domain organizations as revealed by sequence information for a large number of ECM proteins. Table 1 is not complete and additional compilations can be found in Bork (1991, 1992), Baron et al. (1991), Engel (1991), Patthy (1991a,b), Bork and Doolittle (1992) and Kreis and Vale (1993). Such comparisons demonstrate the widespread distribution of domains in different classes of proteins. Examples are the EGF domains in proteins of the ECM, in blood clotting and complement systems and in a number of cell-surface receptors. IgG-domains occur not only in the immunoglobulin family, but also in proteoglycans, in cell-adhesion molecules such as N-CAM, and in receptors recognizing growth factors and carbohydrates. One of the most widespread module is the fibronectin type 3 (F3) domain of which more than 300 variants in about 70 proteins (not counting species redundancies) have been detected so far. They are found in both extracellular and intracellular proteins; for example, in the muscle protein twitchin and in the cytosolic domain of the integrin subunit β4.

Table 1. Domain organization of extracellular matrix proteins

Protein	Domain order†	Reference
fibronectin	$F1_6$ $F2_2$ $F1_3$ $F3_{15\text{-}18}$ $F1_3$	1
tenascin	TL CC EG_{13} $F3_{11\text{-}15}$ FG	2
thrombospondin 1, 2	TA CC PN PR_3 EG_3 EF_7 TC	3
thrombospondin 3, 4	TA CC EG_4 EF_7 TC	3
cartilage oligomeric matrix protein (COMP)	CC EG_4 EF_7 TC	
laminin α1 (Ae)‡	LA EG'_4 EG″ EG'_8 EG" EG'_3 CC LG_5	5
laminin β1 (B1e)	LA EG'_5 LB EG'_8 CC	5
laminin γ1 (B2e)	LA EG'_4 EG" EG'_6 CC	5
kalinin γ2 (B2t)	EG'_3 EG" EG'_4 CC	6
unc 6	LA EG'_3 UA	7
perlecan	PA EG_4 IG $(EG'' EG'_3)_3$ $IG_{14\text{-}21}$ LG EG_2 LG EG_2 LG	8
agrin	KA_8 EG'_2 KA ST_2 EG LG EG_2 LG EG_1 LG	9
nidogen/entactin	N1 EG N2 EG_5 N3	10
fibulin/BM-90	AN_3 EG EG^*_8 FB	11
fibrillin	FA EG_3 CR EG^*_2 TB PR EG EG^*_4 TB EG^*_3 CR EG^*_1 TB EG^*_{12} TB EG^*_2 TB EG^*_7 TB EG^*_5 TB EG^*_7 FB	12
osteonectin/SPARC/BM40	GR KA AL EF	13
procollagen I	PN TH PC	14
collagen IV	7S TH N4	14
collagen VI (α3 chain)	VA_9 TH VA_2 ST F3 PI	14
collagen XII	F3 VA F3 VA $F3_6$ VA $F3_{10}$ VA N9 TH	15
aggrecan	LI LI $KS_{0\text{-}1}$ $C1_{0\text{-}1}$ C2 EG LE CO	16

†Designations of domains are explained in Table 2, linear triple helical and coiled coil domains are underlined.
‡New chain designations for laminins are used (Burgeson et al., 1994), old designations are shown in brackets.
1, Hynes, 1990; 2, Spring et al., 1989; 3, Lawler et al., 1993; 4, Oldberg et al., 1992; 5, Beck et al., 1990; 6, Kallunki et al., 1992; 7, Ishii et al., 1992; 8, Noonan and Hassel, 1993; Kallunki and Tryggvason, 1991; 9, McMahan et al., 1992; 10, Mann et al., 1989; 11, Pan et al., 1993; Agraves et al., 1990; 12, Corson et al., 1993; 13, Engel et al., 1987; 14, Bork, 1992; 15, Yamagata et al., 1991; 16, Mörgelin et al., 1994.

It has to be pointed out that all classifications contain a degree of ambiguity and uncertainty, in some cases because of very low sequence identities, which might not reflect common evolutionary descent. It has been argued that in some cases similar domains were produced by convergent evolution, for example as discussed for EF-hand domains (Kretsinger, 1987).

DOMAINS ARE AUTONOMOUS STRUCTURAL UNITS

Domains may be defined by the sequence blocks which are repeated in the same protein or reoccur in different proteins. Often, however, it is difficult to define the exact starts and ends of domains on this basis. Recognition of a linker sequence which is normally hydrophilic may help in some cases. Domains are often encoded by single exons, but this cannot be an absolute rule since introns can be secondarily introduced into exons during evolution. In the present work a domain is defined as an autonomous, independently-folded, structural unit. The most stringent proof for the structural independence of domains comes from three-dimensional structures. These have been derived by NMR and X-ray diffractions for a number of domains (Table 2; Baron et al., 1991). Earlier indications of a conformational independence of fibronectin and laminin domains were based on circular dichroism studies, which indicated additivity of the spectra of different fragments (Odermatt et al., 1982; Ott et al., 1982). A powerful method for distinguishing individual domains in regions with sequence repeats is based on the resistance of recombinantly prepared fragments against proteolytic susceptibility (Winograd et al., 1991). The structural integrity of separated domains has also been demonstrated for many ECM proteins by the preserved biological functions of domains and fragments. Recently, the three dimensional structures of a pair of fibronectin-type 1 (F1) domains in fibronectin (Williams et al., 1993), a pair of complement control (CO) domains in factor H (Barlow et al., 1993) and a lectin-EGF-module pair (Graves et al., 1994) were resolved. The secondary structure of each module within a pair conformed closely with the structure of the separated single domains, implying that modules fold entirely autonomous within intact proteins.

THE DOMAIN ORGANIZATION OF ECM PROTEINS

The example of the F1 pair in fibronectin demonstrated the potential of NMR for elucidation of the geometry of domain organizations. The NMR technique is limited, however, to structures of smaller molar mass than about 20 000. X-ray analysis is applicable to larger structures but in this case crystallisation of large ECM molecules is a severe problem. Electron microscopy, therefore, is one of the most powerful techniques for elucidation of larger domain organizations (Engel, 1994). Fibronectin (Hynes, 1990; Odermatt et al., 1982), laminin (Beck et al., 1990), thrombospondin (Lawler, 1986) and tenascin (Spring et al., 1989; Erickson, 1993) are examples in which this technique, in combination with hydrodynamic data and sequence analyses, yielded detailed information (Fig. 1). EGF domains are found in all four of these proteins, in linear arrangements with a repeat of 2-2.5 nm per domain. The F1, F2 and F3 domains in fibronectin are also in a linear array which is, however, strongly dependent on ionic strength (Markovic et al., 1983), indicating a solvent-dependent internal association of domains. Likewise, the arms of thrombospondin and cartilage oligomeric matrix protein

Table 2. Domains in extracellular matrix proteins

Code	Name	Structure	Function
AN	Anaphylatoxin	X-ray (Huber et al., 1980)	Occurs in complement components C3, C4, C5
AL	α-helical domain BM-40		Collagen binding (Pottgiesser et al.,1994)
CC	Heptad repeat region	Forms α-helical coiled coil structures with partner chains, X-ray of leucine zipper in GCN4 (Harbury et al., 1993)	Connects 3 laminin chains (Beck et al., 1990), 3 tenascin chains (Spring et al., 1989), 3 thrombospondin (Lawler, 1986), 3 fibrinogen and 5 COMP chains (Efimov et al., 1994)
C1	Chondroitin sulfate binding	Stretched polypeptide chains	Glycosaminoglycan attachment to gly-ser sequences
C2	domains 1, 2	(Mörgelin et al., 1994)	
CO	Complement control domain	NMR (Baron et al., 1991)	Frequent in complement proteins
EF	EF hand-like domain	X-ray structures of cytosolic proteins	Ca^{2+} binding to Ca^{2+} modulated proteins, to osteonectin (Pottgiesser et al., 1994) and thrombospondin (Lawler, 1986)
EG	EGF-like domain	NMR of EGF (Baron et al.,1991), X-ray of EG domain in E-selectin (Graves et al., 1994)	Growth promotion in EGF, putative binding domain for proteins (Davis, 1990)
EG*	EGF domain, Ca2+ binding	NMR of first EG* domain in factor IX (Baron et al., 1991)	Ca^{2+} binding demonstrated for factor IX, for fibrillin (Corson et al., 1993) and other proteins
EG′	Laminin type EGF domain	X-ray and NMR work in progress†	EG′ domain in laminin γ1 chain is a specific binding site for nidogen (Mayer et al., 1993), growth promoting functions for unspecified EG′ domains in laminin were proposed (Panayotou et al., 1989)
EG″	Laminin type extended EGF domain		Cleavage during activation of RGD site in the short arms of murine laminin (Beck et al., 1990)
FA	N-terminal domain in fibrillin		
FB	C-terminal domain in fibrillin		
F1	Fibronectin type 1 domain	NMR of F1 and a pair of F1 domains in fibronectin (Baron et al., 1991; Williams et al., 1993)	N-terminal fragment of 5 F1 domains in fibronectin binds fibrinogen and heparin (Hynes, 1990)
F2	Fibronectin type 2 domain	NMR (Constantine et al.,1992)	The first and second F2 domain in fibronectin are involved in gelatine binding (Hynes, 1990)
F3	Fibronectin type 3 domain	NMR of 10th F3 domain in fibronectin (Baron et al., 1991); X-ray of F3 domain in tenasin (Leahy et al., 1992)	10th F3 domain in fibronectin is the major RGD dependent cell binding domain. Another more C-terminal F3 domain is also involved in cell binding (Hynes, 1990)
FG	Fibrinogen γ domain		Cell binding in tenascin (Joshi et al., 1993)
GR	Glutamic acid rich domain		
IG	IgG domain	X-ray of many immunoglobulins	Antigen binding and other recognition functions in IgG
KA	Kazal inhibitor-like domain	X-ray of Kazal inhibitor of Kazal family (Bode and Huber, 1992)	Homology with follistatin, ovomucoid. Kazal inhibitors inhibit proteases by binding to their active site but no such affinity was found for KA domains in ECM proteins
KS	Keratan sulfate binding domain	Attachment of keratan sulfate	
LE	Mammalian lectin domain	X-ray of LE in E-selectin (Graves et al., 1994), mannose binding protein (Weis et al., 1992)	LE in aggrecan binds saccharides with low activity but physiological target unknown
LG	C-terminal domains of laminin α-chain		Homology with steroid binding hormone (Beck et al., 1990), splice variants in agrin cluster acetylcholine receptor (McMahan et al., 1992)‡
LA	N-terminal laminin domain		Involved in self-association (Beck et al., 1990)
LI	Link protein domain		The first LI domain in aggrecan binds hyaluronan, the second not (Mörgelin et al., 1994)
N1	N-terminal nidogen domain		
N2	Central nidogen domain		Binds collagen IV and perlecan (Beck et al., 1990)
N3	C-terminal nidogen domain		Bind to a distinct EG′ domain of laminin γ1-chain (Mayer et al., 1993)
N4	Non-collagenous domain in collagen IV		
N9	Non-collagenous domain in collagen IX		
PA	N-terminal perlecan domain		
PC	C-terminal procollagen I domains		
PN	N-terminal procollagen I domains		
PI	BPTI-like domain	X-ray of of inhibitors of the class of the bovine pancreatic trypsin inhibitor family (Bode and Huber, 1992)	Small inhibitors inhibit specific proteases by binding to their active site but no such affinity was found for PI domains in ECM proteins
PR	Properdin domain		Binding to membranes?
ST	Serine/threonine rich domain		
TB	TGF-β binding protein domain		Proposed to be involved in TGF-β binding (Corson et al., 1993)
TA	N-terminal domain of thrombospondin		
TC	C-terminal domain of thrombospondin		Involved in cell binding (Lawler, 1986)
TH	Gly-X-Y repeat region	Three chains form a collagen triple helix	Connects 3 chains in various collagens
TL	N-terminal domain of tenascin		Probably involved in linking trimers to hexamers (Erickson, 1993)
UA	C-terminal unc-6 domain		
VA	von Willebrand factor A domain		VA domain in von Willebrand factor binds collagen (Colombatti and Bernaldo, 1991)
7S	N-terminal domain in collagen IV		Connects 4 collagen IV molecules in an antiparallel arrangement

For simplicity all domains are designated by a code of two letters or a letter and a number. Very obvious variants of domains within each homology class are indicated by dashes or asterisks. The subscripts indicate the number of repeated domains with internal homology. Ranges of numbers indicate the existence of splice variants.

†R.Huber and R. Timpl, personal communication. ‡Binding to α-dystroglycan (Sealock and Froehner, 1994).

(Mörgelin et al., 1992) are extended. The stretched arms of laminin and fibronectin exhibit a limited flexibility comparable with that of actin filaments (Engel et al., 1981); hence these domains are not loosely connected but interact with each other (as exemplified by the rigid end to end structure of the F1 pair). In contrast,only small constraints on the flexibility of domains were seen for the CO- and the LE-EGF-pairs.

The extended arrangements mentioned so far are formed by sequential arrangements of small globular domains in a single chain. In two types of domain, however, long linear structures are formed by several chains. These are the collagen triple helices formed by three chains with Gly-X-Y repeats, and the α-helical coiled-coil structures in which two to five chains with heptad repeats of nonpolar residues are connected (Cohen and Parry, 1990, 1994; Lupas et al., 1991). The length of these structures is highly variable. Coiled-coil structures in COMP and thrombospondin are not longer than 50 residues/chain (7.5 nm; Efimov et al., 1994). They are of similar size in tenascin (Spring et al., 1989) but longer, consisting of 600 residues (76 nm), in laminin (Beck et al., 1990). Collagen triple helices range from 45 residues/chain in the N-terminal small triple helix of collagen III (Bruckner et al., 1978) to 8 000 residues/chain (2400 nm) in some worm collagens (Gaill et al., 1991). Thus, an amazing diversity of forms of ECM proteins can be built up from the modular pool.

FUNCTIONAL PREDICTIONS FOR MODULES BASED ON PRIMARY SEQUENCE HOMOLOGY ARE OFTEN WRONG

Elucidation of the functions of individual domains in ECM proteins is a challenging but very time consuming and difficult task. An outstanding success was the identification of the cell binding site containing Arg-Gly-Asp in fibronectin. In the pioneering work by Ruoslahti (1988), this site was identified in the 10th F3 domain of fibronectin. An exposed and flexible three dimensional structure has been recently demonstrated for the Arg-Gly-Asp region both in this domain (Baron et al., 1991) and in disintegrins (Blobel and White, 1992). When it was found that cell attachment by several other ECM proteins could be inhibited by Arg-Gly-Asp peptides, it was initially thought that a universally valid principle had been discovered. As a consequence, many putative cell attachment sites were predicted from sequence data. It is now realized that this does not hold true: many cell attachment processes are Arg-Gly-Asp independent and many major cell attachment sites do not contain this sequence. It was even found that an F3 domain in tenascin, which contains an Arg-Gly-Asp sequence at a similar location as the classic domain in fibronectin is not involved in attachment, although in the isolated recombinantly prepared domains the tripeptide sequence was active (Aukhil et al., 1993). Instead, another domain mediates Arg-Gly-Asp independent cell binding in native tenascin (Spring et al., 1989). As it was pointed out by Ruoslahti (1988) and Hynes (1990), but ignored by many others, attachment is usually highly conformation dependent (Deutzmann et al., 1990) and, as for fibronectin (Hynes, 1990), more than one binding site may be involved.

Another example of inaccurate prediction of functions based on sequence similarity relates to the EGF domains. It is an appealing concept that some of the EGF-like domains in ECM proteins may act as localized signals for growth and differentiation, which may act in a specific and vectorial way on adjacent cells. Indeed, growth-promoting functions have been experimentally shown for laminin, thrombospondin and tenascin (Engel, 1989). For laminin, which is amongst the first ECM molecules expressed in mammalian embryonic development, it was possible to localize this function to fragment P1 (Panayotou et al., 1989) which comprises short-arm regions of the α, β and γ chains with about 25 EGF-like repeats in total. Unambigous proof is missing, however, for an EGF domain being the active functional site in the very large fragment P1. Furthermore, it is clear that not all EGF domains in ECM and other proteins exhibit growth factor-like functions. The best demonstrated function of the laminin type EGF (EG′) domains (Table 2) is to provide a very specific binding site for the C-terminal nidogen domain N3 (Mayer et al., 1993). The three-dimensional structure of the nidogen binding EGF-domain will

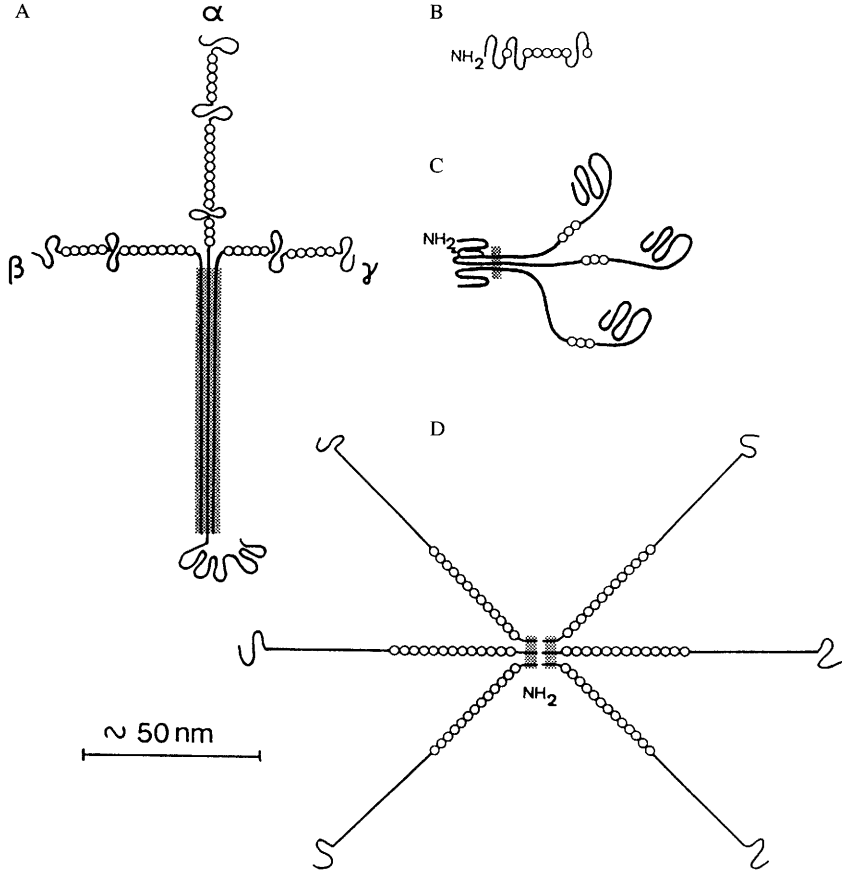

Fig. 1. Schematic representations of the shapes of (A) laminin, (B) nidogen/entactin, (C) thrombospondin and (D) tenascin, according to electron microscopic information. EGF-like domains are represented by open circles, hatched fields mark coiled-coil structures. The proteins are approximately drawn to scale (for references see text).

be known soon and it is hoped that details of its specific function will be explored. Other functions, like Ca^{2+} binding of specialized EGF domains (EG*), have been demonstrated (Table 2) and have been correlated with the three dimensional structure of the EG* domains (Baron et al., 1991). The functions of most other EGF-like domains remain unexplored.

Functional predictions that are entirely based on recognition of a general sequence motif are usually wrong. Very specific information like calcium-binding motifs in EGF- or EF-hand domains might be helpful but even in these cases there have been many disappointing experiences. We urge, therefore, that the frequently used term 'putative functional domains' should be avoided, since it can lead to confusion when 'putative' is inadvertantly omitted (eg. in the next review).

Another argument for the functional promiscuity of domains comes from estimates of the number of protein families. Recent genome sequencing efforts show that about one third of sequenced open reading frames belong to families that already have members in the databanks. From these data Chothia (1992) estimated that about 1500 different protein families exist. Even if there are ten times more families, because of biases in the databanks, the number of functions greatly excedes the number of basic protein structures. Thus the prototype of each family is modelled differently to fulfil specific functions. Extracellular modules provide examples of great functional variety being achieved from a few basic structures. Just as no one would claim to predict the antigen from the sequence of an antibody, we feel the elucidation of functions of protein modules should rely principally on experimental effort, not sequence comparisons.

DOMAINS MAY FUNCTION INDEPENDENTLY OR IN COMBINATION WITH OTHER DOMAINS

Perhaps the most important function of coiled-coil and collageneous domains is to connect subunits within a single molecule, in which they may exhibit a concerted function. This is clearly demonstrated by thrombospondins (Lawler 1986; Lawler et al., 1993) and COMP (Mörgelin et al., 1992) in which 3 or 5 identical chains are combined. These all point in the same direction and hence the C-terminal cell binding domains of these molecules and other domains are brought in close vicinity (Fig. 1). This alignment may be important for simultaneous recognition of multiple receptor sites at the cell surface. Although details of the binding mechanism have not yet been explored, the situation may be comparable to the binding of the hexameric 'flower bouquet'-shaped first component of complement C1q, that binds to clusters of IgG. In this example it was demonstrated that sufficient binding strength is only produced by multivalent binding (Tschopp et al., 1980). This affords a mechanism for discriminating between clustered and isolated IgG molecules at a cell surface.

In many collagens, several globular domains are combined by association of three chains in the collagen triple helix (for example, collagens IV, VI and XII; Table 1). Von Willebrand type A (VA) domains are involved in the self-assembly of some collagens and have frequently been designated as collagen-binding domains, although direct proof for this activity is missing in most cases (Colombatti and Bonaldo, 1991).

Laminin is comparable, in that three different chains α, β and γ are connected by a coiled-coil domain. Many genetically distinct variants of these chains have been found (Paulsson, 1993) and these are combined to give distinct laminin isoforms. Some isoforms are transiently expressed at restricted sites, suggesting specialized functions. The assembly of the three different chains is highly specific and correct assembly is crucial for cell binding of laminin by α6β1 integrin and for the promotion of neurite outgrowth (Hunter et al., 1992, Deutzmann et al., 1990; Sung et al., 1993).

THE TIME COURSE OF EVOLUTION OF ECM PROTEINS

Doolittle (1985, 1992) attempted to group proteins according to their time of invention. He classifies ECM proteins as very recent inventions, each of which is found in animals or plants but not in both, nor in prokaryotes. This suggests that ECM proteins arose around the time that plants and animals diverged, perhaps 1 billion years ago (Doolittle, 1985) It has been proposed that modern mosaic proteins are the result of efficient mechanisms of exon shuffling (Patthy, 1991b). For several ECM proteins for which sufficient sequences from phylogenetically distant organisms were available, phylogenetic trees were constructed. The construction of the dendrograms utilized the method of maximum parsimony, which determines the tree requiring the minimum number of base substitutions (alternative phylogenetic reconstruction methods are possible) As an example, the phylogenetic tree of the thrombospondin gene family is shown (Fig. 2; modified from Lawler et al., 1993, by addition and inclusion of COMP). The dendrogram is based on a comparison of the C-terminal six EF domains and the TC domain (Table 1). Lawler et al. (1993) were able to assign a very rough time scale to the dendrogram by calibration with two phylogenetic events. This is possible by

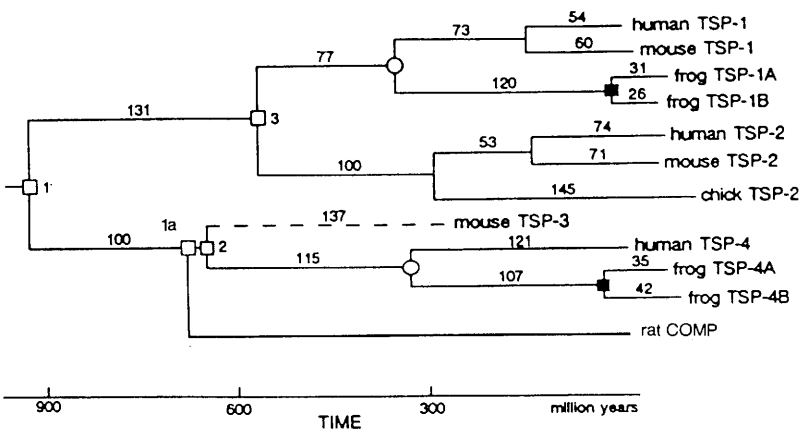

Fig. 2. Phylogenetic tree of the thrombospondin (TSP) gene family. Data for TSPs are according to Lawler et al. (1993) and the branch for cartilage oligomeric matrix protein (COMP) has been added on the basis of UPGMA distance matrix (Sneath and Sokal, 1973) and parsimony analyses (Felsenstein, 1993). The open squares mark the putative positions of gene duplications.

assuming a constant average rate of evolutionary divergence for the protein region under consideration. Different proteins change at very different rates but each rate is approximately constant (Doolittle, 1992); this rate constancy may apply to individual domains within mosaic proteins,but not for the entire protein. From Fig. 2, we can infer that the C-terminal portions of thrombospondins and COMP had a common ancestor earlier than 900 million years ago. At this time a gene duplication (open box 1) resulted in the two branches: one with a precursor for thrombospondins 1 and 2, and the other with a precursor for thrombospondin 3 and 4 plus COMP. One can also deduce from the phylogenetic tree, and the domain distribution in different branches, that either the N-terminal domains PR_3 EG of thrombospondins 1 and 2 were inserted into the thrombospondin 1/2 branch between 900 and 600 million years ago, or alternatively, they were already present in the common precursor and the thrombospondin type 3/4/COMP branch subsequently lost these domains. The resolution of such phylogenetic trees is insufficient to resolve whether COMP diverged from thrombospondins 3 and 4 before or after the branching off of thrombospondin type 3. However, an early branching of COMP would support a model that evolution proceeded in the direction of increasing domain complexity. It is important to note, however, that phylogenetic trees derived from certain domains of a multidomain protein are not able to predict the history of other domains in the same protein, hence it remains unclear whether the three EGF domains common to all proteins (Table 1) were present in the precursor. It will be interesting to scrutinize the phylogeny of different domains within one protein and hence gain more insight into the pattern and timing of domain acquisition within multidomain proteins.

MECHANISMS OF THE EVOLUTION OF ECM PROTEINS

New proteins come from old proteins as the result of gene duplications followed by base substitutions (Doolittle, 1992). This very general statement also applies to mosaic proteins. It is obvious from Table 1, however, that in their case individual domains can also be rearranged extensively, somewhat like mobile elements (Doolittle,1992).

Mechanisms of gene duplication by unequal crossing-over between sister chromosomes containing the genes are described in textbooks. Unequal cross-overs can also readily extend tandemly repeated genes into long series. Duplications, deletions, inversions, conversions, slippages and translocations of DNA segments can arise as the result of erratic rejoining of fragments. The genomic sequence of chromosome III of *C. elegans* (Wilson et al., 1994) and a comparative study of large DNA sequences of mouse and man (Koop and Hood, 1994) revealed an intriguing view on gene organization, with evidence for duplications, inversions and other gene rearrangements. Gene rearrangements are rare events catalyzed by the enzymes that mediate normal recombination processes; in the example of thrombospondins they resulted in gene duplications at time intervals in the range of 100 million years (see Fig. 2). For mosaic proteins it is generally believed that these processes are speeded up by the presence of introns. The most trivial reason for the higher speed of this process is the possibility of breaking and rejoining the DNA anywhere in the long introns on either side of an exon. Thus a large number of different possible breakages could lead to exon shuffling, in each case the exon is left intact, which in turn encodes a stable protein domain in the majority of extracellular modules.

Exon shuffling could involve transposable elements (which make up about 10% of the genome of higher eukaryotes); examples can be seen 'in flagranti' in several human genetic disorders. Duplication or deletion of exons are also the cause of several human genetic disorders (Bates and Lehrach, 1994, Makalowski et al., 1994). The possibility of an exon variation at the DNA level by reverse transcription of an alternative splice variant at the RNA level may also be considered. Reverse transcription may occur by mammalian enzymes with reverse transcriptase activity or by the help of virus systems (Fink, 1987).

Details of alternative possible mechanisms for exon shuffling have been discussed by Rogers (1990) and Patthy (1991b). Several examples strongly suggest that exons can be inserted into preexisting introns. However, domain shuffling is clearly not the result of just one mechanism, nor is it the only process operating in multidomain protein evolution. This becomes evident with the increasing number of observations in which exons do not correspond to protein domains, or in which domains consist of several exons (e.g. F3 domains are encoded by two exons, yet no half F3s has been found to date). Even harder to reconcile with a dominant role for exon shuffling is the observation that exon-intron boundaries, within the same domain organization, can differ from one species to another. Extensive rearrangements following the presumed duplication of a common primordial gene were shown for the genes for β and γ chains of laminin (Kallunki et al., 1991).

An example in which the relative contributions of exon shuffling and other processes were compared relates to the EF-hand calcium modulated proteins. Extensive analysis revealed a random distribution of introns over 'domain' and 'interdomain' space, and that some introns were acquired after a four domain precursor was formed. It was therefore concluded that in the evolution of the widely distributed EF-hand protein family, exon shuffling played little if any role (Kretsinger and Nakayama, 1993).

The evolution of genes for the modular ECM proteins may have been further complicated by horizontal gene transfer; for example, Bork and Doolittle (1992) suggested that bacteria may have acquired a F3 domain from animals.

WHAT WERE THE SELECTION PRESSURES AFFECTING THE EVOLUTION OF ECM PROTEINS?

Unequal crossing-over may lead to either an increase or a decrease in the number of repeated domains. This suggests that the large number of repeated domains present in ECM proteins may be the result of natural selection. One reason for the large number of domains in an ECM protein may be the need to prevent diffusion of domains with specific activities into the otherwise open extracellular space: this could be achieved simply by making the proteins very large. For example, this allows localization of domains with cell signalling activity at specific sites, and the variability of the extracellular environment by time-dependent and vectorial expressions of different proteins (Engel, 1989). Another common feature of the large

and extended ECM proteins is their ability to bridge between distant sites, for example between cellular receptors and other parts of the matrix. Clearly the development of specialized assembly domains was a prerequisite to develop multifunctional large molecules with the potential of forming higher macromolecular organisations. The α-helical coiled-coil domains are also found in many cytoskeletal proteins but collagen triple helices are specific for extracellular proteins. In addition to other functions, collagen triple helices are essential for the formation of collagen fibres, cuticle structures and networks. They contribute essentially to the mechanical properties of tissues of larger organisms.

Of course, ECM proteins contain a large number of domains with much more specific functions than spacing, assembly or support; a few are listed in Table 2. In addition, it must be stressed that such specific functions have been elucidated for only a small percentage of the known domains.

This rather simplified interpretation of selection pressures is apparently contradictory to the complete lack of phenotype resultant from genetic elimination of certain ECM proteins (even those implicated in important functions). In contrast to fibronectin, which is absolutely required in early stages of embryonic development, no phenotype was detectable in transgenic mice after knock-outs of tenascin and S-type lectin (George et al., 1993, Poirier and Robertson, 1993, Saga et al., 1992). One explanation may be that hitherto unrecognized subtle functions of these proteins may cause a small increase of fitness, which is not obvious in the phenotype, and would only be apparent in the appropriate population size and natural environment (as seen for transgenic mice lacking metallothioneins; Michalska and Choo, 1993). Alternatively, there may be functional redundancies between some ECM proteins with similar domains, which can at least in part fulfil the function of the deficient protein. Even in this case, however, selective advantages may be necessary to maintain such a redundancy in a population. These could either be selection for a subtle divergent function, an increased fidelity for a certain process or an enhanced efficiency of a cumulative function.

REFERENCES

Agraves, W. S., Tran, H., Burgess, W. H. and Dickerson, K. (1990). Fibulin is an extracellular matrix and plasma glycoprotein with repeated domain structure. *J. Cell Biol.* **111**, 3155-3164.

Aukhil, I., Joshi, P., Yan, Y. and Erickson, H. P. (1993). Cell- and heparin-binding domains of the hexabrachion arm identified by tenascin expression proteins. *J. Biol. Chem.* **268**, 2542-253.

Barlow, P. N., Steinkasserer, A., Norman, D. G., Kieffer, B., Wiles, A. P., Sim, R. B. and Campbell, I. D. (1993). Solution structure of a pair of complement modules by nuclear magnetic resonance. *J. Mol. Biol.* **232**, 268-284.

Baron, M., Nordman, D. G. and Campbell, I. D. (1991). Protein modules. *Trends Biochem.* **16**, 13-17.

Bates, G. and Lehrach, H. (1994). Trinucleotide repeat expansions and human genetic disease. *Bioessays* **16**, 277-284

Beck, K., Hunter I. and Engel, J. (1990). Structure and function of laminin: Anatomy of a multidomain glycoprotein. *FASEB J.* **4**, 148-160.

Blobel, C. P. and White, J. M. (1992). Structure, function and evolutionary relationship of proteins containing a disintegrin domain. *Curr. Opin. Cell Biol.* **4**, 760-765.

Bode, W. and Huber, R. (1992). Natural protein proteinase inhibitors and their interaction with proteinases. *Eur. J. Biochem.* **204**, 433-451.

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett.* **286**, 47-54.

Bork, P. (1992). The modular architecture of vertebrate collagens. *FEBS Lett.* **307**, 49-54.

Bork, P. and Doolittle, R. F. (1992). Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl. Acad. Sci. USA* **89**, 8990-8994.

Bruckner, P., Bächinger, H. P., Timpl, R. and Engel, J. (1978). Three conformationally distinct domains in the amino-terminal segment of type III procollagen and its rapid triple helix = coil transition. *Eur. J. Biochem.* **90**, 595-603.

Burgeson, R. E., Chiquet, M., Deutzmann, R., Ekblom, P., Engel, J., Kleinman, H., Martin, G. R., Meneguzzi, G., Paulsson, M., Sanes, J., Timpl, R. Tryggvason, K., Yamada, Y. and Yurchenco, P. D. (1994). A new nomenclature for the laminins. *Matrix Biol.*, in press.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature* **357**, 543-544

Cohen, C. and Parry, D. A. D. (1990). α-helical coiled- coils and bundles: How to design an α-helical protein. *Proteins* **7**, 1-15.

Cohen, C. and Parry, D. A. D. (1994). Alpha-helical coiled coils: more facts and better predictions. *Science* **263**, 488-489.

Colombatti, A. and Bonaldo, P. (1991). The superfamily of proteins with von Willebrand factor type A-like domains: one theme common to components of extracellular matrix. *Blood* **77**, 2305-2315.

Constantine, K. L., Madrid, M., Banyai, L., Trexler, M., Patthy, L. and Llinas, 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.

Corson, G. M., Chalberg, S. C., Dietz, H. C., Charbonneau, N. L. and Sakai, L. Y. (1993). Fibrillin binds calcium and is coded by cDNAs that reveal a multidomain structure and alternatively spliced exons at the 5′ end. *Genomics* **17**, 476-484.

Davis, C. G. (1990). The many faces of epidermal growth factor repeats. *New Biologist* **2**, 410-419.

Deutzmann, R., Aumailley, M., Wiedemann, H., Pysny, W., Timpl, R. and Edgar, D. (1990). Cell adhesion spreading and neurite stimulation by laminin fragment E8 depend on maintenance of secondary and tertiary structure in its rod and globular domain. *Eur. J. Biochem.* **191**, 513-522.

Doolittle, R. F. (1985). The geneology of some recently evolved vertebrate proteins. *Trends Biochem.* **10**, 233-237.

Doolittle, R. F. (1992). Reconstructing history with amino acid sequences. *Protein Science* **1**, 192-200.

Doolittle, R. F., Feng, D. F., Johnson, M. S. and McClure, M. A. (1986). Relationship of human protein sequence to those of other organisms. *Cold Spring Harbor Symp Quant. Biol.* **51**, 447-455.

Efimov, V. P., Lustig, A. and Engel, J. (1994). The thrombospondin-like chains of cartilage oligomeric matrix protein are assembled by a five-stranded α-helical bundle between residues 20 and 83. *FEBS Lett.* **341**, 54-58.

Engel, J. (1989). EGF-like domains in extracellular matrix proteins: localized signals for growth and differentiation? *FEBS Lett.* **251**, 1-7.

Engel, J. (1991). Common structural motifs in proteins of the extracellular matrix. *Curr. Opin. Cell Biol.* **3**, 779-785.

Engel, J. (1994). Electron microscopy of extracellular matrix components. *Meth. Enzymol.* (in press).

Engel, J., Odermatt, E., Engel, A., Madri, J. A., Furthmayr, H., Rohde, H. and Timpl, R. (1981). Shapes, domain organizations and flexibility of laminin and extracellular matrix. *J. Mol. Biol.* **150**, 97-120.

Engel, J., Taylor, W., Paulsson, M., Sage, H. and Hogan B. (1987). Calcium binding domains and calcium induced conformational transition of SPARC (osteonectin, BM-40), an extracellular glycoprotein expressed in mineralized and non mineralized tissues. *Biochemistry* **26**, 6958-6965.

Erickson, H. P. (1993). Tenascin C, tenascin R and tenascin Y: a family of talented proteins in search of functions. *Curr. Opin. Cell Biol.* **5**, 869-876.

Felsenstein, J. (1993). PHYLIP (Phylogeny Interference package) Version 3. 5 (Computer software package distributed by the author, Dept. Genetics, University of Washington, Seattle.

Fink, G. R. (1987). Pseudogenes in yeast ? *Cell* **49**, 5-6.

Gaill, F., Wiedemann, H., Mann, K., Kühn, K., Timpl, R. and Engel, J. (1991). Molecular characterization of cuticle and interstitial collagens from worms collected at deep sea hydrothermal vents. *J. Mol. Biol.* **221**, 209-223.

George, E., Georges-Labouesse, E., Patel-King, R., Rayburn, H. and Hynes, R. (1993). Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin. *Development* **119**, 1079-1091.

Graves, B. J., Crowther, R. L., Chandran, C., Rumberger, J. M., Li, S., Huang, K.-S., Pesky, D. H., Familletti, P. C., Wolitzky, B. A. and Burns,

D. K. (1994). Insight into E-selectin/ligand interaction from the crystal structure and mutagenesis of the lec/EGF domains. *Nature* **367**, 532-538.

Harbury, P. B., Zhang, T., Kim, P. S. and Alber, T. (1993). A switch between two-, three- and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401-1407.

Huber, R., Scholze, H., Paques, E. P. and Deisenhofer, J. (1980). Crystal structure analysis and molecular model of human C3a anaphylatoxin. *Hoppe-Seylers Z. Physiol. Chem.* **361**, 1389-1399.

Hunter, I., Schulthess, T. and Engel, J. (1992). Lamininin chain assembly by triple and double stranded coiled-coil structures. *J. Biol. Chem.* **267**, 6006-6011.

Hynes, R. O. (1990). *Fibronectins* (ed. A. Rich), New York/Berlin: Springer Verlag.

Ishii, N., Wadsworth, W. G., Stern, B. D., Culotti, J. G. and Hedgecock, E. M. (1992). UNC-6, a laminin-related protein guides cell and pioneers axon migration in C. elegans. *Neuron* **9**, 873-881.

Joshi, P., Chung C. Y., Aukhil, I. and Erickson, H. P. (1993). Endothelial cells adhere to the RGD domain and the fibrinogen-like terminal knob of tenascin. *J. Cell. Sci.* **106**, 389-400.

Kallunki, P. and Tryggvason, K. (1991). Human basement membrane sulfate proteoglycan core protein: A 467 kD protein containing multiple domains resembling elements of the low density lipoprotein receptor, laminin, neural cell adhesion molecules and epidermal growth factor. *J. Cell Biol.* **116**, 559-571.

Kallunki, P., Sainio, K., Eddy, R., Byers, M., Kallunki, T., Sariola, H., Beck, K., Hirvonen, H., Shows, T. B. and Tryggvason, K. (1992). A truncated laminin chain homologous to the B2 chain: structure, spatial expression, and chromosomal assignment. *J. Cell Biol.* **119**, 679-693.

Kallunki, T., Ikonen, J., Chow, L. T., Kallunki, P. and Tryggvason, K. (1991). Structure of the human laminin B2 chain reveals extensive divergence from the laminin B1 chain gene. *J. Biol. Chem.* **266**, 221-228.

Koop, B. and Hood, L. (1994). Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genetics* **7**, 48-53.

Kreis, T. and Vale, R. (1993). *Guidebook to the Extracellular Matrix and Adhesion Proteins.* Oxford: Oxford University Press.

Kretsinger, R. H. (1987). Calcium coordination and the calmodulin fold: divergent versus convergent evolution. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 499-510.

Kretsinger, R. H. and Nakayama, S. (1993). Evolution of EF-hand calcium-modulated proteins. IV. Exon shuffling did not determine the domain composition of EF-hand proteins. *J. Mol. Evol.* **36**, 477-488.

Lawler, J. (1986). The structural and functional properties of thrombospondin. *Blood* **67**, 1197-1209.

Lawler, J., Duquette, M., Urry, L., McHenry, K. and Smith, T. F. (1993). The evolution of the thrombospondin gene family. *J. Mol. Evol.* **36**, 509-516.

Leahy, D. J., Hendrickson, W. A., Aukhil, I. and Erickson, H. P. (1992). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science* **258**, 987-991.

Lupas, A., Van-Dyke, M. and Stock J. (1991). Predicting coiled coils from protein sequence. *Science* **252**, 1162.

Makalowski, W., Mitchell, G., and Labuda, D. (1994). Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* **10**, 188-193.

Mann, K., Deutzmann, R., Aumailley, M., Timpl, R., Raimondi, L., Yamada, Y., Pan, T., Conway, D. and Chu, M.-L. (1989). Amino acid sequence of mouse nidogen, a multidomain basement membrane protein with binding activity for laminin, collagen IV and cells. *EMBO J.* **8**, 65-72.

Markovic, Z., Lustig, A., Engel, J., Richter, H. and Hörmann, H. (1983). Shape and stability of fibronectin in solutions of different pH and ionic strength. *Hoppe-Seyler's Z. Physiol. Chem.* **364**, 1795-1804.

Mayer, U., Nischt, R., Poschl, E., Mann, K., Fukuda, K., Gerl, M., Yamada, Y. and Timpl, R. (1993). A single EGF-like motif of laminin is responsible for high affinity nidogen binding. *EMBO J.* **12**, 1879-1885.

McMahan, U. J., Horton, S. E., Werle, M. J., Honig, L. S., Kröger, S., Ruegg, M. A. and Escher, G. (1992). Agrin isoforms and their role in synaptogenesis. *Curr. Opin. Cell Biol.* **4**, 869-874.

Michalska, A. E. and Choo, K. H. A. (1993). Targeting and germ-line transmission of a null mutation at the metallothionein I and II loci in mouse. *Proc. Natl. Acad. Sci. USA* **90**, 8088-8092.

Moergelin, M., Heinegard, D., Engel, J. and Paulsson, M. (1992). Electron microscopy of native COMP (cartilage oligomeric matrix protein) purified from the swarm rat chondrosarcoma reveals a five-armed structure. *J. Biol. Chem.* **2367**, 6137-6141.

Moergelin, M., Heinegard, D., Engel, J. and Paulsson, M. (1994). The cartilage proteoglycan aggregate: Assembly through combined protein-carbohydrate and protein-protein interactions. *Biophys. Chem.* (in press).

Noonan, D. M. and Hassel, J. R. (1993). Proteoglycans of basement membranes. In *Molecular and Cellular Aspects of Basement Membranes* (ed. Rohrbach, D. H. and Timpl, R.) pp. 189-210, New York/London: Academic Press.

Odermatt, E., Engel, J., Richter, H. and Hörmann, H. (1982). Shape, conformation and stability of fibronectin fragments determined by electron microscopy, circular dichroism and ultracentrifugation. *J. Mol. Biol.* **159**, 109-123.

Oldberg, A, Antonsson, P., Lindblom, K. and Heinegard, D. (1992). COMP (cartilage oligomeric matrix protein) is structurally related to the thrombospondins. *J. Biol. Chem.* **267**, 22346-22350.

Ott, U., Odermatt, E., Engel, J., Furthmayr, H. and Timpl, R. (1982). Protease resistance and conformation of laminin. *Eur. J. Biochem.* **123**, 63-72.

Pan, T.-C., Kluge, M., Zhang, R.-Z., Mayer, U., Timpl, R. and Chu, M.-L. (1993). Sequence of extracellular mouse protein BM-90/fibulin and its calcium dependent binding to other basement-membrane ligands. *Eur. J. Biochem.* **215**, 733-740.

Panayotou, G., End, P., Aumailley, M. Timpl, R. and Engel, J. (1989). Domains of laminin with growth-factor activity. *Cell* **56**, 93-101.

Patthy, L. (1991a). Exons – original building blocks of proteins? *Bioessays* **13**, 187-192.

Patthy, L. (1991b). Modular exchange principles in proteins. *Curr. Opin. Struct. Biol.* **1**, 351-361.

Paulsson, M. (1993). Laminin and collagen IV variants and heterogeneity in basement membrane composition. In *Molecular and Cellular Aspects of Basement Membranes* (ed. Rohrbach, D. H. and Timpl, R.) pp. 177-185, New York/London: Academic Press.

Poirier, F. and Robertson, E. (1993). Normal development of mice carrying a null mutation in the gene encoding the L14 S-type lectin. *Development* **119**, 1229-1236.

Pottgiesser, J., Maurer, P., Mayer, U., Nischt, R., Mann, K., Timpl, R., Krieg, T. and Engel J. (1994). Changes in calcium and collagen IV binding caused by mutations in the EF hand and other domains of extracellular matrix protein BM-40 (SPARC, osteonectin). *J. Mol. Biol.* (in press).

Rogers, J. H. (1990). The role of introns in evolution. *FEBS Lett.* **268**, 339-343.

Ruoslahti, E. (1988). Fibronectin and its receptors. *Ann. Rev. Biochem.* **57**, 375-413.

Saga, Y., Yagi, T., Ikawa, Y., Sakakura, T., and Aizawa, S. (1992). Mice develop normally without tenascin. *Genes Dev.* **6**, 1821-1831.

Sealock, R. and Froehner, S. C. (1994). Dystrophin-associated proteins and synapse formation: is a-dystroglycan the agrin receptor? *Cell* **77**, 617-619.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy.* San Francisco: Freemann.

Spring, J., Beck, K. and Chiquet-Ehrismann, R. (1989). Two contrary functions of tenascin: Dissection of the active sites by recombinant tenascin fragments. *Cell* **59**, 325-334.

Sung, U., O'Rear, J. J. and Yurchenco, P. D. (1993). Cell and heparin binding in the distal long arm of laminin: identification of active and cryptic sites with recombinant and hybrid glycoprotein. *J. Cell. Biol.* **123**, 1255-1268.

Tschopp, J., Schulthess, T., Engel, J. and Jaton, J.-C. (1980). Antigen-independent activation of the first component of complement C1 by chemically cross-linked rabbit IgG-oligomers. *FEBS Lett.* **112**, 152-154.

Weis, W. I., Drickamer, K. and Hendrickson, W. A. (1992). Structure of a C-type mannose-binding protein complexed with an oligosaccharide. *Nature* **360**, 127-134.

Williams, M. J., Phan, I., Baron, M., Driscoll, P. C. and Campbell, I. D. (1993). Secondary structure of a pair of fibronectin type 1 modules by two-dimensional nuclear magnetic resonance. *Biochemistry* **32**, 7388-7395.

Wilson, R. et al. (1994). 2. 2 MB of contiguous nucleotide sequence from chromosome III of *C. elegans. Nature* **368**, 32-38.

Winograd, E., Hume, D. and Branton, D. (1991). Phasing the conformational unit of spectrin. *Proc. Natl. Acad. Sci. USA* **88**, 10788-10791.

Yamagata, M., Yamada, K. M., Yamada, S. S., Shinomura, T., Tanaka, H., Nishida, Y., Obara, M. and Kimata, K. (1991). The complete primary structure of type XII collagen shows a chimeric molecule with reiterated fibronectin type III motifs, von Willebrand factor A motifs, a domain homologous to a non-collagenous region of type IX collagen and short collagenous domains with an Arg-Gly-Asp site. *J. Cell Biol.* **115**, 209-221.