The extracellular matrix is formed by complex and intricate networks within which molecules are precisely organized. These molecular networks determine the specific histoarchitecture of tissues and provide cells with information and a scaffold. Most of the structural extracellular matrix molecules — collagens, noncollagenous glycoproteins, and proteoglycans — are chimeric and share common domains. Studies of the interactions between extracellular matrix molecules and mapping of the interaction sites to defined structural modules have led to the concept that the function of the extracellular matrix relies largely in the polymers that they form. Furthermore, determination of the tertiary structure of protein motifs involved either in the assembly of the various molecules into polymers or in cell–extracellular matrix interactions has recently opened the field of structural biology of the extracellular matrix.

Key words Collagens · Noncollagenous glycoproteins · Integrins · Cell-matrix interactions

Abbreviations COL Collagenous · ECM Extracellular matrix · EGF Epidermal growth factor · FACIT Fibril-associated collagen with interrupted triple helix · Fn Fibronectin · NC Noncollagenous · vWF-A Von Willebrand factor A domain

Introduction

The extracellular matrix (ECM) is constituted by diverse composite structures which determine the histoarchitecture specific to every organ and provide cells with biological information and a mechanical scaffold for adhesion and migration. Elucidation of the molecular composition of these complex structures began slowly with the discovery, characterization, and purification of the constituent ECM components extracted from cells or tissues, and is nowadays developing much more rapidly by using molecular biology approaches. The latter allows the identification and generation by recombinant technology of full-length or partial polypeptides corresponding to proteins which have, or have not, been purified from tissues, thus allowing further studies of their functions. It has also revealed that most of the ECM molecules are chimeric and share similar structural domains which made the traditional grouping of matrix molecules into collagens, struc-
tural glycoproteins, and proteoglycans difficult. Indeed some collagens, such as collagens IX, XII, and XIV, contain glycosaminoglycan chains or very large noncollagenous domains representing up to 90% of their molecular mass. In addition, domains typical for ECM components, such as the collagen triple helix, are present in molecules which do not belong to the ECM, such as the macrophage scavenger receptor, the acetylcholinesterase receptor, and the C1q component of complement. Furthermore, it has recently been shown that a collagen molecule, collagen XVII, contains a transmembrane domain and is inserted in the cell plasma membrane.

Purification of the authentic or recombinant molecules has opened the way for studies of interactions and structure-function relationships, from which the concept has evolved that function of the ECM relies not in the necessary single molecules but rather in the polymers that they form. This has been strengthened by investigations of material from patients with inherited diseases and from transgenic mice in which a genetic defect in any one of the components of a given polymer leads to similar clinical symptoms. The specific interactions between protein motifs leading to the assembly of the different molecules into homotypic and heterotypic polymers or involved in cell-ECM interactions have been precisely located. When small enough, the protein motifs have been subjected to structural analysis by magnetic resonance or crystallography and the precise spatial orientation of reactive sites resolved.

Because of the extensive literature on the subject only some of the components of the ECM are described here. The review is limited to molecules for which comprehensive concepts have been developed regarding their organization into supramolecular complexes and biological functions.

The constitutive molecules of the ECM

Collagens, the most abundant structural components of the ECM, are homo- or heterotrimeric molecules whose subunits, the α chains, are distinct gene products. Up to now 34 different α chains differing in the primary sequence have been cloned and sequenced (Table 1), several of them being differentially spliced, which adds to the diversity of the collagen family [1–3]. The sequence of the various α chains contains a variable number of the classical Gly-X-Y repetitive motifs which form the so-called collagenous (COL) domains and noncollagenous portions (NC domains) of variable length and location. The COL portion(s) of three homologous or heterologous α chains are folded together into a helix with a coiled-coil conformation which constitutes the characteristic basic structural motif of collagens. Up to now 19 different combinations of the α chains (collagen types I–XIX) have been identified or predicted to exist in the superfamily of vertebrate collagens (Table 1).

A characteristic property of collagens is to form highly organized polymers. On this basis the different collagen types can be grouped roughly into two main classes of molecules, the fibril-forming collagens (collagens I, II, III, V, and XI), which constitute a rather homogeneous subgroup, and the heterogeneous class of nonfibrillar collagens. In addition to their intrinsic property to organize into distinct supramolecular assemblies, a major difference between the two subgroups is that for the fibril-forming collagens almost the entire length of the molecules (approx. 300 nm) is constituted by a single COL domain, while nonfibrillar collagen molecules contain one or several COL and NC domains of variable size according to the different collagen types (Fig. 1). On this basis several subfamilies can be distinguished, such as network-forming collagens (basement membrane collagens IV and collagens VIII and X), microfibrillar collagen VI, fibril-associated collagen with interrupted triple helix (FACIT) collagens (IX, XII, XIV, XVI, XIX), and the recently named multiplexins (multiple triple-helix domain and interruptions) XV and XVIII.

Most of these collagens have a specific distribution. Fibril-forming collagens, collagen VI, and FACIT, are expressed in interstitial connective tissue. Collagen VI is a heterotrimer with the chain composition α1(VI)α2(VI)α3(VI). The dumbbell-shaped molecule has a short triple helical domain of 105 nm flanked by NC globular regions contributing two-thirds of the molecular mass [1]. These domains consist of repetitive motifs with homology to the A domain of von Willebrand factor (vWF-A),

### Table 1 The collagen family and subfamilies

<table>
<thead>
<tr>
<th>Type</th>
<th>Chains</th>
<th>Molecular composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibril-forming collagens</td>
<td>α1(I), α2(I)</td>
<td>[α1(I)]2, [α2(I)]2</td>
</tr>
<tr>
<td></td>
<td>α1(II)</td>
<td>[α1(II)]3</td>
</tr>
<tr>
<td></td>
<td>α1(III)</td>
<td>[α1(III)]3</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>[α1(V)α2(V)], [α1(V)α2(V)]β, [α1(V)α2(V)α3(V)]β</td>
</tr>
<tr>
<td></td>
<td>XI</td>
<td>[α1(XI)α2(XI)]β, [α1(XI)α2(XI)]β, [α1(XI)α2(XI)]β, [α1(XI)α2(XI)]β</td>
</tr>
<tr>
<td>Basement membrane collagens</td>
<td>α1(IV), α2(IV)</td>
<td>[α1(IV)]2, [α2(IV)]2</td>
</tr>
<tr>
<td></td>
<td>α3(IV)</td>
<td>?</td>
</tr>
<tr>
<td>FACIT</td>
<td>α1(IX), α2(IX), α3(IX)</td>
<td>[α1(IX)α2(IX)α3(IX)]β</td>
</tr>
<tr>
<td></td>
<td>α1(XII)</td>
<td>[α1(XII)]3</td>
</tr>
<tr>
<td></td>
<td>α1(XIV)</td>
<td>[α1(XIV)]3</td>
</tr>
<tr>
<td></td>
<td>α1(XVI)</td>
<td>[α1(XVI)]3</td>
</tr>
<tr>
<td></td>
<td>α1(XIX)</td>
<td>?</td>
</tr>
<tr>
<td>Multiplexins</td>
<td>α1(XV)</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>α1(XVIII)</td>
<td>?</td>
</tr>
<tr>
<td>Orphans</td>
<td>α1(VI), α2(VI), α3(VI)</td>
<td>[α1(VI)α2(VI)α3(VI)]β</td>
</tr>
<tr>
<td></td>
<td>α1(VII)</td>
<td>[α1(VII)]3</td>
</tr>
<tr>
<td></td>
<td>α1(VIII), α2(VIII)</td>
<td>[α1(VIII)]3, [α1(VIII)]β</td>
</tr>
<tr>
<td></td>
<td>α1(X)</td>
<td>[α1(X)]3</td>
</tr>
<tr>
<td></td>
<td>α1(XII)</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>α1(XVII)</td>
<td>[α1(XVII)]3</td>
</tr>
</tbody>
</table>

...
with two motifs at the C-terminus of all three chains and one, one, and nine motifs at the N-terminus of \( \alpha_1 \)(VI), \( \alpha_2 \)(VI) and \( \alpha_3 \)(VI) chains, respectively. In addition, the \( \alpha_1 \)(VI) sequence presents one module with homology to the fibronectin (Fn) type III repeats, and one Kunitz-type protease-inhibitor module [4, 5].

In sharp contrast to the fibril-forming collagens, FACIT (collagens IX, XII, XIV, XVI, and XIX) are multimeric molecules formed by two or more very short helical rods interspaced with small nontriplet helical domains and by a large aminoterminal NC region. For collagen XII and XIV the NC portion of the \( \alpha \) chain consists of several single or successive FnIII motifs interspaced by vWF-A motifs and forms 60-nm extended arms. The aminoterminal domain of collagen IX is smaller, and a glycosaminoglycan is branched on the \( \alpha_2 \)(IX) chain [2].

The most abundant structural component of basement membranes is collagen IV. Six genetically different \( \alpha \) chains, \( \alpha_1 \)(IV), \( \alpha_2 \)(IV), \( \alpha_3 \)(IV), \( \alpha_4 \)(IV), \( \alpha_5 \)(IV), and \( \alpha_6 \)(IV), have been cloned and sequenced [1]. Only collagen IV with the composition \([\alpha_1 \text{(IV)}]_2[\alpha_2 \text{(IV)}]_2\) has been purified, and other combinations between \( \alpha \) chains are not definitively established. Sequencing of the various cDNAs predicts, however, that organization of the various molecules should be similar to that described for \([\alpha_1 \text{(IV)}]_2[\alpha_2 \text{(IV)}]_2\), which is referred to as collagen IV here. The molecule contains a carboxyterminal NC domain, NC1, and a relatively long triple-helical domain of 400 nm with several small interruptions allowing for a rather flexible helical rod. Two other collagens, VII and XVII, are associated with specialized structures of basement membranes underlying stratified epithelia, the anchoring fibrils and the hemidesmosomes, respectively. Collagen VII is a homotrimer \([\alpha_1 \text{(VII)}]_3\) and its triple helix is the longest, 450 nm, described for vertebrate collagens [6]. The amino acid sequence of the COL portion presents several imperfections in the repetition of the Gly-X-Y repeats. The sequence predicted for the NC aminoterminus of the \( \alpha_1 \)(VII) chain contains two vWF-A motifs, ten FnIII repeats, and a stretch of amino acid homologous to cartilage matrix protein. This region is separately folded into an arm of 36 nm for each individual \( \alpha_1 \)(VII) chain. An additional small globular domain, NC2, is present at the carboxyterminus of the precursor molecule and is involved in the assembly of collagen VII molecules (see below).

For many years antigens involved in cutaneous blistering autoimmune diseases, including the bullous pemphigoid antigen of 180 kDa, have been shown to be present in hemidesmosomes, which are specialized structures of the epidermal-dermal junction securing the epidermis to the upper layer of the basement membrane. Sequencing of the corresponding cDNA has shown this antigen to be a very special collagen, collagen XVII [7, 8]. The deduced amino acid sequence predicts a transmembrane protein with a type II orientation, i.e., the carboxyterminal portion being extracellular. The predicted carboxyterminal sequence of the \( \alpha_1 \)(XVII) chain is constituted by 13–15 distinct NC segments interspaced by small NC sequences and forming the ectodomain. Adjacent to this portion there are eight heptad repeats, which could be involved in the formation of coiled-coil structures, followed by a stretch of amino acids compatible for a transmembrane association. The aminoterminal intracellular region of the \( \alpha_1 \)(XVII) chain contains four 24-residue repeats of unknown function. Rotary shadowing images of collagen XVII indicate folding of the aminoterminal domains into globules and of the ectodomain into a 70-nm rod terminated by a flexible tail [9].

The multiplexins (collagens XV and XVIII) are also localized to basement membranes. These are characterized by a central and interrupted COL domain flanked by relatively large amino and carboxy NC regions [9].

Several classes of proteoglycans with different sizes and protein cores are present in the ECM. Proteoglycans can be considered to constitute a distinct subset of NC glycoproteins containing glycosaminoglycan side chains. Sequencing of many core protein cDNAs and analysis of the glycosaminoglycan side chains has, however, revealed an extreme diversity in the proteoglycan family and complicated the assignment of some molecules into this family. Indeed, based on the presence of homologies in the protein cores, molecules such as the link protein or TSG-6, which are devoid of glycosaminoglycan side chains, belong to the proteoglycan family [10]. Other members of the family, including decorin or biglycan, are expressed as isoforms either with or without glycosaminoglycan chains [11]. Based on typical features in the sequence of the protein core a simplified classification into two subfamilies has recently been proposed [12] and is used here (Table 2). The classification distinguishes the small leucine-rich proteoglycans and the modular proteoglycans with two subgroups, the nonhyalurano-binding and the hyaluronan- and lectin-binding proteoglycans or hyalectans.

The small leucine-rich proteoglycan group comprises decorin, biglycan, fibromodulin, lumican, and epiphycan. They have a small protein core (36–42 kDa) with an aminoterminal domain for glycosaminoglycan side chains branching (one to four chains depending on the member), a central region made by the leucine-rich repeats and flanked by cysteine clusters, and a carboxyterminus...
domain. The protein core of decorin, lumican, or fibromodulin is compact and horseshoe-shaped, a conformation suitable for protein-protein interactions [13]. The hyalectan subgroup comprises versican, aggrecan, neurocan, and brevican. They have a large protein core (100–370 kDa) consisting of a central region for glycosaminoglycan chain branching, a carboxyterminal region with epidermal growth factor (EGF)-like repeats, stretches of amino acids homologous to vertebrate lectins or to complement regulatory protein, and a hyaluronan binding domain at their aminoterminus [12]. Two members of this subfamily, versican and CD44, have a widespread distribution, while aggrecan is expressed mostly in cartilage, and neurocan or brevican are brain tissue specific.

Recently the core protein of a chondroitin sulfate proteoglycans, bamacan, has been characterized [17]. The polypeptide is predicted to be folded into a central rod of potential coiled-coil conformation, with globular domains at both ends.

In addition to collagens and proteoglycans, many N-glycoproteins are building blocks of the ECM. One of the best studied is Fn [18]. The molecule is a dimer, the subunits of which (250–280 kDa) are disulfide linked at the carboxyterminus and form arms of 50 nm. The amino acid sequence of Fn presents three types of repetitive motifs: 12 motifs of type I homology (FnI), 2 motifs of type II homology (FnII), and 15–17 motifs of type III homology (FnIII; Fig. 2). The latter contribute up to half of the molecule and are present in variable number due to complex alternative splicing of a single Fn gene, giving rise to at least 20 variants in human. The tertiary structure of several Fn repeats has been determined by magnetic resonance spectroscopy or by crystallography. In particular, type III modules consist of two layers of β sheets, one with three antiparallel strands, and the other with four antiparallel strands, enclosing a hydrophobic core, an arrangement similar to that of immunoglobulin folds [19, 20]. The data indicate that the different modules are independently folded and arranged as pearls on a necklace. Most importantly, despite the complex folding into β sheets, the structure can undergo conformational changes by reversible unfolding of the FnIII modules [21, 22].

Other components rich in FnIII modules are the tenascins, which, however, are not strictly categorized as structural components of the ECM. Tenasin appears early during embryogenesis, and its synthesis is switched off in mature tissue. However, it reappears during wound healing and in some stromal tumors [23, 24]. Tenascins are formed by the association of three subunits, folded together at their aminoterminus into a triple-stranded kDa protein core consists of nine aminoterinal Kazal-type protease inhibitor motifs, two and four EGF-like repeats, the latter interspaced by three lamininlike G domains.

Table 2 Structural glycoproteins and proteoglycans of the ECM

<table>
<thead>
<tr>
<th>Glycoproteins</th>
<th>Proteoglycans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interstitial connective tissue</td>
<td>Small leucine-rich proteoglycans</td>
</tr>
<tr>
<td>Fibronectins</td>
<td>Decorin</td>
</tr>
<tr>
<td>Tenascins</td>
<td>Biglycan</td>
</tr>
<tr>
<td>Fibromodulin</td>
<td>Lumican</td>
</tr>
<tr>
<td>Elastin</td>
<td>Epiphycan</td>
</tr>
<tr>
<td>Basement membranes</td>
<td>Modular proteoglycans: nonhyaluronan binding</td>
</tr>
<tr>
<td>Laminins</td>
<td>Perlecan</td>
</tr>
<tr>
<td>Nidogen/entactin</td>
<td>Agrin</td>
</tr>
<tr>
<td>Fibulin</td>
<td>Testican</td>
</tr>
<tr>
<td>Modular proteoglycans: hyalectans</td>
<td>Modular proteoglycans: nonhyaluronan binding</td>
</tr>
<tr>
<td>Decorin</td>
<td>Perlecan</td>
</tr>
<tr>
<td>Biglycan</td>
<td>Agrin</td>
</tr>
<tr>
<td>Fibromodulin</td>
<td>Testican</td>
</tr>
<tr>
<td>Lumican</td>
<td>Modular proteoglycans: nonhyaluronan binding</td>
</tr>
<tr>
<td>Epiphycan</td>
<td>Modular proteoglycans: hyalectans</td>
</tr>
<tr>
<td>(hyaluronan- and lectin-binding)</td>
<td>(hyaluronan- and lectin-binding)</td>
</tr>
<tr>
<td>Aggrecan</td>
<td>Aggrecan</td>
</tr>
<tr>
<td>Versican</td>
<td>Versican</td>
</tr>
<tr>
<td>Neurocan</td>
<td>Neurocan</td>
</tr>
<tr>
<td>Brevican</td>
<td>Brevican</td>
</tr>
</tbody>
</table>

Fig. 2 Modular representation of one Fn chain. Domains or sequences interacting with other molecules are indicated. EIIIA, EIIIB, and IICCS represent modules which can be spliced out or inserted during RNA processing.
coiled coil stabilized by disulfide bonds and by further assembly of two trimers into a hexamer. Located next to the oligomerization domain is a series of EGF-like motifs, followed by a row of FnIII repeats, the number of which varies with alternative spliced variants and tenascin type. The particularly well conserved carboxyterminal sequence contains motifs as those found in the globular part of β and γ chains of fibrinogen, and a putative divalent cation binding site. Oligomerization and module arrangement of tenascin chains result in a molecule with six extended arms of 75 nm. It is not known how tenascins are integrated into the ECM, and there is no evidence for self-binding.

Fibrillins are a family of proteins consisting almost entirely of EGF-like motifs, 47 motifs with six cysteine and 8 motifs with eight cysteines [26, 27]. Several motifs bind calcium which probably serves to stabilize the structure of fibrillin [28] which appears as a flexible and extended molecule of 148 nm [29]. The repetitive motifs are interrupted once by a short sequence postulated to be a hinge region allowing bending of the molecule.

The most typical and abundant NC glycoproteins of basement membranes are laminins. They are very large molecules constituted by the association of three genetically distinct polypeptides, the α, β, and γ chains. Five α (α1–α5), three β (β1–β3), and two γ (γ1–γ2) chains have been cloned and sequenced, and more than ten different laminin isoforms have been predicted to exist [30]. However, due to difficulties in the purification of the proteins most available data are still derived from studies with laminin 1 (α1β1γ1), the first identified isoform (Fig. 3). This has been isolated from the Engelbreth-Holm-Swarm tumor transplantable to mice [14] and is therefore frequently referred to as EHS laminin. The α1 (approx. 400 kDa), β1 and γ1 (each approx. 200 kDa) chains are associated into a cross-shaped molecule with three short arms – one with a length of 43 nm and two of 34 nm – and a 77-nm long arm. The aminoterminus of each chain forms the short arms of the cross, and the corresponding deduced sequence presents cysteine-poor regions folded into globular domains and EGF-like repeats aligned in rods. The β1 and γ1 chains contain two globes and two rods while the α1 chain contains three of each. The long arm of the molecule is formed by the carboxyterminal part of the α1, β1, and γ1 chains, whose sequences are constituted by repeating heptad motifs allowing folding of the three chains into a coiled-coil α helix. At the carboxyterminus the α1 chain is longer than the β1 or γ1 chains and is folded into five globes, the G domains [14].

Other isoforms result from the association of either one of the long, α2 or α5 (350–380 kDa), or short, α3 or α4 (190–200 kDa), variants of the α chain with the γ1 and either the β1 or the β2 chain, the latter being a further variant of the β chain. In addition, two shorter chains, β3 and γ2, are associated with the α5 chain to form laminin 5 [31]. The deduced sequence of each laminin chain shows that the approx. 600 residue carboxyterminal region consists of repeated heptad peptides suitable for trimer assembly into a coiled-coil structure. This has been confirmed by circular dichroism spectra which indicate that α helical structures account for about 30% of the conformation of laminins 2, 4, and 5 [32, 33]. Similarly to the laminin α1 chain, the carboxyterminus of all α chain variants contains five homologous domains, the G domains. For the α2 and α5 chains this region might be processed, which results in the cleavage of the two last G domains, G4 and G5 [31, 34].

In contrast to a similar overall organization of the carboxyterminal region, there is a considerable variability in the aminoterminal portion of the various laminin chains. The α1, α2, and α5 chains contain three dissimilar globular domains intercalated by 17 or 21 EGF-like modules, while the α3 and α4 chains contain only one small globular domain and three EGF-like modules. The β1, β2, and γ1 chains contain 13 and 11 EGF-like modules, respectively, and two globular domains, whereas one globular domain and 6 or 8 EGF-like repeats are present in the β3.
or the $\gamma_2$ chain, respectively. The structural differences are reflected by the gross shape of the laminin isoforms viewed by rotary shadowing showing particles with three short arms, two short arms, or no short arms.

The different laminin chains have a specific spatio-temporal expression and a tissue-specific localization which has been demonstrated for several laminin variants [35]. The $\alpha_2$-chain containing isoforms are present essentially in basal laminae of muscle and motor neurons synapses, while the $\alpha_3$-chain containing isoforms are restricted to the basement membranes underlying stratified epithelia. At an ultrastructural level $\alpha_5$-chain containing isoforms have been shown to be associated with the anchoring filaments of the dermo-epidermal junction originating at the hemidesmosomes and spanning the lamina lucida towards the lamina densa [31, 36].

In addition to laminins, smaller NC glycoproteins are present in basement membranes. Nidogen/entactin, a single-chain molecule of 150 kDa, consists of two amino-terminal globular domains, $G_1$ and $G_2$, separated by a short stretch of amino acids, and one carboxyterminal globular domain, $G_3$, rich in the YWTD motif found in the low-density lipoprotein receptor. Domains $G_2$ and $G_3$ are connected by a 15-nm rod constituted by the repetition of five EGF-like motifs [37]. Fibulins are represented by two isoforms, 1 and 2. Both are rich in EGF-like repeats (9–11) forming a central rod flanked by globular domains. Fibulin-1 is a single-chain molecule while fibulin-2 associates into a trimer [38].

### Polymers of the interstitial connective tissue

The fibril-forming collagens are synthesized as precursors, the procollagens, containing large globular N- and C-terminal propeptides which are enzymatically removed after secretion into the extracellular space. This processing into mature collagen results in molecules constituted by a 300-nm helical rod terminated by very short NC sequences, the telopeptides, for the most abundant collagens I, II, and III, or by a small globule for the closely related minor collagens V or XI [3]. This step is followed by fibrillogenesis in which collagen molecules assemble longitudinally head-to-tail and aggregate laterally in a quarter-staggered manner (Fig. 4). This results in the formation of two major tissue-specific fibrillar polymers: the collagen II containing fibrils typical of cartilage and the collagen I containing fibrils of most other interstitial connective tissues, both exhibiting the characteristic striation with a 67-nm periodicity. In terms of collagen composition and fibril morphology, these two networks are heterogeneous due to mixing of different proportions of collagen I, III, and V, within collagen I containing fibrils or collagens II and XI in collagen II containing fibrils [39, 40]. The current hypothesis postulates that collagens V and XI constitute the core of the fibrils, and that collagens I and III (fully or partially processed), or collagen II polymerize around this core (Fig. 4). The abundant collagens I, II, and III, endow the fibrils with mechanical strength, while collagens V and XI seem to regulate fibril morphogenesis in a tissue-specific manner. In addition to the central 300-nm collagen rod, collagens V and XI retain part of the amino-terminal domain which projects onto the surface of the fibrils, a feature which could regulate fibril growth by limiting further lateral aggregation of collagen I/III or II monomers [41, 42]. Using a dominant-negative strategy to decrease secretion of collagen V, and so to perturb the collagen I/V ratios within fibrils, it has recently been shown that collagen V indeed regulates fibrillogenesis, the amount of collagen V within fibrils being inversely proportional to their diameter [43].

Alternatively, fibril morphology can be determined by the presence of other tissue-specific COL or NC molecules associated with the fibrils such as FACIT proteins or proteoglycans. Assembly between FACIT molecules has not been found, but their association to collagen fibrils is suggested by the immuno-electron microscopy observation that they periodically decorate the fibrils [44]. Interactions between the COL tail of the FACIT and collagen fibrils have been observed to occur either directly by formation of covalent bonds between collagens II and IX [45, 46] or indirectly by presumably the involvement of decorin which binds both collagen I [13] and collagens XII or XIV [47]. These associations leave free the very large aminoterminal NC portions of the FACIT which can extend out of the fibrils (Fig. 4). A function of the collagen XII and XIV could therefore be to contribute to the spatial organization of the interstitial collagen fibrils via further interactions between their NC aminoterminal domains and proteoglycans [48]. Moreover, procollagen I N proteinase, one of the enzymes required for processing of procollagens into collagens, binds specifically to collagen XIV, which suggests that collagen XIV regulates fibril assembly by immobilizing the enzyme in the close vicinity of collagen I fibrils [49].

The backbone of another distinct polymer, the microfibrillar network with a typical 100-nm periodicity, is constituted by collagen VI. The helical portions of two collagen VI monomers are associated in an antiparallel fashion, and two dimers are associate into tetramers.

![Fig. 4 Model for a collagen I containing fibrils. The constitutive molecules are indicated](image-url)
which in turn associate end-to-end to form microfibrillar structures [1]. Interactions between collagen VI and glycosaminoglycans have been observed [50–52], which might be responsible for connecting the collagen VI network with other polymers of the ECM.

Two other prominent meshworks of the ECM are based on the assembly of NC structural glycoproteins, Fn for one and elastin and fibrillin for the other. In vivo, Fn forms fibrils stabilized by disulfide bridges [18]. However, in contrast to the classical self-assembly of collagen, Fn polymerization is cell driven. Several regions of the molecule are involved in the assembly process: the aminoterminal 29-kDa heparin-binding domain, the RGD-containing cell-binding domain, and the first FnIII repeat [53, 54]. Other structural domains of Fn are involved in interactions with heparin and fibrin. Next to the aminoterminal heparin-binding region is a collagen binding site likely to be involved in vivo in the association of the Fn polymers with various COL structures which could play a role in the maintenance of tissue architecture [55]. Fibrillin monomers polymerize into ordered aggregates constituting the 10- to 12-nm diameter microfibrils associated with elastin. Assembly of the elastin/microfibril system is likely to require several other small proteins, in particular the microfibril-associated glycoproteins) which are associated with the bead region of fibrillin-containing microfibrils [56, 25].

A large body of evidence indicates that proteoglycans provide a glue between the various COL and glycoprotein networks by developing specific interactions with single components. The small leucine-rich proteoglycans characteristically bind to other ECM molecules. The protein core of decorin or fibromodulin binds to different sites on collagen I or II containing fibrils [57]. As mentioned above, this seems to be critical for the association of the FACIT to collagen I containing fibrils and for the regulation of fibril growth. In cartilage matrix a similar function could be supported by fibromodulin. Interactions of the protein core of decorin have also been reported with collagen VI [50], Fn, and thrombospondin [58, 59], while biglycan interacts with collagen via the glycosaminoglycan side chains [60]. In this way proteoglycans regulate matrix assembly, stabilize the ECM architecture, and may play a role in the spatial arrangement of the structural polymers.

**Polymers of the basement membranes**

Collagen IV molecules have the property to self-assemble into large polymers by dimerization and covalent cross-linking of the NC1 domains and by parallel and antiparallel overlapping on 30 nm of four carboxyterminal 7S domains [14]. The dimers are stabilized by intermolecular disulfide bonds and the tetramers by lysine-derived intra- and intermolecular cross-links. In addition, part of the major triple helices aggregate laterally and form a supercoiled structure [61]. This results in the formation of insoluble and large polymers which constitute a stable scaffold for the basal lamina and an anchoring support for cells and other constituents of the basement membrane. Another polymeric scaffold is provided by laminin 1 which has the propensity in the presence of calcium to self-associate into polymers by mean of interactions between the most aminoterminal globular domain of the short arms, domain VI [62]. In addition, laminin 1 forms a stable high-affinity equimolecular complex with nidogen upon an interaction between the EGF-like motif III4 of the laminin γ1 chain and the carboxyterminal G3 domain of nidogen [63]. Crystal structure analysis of this EGF-like motif indicates that the crucial residues for nidogen binding are close together and exposed at the surface of the module [64]. Furthermore, the G3 aminoterminal globule of nidogen binds with high affinity to collagen IV [37, 65], the core protein of perlecan [66], or to fibulin-1, while nidogen G3 domain has the property to interact with fibulin-2 [67]. Nidogen therefore appears very critical for connecting the collagen IV and laminin scaffolds and for anchoring other components such as perlecan or fibulins (Fig. 5).

This model of association based on laminin polymerization via the short arm’s domain VI and on the bridging role of nidogen is predicted to apply to basement membranes where domain VI- and γ1 chain-containing laminin isoforms are present. A different scheme of assembly (Fig. 5) was recently developed for laminin 5 (α2β1γ1), whose γ1 chain does not contain a high affinity nidogen binding site, and which forms a covalently linked complex with laminin 6 (α5β1γ1) or laminin 7 (αγβγ). Only the latter contain a potential nidogen binding site, and, moreover, polymerization domain VI is present only once in laminin 5 and twice on laminin 6 or 7 [68]. It therefore remains to be demonstrated whether this is sufficient to allow polymerization as has been observed for laminin 1.

On a molecular basis little is known about how basement membranes are connected to the underlying interstitial connective tissue. Based on biochemical observa-
tions, fibulin-2 binds to the G2 domain of nidogen, to Fn, and to fibrillin, while fibulin-1 has affinity for the G domain of laminin, the G2 domain of nidogen, and for fibrillin [67, 69, 70]. The fibulins could therefore be involved in connecting basement membrane networks to the Fn fibrils or the fibrillin-associated microfibrils of the interstitium (Fig. 5). For basement membranes underlying stratified epithelia the collagen VII containing anchoring fibrils play a major role in securing the basement membrane to the underlying mesenchyma, as has been deduced from studies of patients with epidermolysis bullosa dystrophica [71]. After removal of the NC2 domain, collagen VII molecules dimerize by antiparallel overlapping of about 60 nm of the carboxyterminal region of the triple helices [72]. The dimers are associated laterally, which results in the formation of 800-nm-long bundles, a size which corresponds to that observed for anchoring fibrils. The aminoterminal domains located at both ends of the bundles are presumably responsible for the fanlike appearance of the extremities of the anchoring fibrils and provide potential targets for interactions with other ECM components such as collagen IV [6] and the recently identified GDA-J/F3 antigen, a small 45- to 50-kDa protein [73]. Furthermore, the anchoring fibrils extend from the basis of the lamina densa out into the mesenchyma where they intermingle with interstitial collagen fibrils. Whether direct interactions between collagen VII and interstitial components exist is not known.

From its transmembrane localization it can be presumed that collagen XVII serves to anchor basal epithelial cells to the basement membrane, a prediction which is supported by recent investigations of patients with bullous diseases and of genetically engineered mice [71]. This collagen could also participate in the assembly and stabilization of basement membrane complexes since, as for other collagens, the major triple helical rod of several collagen XVII molecules could interact laterally, on the one hand, and, on the other, serve to project the C-terminal interrupted and flexible COL region in an orientation suitable for interactions with other basement membrane components. Collagen XVII may also function as a signal transducer since the sequence of the intracellular domain contains several tyrosine residues that can be phosphorylated.

Various structural domains with different functions are therefore present and shared by the constituents of the ECM. Collagen and coil-coiled α helices serve to assemble and maintain together the polypeptide chains of molecules. In addition, multiple lateral interactions between collagen helices are critical for the formation of supramolecular structures such as microfibrils and fibrils with mechanical stability and resistance to tension and pressure. Rows of successive EGF-like motifs or FnIII repeats form rodlike elements with a certain degree of flexibility and which may represent spacers between other biologically active domains involved in specific interactions to connect the components of the ECM, as is the case for the nidogen-binding motif of the EGF-like module γ1–III4 of laminin. This is likely to bring each individual component in a defined conformation and to affect the spatial orientation of their different domains rendering them accessible or not for biological interactions. Besides protein-protein interactions responsible for connecting the components of the extracellular space, many of the components develop other interactions with cell surface molecules and contribute to the attachment of cells to the ECM.

Cell binding to the ECM

Specific interactions between cells and the ECM are mediated by transmembrane molecules, mainly integrins and perhaps also members of the collagen family, proteoglycans, or cell-surface-associated components. These interactions lead to a direct or indirect control of cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis. In addition, integrins function as mechanoreceptors and provide a force-transmitting physical link between the ECM and the cytoskeleton [74, 75]. Integrins are transmembrane α/β noncovalently associated heterodimers and formed by noncovalent association of one α/β subunit [76]. Twenty-two combinations of 16 different α and 8 different β subunits have been identified. Each subunit has a large extracellular domain, a transmembrane segment, and a short intracellular tail of 35–50 residues, except β4 whose intracellular domain is of approx. 1000 amino acids. The ligand binding site is contributed by portions of both the α and β subunit extracellular domains, while the cytoplasmic domains interact directly or indirectly with several cytoskeletal-linker molecules and actin (Fig. 5). In addition, ligand binding is concomitant with integrin clustering and is rapidly followed by recruitment of cytoskeletal-linker molecules into adhesion complexes and anchorage of actin stress fibers to these complexes. A cascade of phosphorylation reactions occurs simultaneously to the rearrangement of the cytoskeleton which probably leads to transmission of signals by conformational changes and establishment of interactions between the adhesion complex-associated molecules [77].

The use of proteolytic fragments, synthetic or recombinant (poly)peptides replicating specific or mutated sequences or domains of authentic ECM molecules has allowed the precise localization of cell binding sites on multidomain matrix macromolecules. This has also been instrumental in demonstrating that multiple and interactive modules and peptides communicate with cells. Structure and conformation of some ligand motifs have now been clarified by magnetic resonance imaging and X-ray crystallography analyses.

Cell adhesion promoting activity was first observed for Fn, dissected down to a 11-kDa peptide, and mapped to an RGDS (Arg-Gly-Asp-Ser) sequence located in the tenth FnIII unit by using overlapping synthetic peptides [78]. However, for a full activity similar to that of authentic Fn the RGDS peptide needs the presence of synergistic sequences located on the eighth and ninth FnIII
motifs (Fig. 2). The RGD-dependent cell binding region of Fn is recognized by the classical Fn receptor, the $\alpha_\text{IIb} \beta_3$ integrin, and by the promiscuous $\alpha_\text{IIb} \beta_1$, $\alpha_\beta_3$, and $\alpha_\beta_6$ integrins [76]. Other cell binding sequences of Fn are contributed by the IDAPS sequence of the heparin-binding domain (12th–14th FnIII motifs) and the REDV and LDVPS sequences present in, respectively, the CS5 and CS1 peptides of the adjacent alternatively spliced IICCS region [79]. The two first are of low affinity and the third of high affinity for the $\alpha_\text{IIb} \beta_3$ integrin. In addition, the CS1 site is a ligand for the $\alpha_\text{IIb} \beta_1$ integrin. Additional interactions between cell surface heparan sulfate and chondroitin sulfate proteoglycans with the heparin-binding domain are thought to reinforce or to stabilize the integrin-mediated interactions or, alternatively, to regulate integrin activation state or conformation [80, 81]. Tenascin, which is also rich in FnIII modules has both adhesive and counteradhesive activities. Adhesive activities have been localized to the RGD sequence present in the 3rd FnIII motif, a ligand for the $\alpha_\text{IIb} \beta_3$ or $\alpha_\beta_1$, and to the carboxyterminal domain (Fbg domain) which interacts with phosphatidylinositol-anchored heparan sulfate proteoglycans. Counteradhesive activity has been mapped to the EGF-like motif-containing aminoterminal domain and to the seventh and eighth FnIII motifs [82]. By interacting with F11/contactin, a cell surface molecule of the immunoglobulin superfamily, tenasin also promotes neurite outgrowth.

Close structural analysis of tenth FnIII module of Fn [19, 20] and of the third FnIII motif of tenasin [83] reveals that the RGD sequence is inserted at the apex of a flexible loop emerging from two rather rigid $\beta$ strands, and that the acidic and basic side chains point away in opposite directions. The tripeptide therefore protrudes out of the surface of the module in a stereochemical presentation well adapted for easy and fast recognition by the receptor. Interestingly, the RGD-containing loop is more extended and flexible in Fn than in tenasin, which may explain the difference in specificity. Furthermore, compact folding of FnIII modules into rigid $\beta$ sheets probably allows regulatory interactions with adjacent motifs. The activity of the RGD sequence located in tenth FnIII motif of Fn is positively affected by the eighth and ninth FnIII modules, located approx. 20 kDa away, but spatially very close to the tripeptide [84]. In contrast, the second FnIII motif of tenasin is thought to hamper accessibility to the RGD sequence located in the third FnIII module. Most interestingly, it has been postulated that FnIII modules can stretch by reversible unfolding [21]. Based on these observations one can speculate that a potential regulation of the synergy occurs by dynamic folding and unfolding of the cell binding sites.

Laminin 1 contains multiple cell binding sites contributed by different domains of the molecule (Fig. 3). The first identified cell binding domain of laminin 1, the pepsin resistant fragment P1, corresponds to the center of the cross formed by the three short arms and lacking most of the globular domains [85]. The cell binding site is provided by an RGD sequence located on one EGF-like module of domain IIIa of the mouse laminin $\alpha_\text{1}$ chain and is the target for the promiscuous RGD-binding integrins such as $\alpha_\text{IIb} \beta_1$ or $\alpha_\beta_3$. The native RGD sequence is, however, cryptic and becomes accessible to cells after proteolytic degradation of the adjacent domain IVa. Another cell binding site, corresponding to fragment E$_{1-4}$ and available to cells on the intact short arms, is a ligand for $\alpha_\text{IIb} \beta_1$ or $\alpha_\beta_3$ integrins [85].

A major cell binding site of laminin 1 is located on fragment E$_{1-2}$ and is a ligand for $\alpha_\text{IIb} \beta_1$ or $\alpha_\beta_3$ integrins (Fig. 3). Fragment E$_{1-2}$ is contributed by the carboxyterminal region of the triple-stranded helix formed by the $\alpha_1$, $\beta_1$, and $\gamma_1$ chains and by the first three G domains of the $\alpha_1$ chain. Although the cell binding site is probably located on the G domains, the presence of the coiled-coil structure is required for activity and is thought to maintain a correct folding of the G domains [85]. Such complex requirements have impaired dissecting further down the cell binding site and highlights the crucial role of a correct folding of noncontiguous sequences on the spatial organization and consequent biological activity of adhesion motifs.

Another cell binding site of laminin is provided by the heparin-binding fragment E$_3$ corresponding to the most carboxyterminal G$_4$ and G$_5$ domains. This cell binding site was called “minor” because it induces, at least in vitro, adhesion of only some types of cells, no spreading, and is integrin independent [86, 87]. The cellular receptor for this fragment is $\alpha$-dystroglycan, which is the extracellular part of a large oligomeric complex associated with dystrophin, and which is also a receptor for agrin [88].

Although the data are still more fragmentary for laminin isoforms, it is well established that they induce cell adhesion and contain at least one conformation-dependent cell binding site on the long arm of which $\alpha$ helical structure is required for activity. Laminin isoforms are ligands for both the $\alpha_\text{IIb} \beta_3$ and $\alpha_\beta_1$ integrins, whose affinities differ, however, between isoforms, which suggests that the adhesion motifs are probably not strictly identical [85]. Another integrin, $\alpha_\beta_4$, which is localized at the basal side of epithelial cells, is associated in vivo with laminins. However, in vitro data indicate that this integrin has very low or no affinity for laminin ligands. The reason of this controversy could be that binding of $\alpha_\beta_4$ integrin to laminins is not involved in the initial recognition of the ligand but is necessary for stabilization of the interactions [89].

Synthetic peptides representing short linear sequences of various laminin chains have been shown to inhibit or potentiate laminin-induced cellular responses [90]. The molecular mechanisms underlying these effects are, however, not clear, and it remains to be definitively established whether the synthetic peptides really mimic the activity of the authentic laminin molecule since at least in vitro several of these short sequences are cryptic in the authentic molecule [91, 92].

Several members of the collagen superfamily are endowed with cell adhesion promoting activity. For fibrillar collagens I, II, III, and V basement membrane collagen
IV, and microfibrillar collagen VI, the cell adhesion activity has been mapped to the triple helical rod [93, 98]. Interestingly, although collagens contain many RGD sequences, cell adhesion to native collagens is not inhibited by RGD peptides and is mediated by the RGD-independent integrins α1β1 and α2β1. Because short COL peptides unfold at physiological temperature, it has been difficult to map cell binding sites on the collagen helices, except for the disulfide stabilized cell adhesion peptide of collagen IV in which two aspartate residues within each α1(IV) chain and one arginine residue within the α2(IV) chain were found to form the motif for α1β1 integrin binding [99]. That the same integrins recognize all the different collagen ligands is a rather intriguing observation because on a linear sequence basis collagen homologies are restricted mainly to the Gly-X-Y triplets. However, folding of the chains into a helix may result in similar arrays of amino acid side chains at the surface of the rods.

Surprisingly, unfolding of the triple helices abolishes the binding of the α1β1 and α2β1 integrins and uncovers new RGD-dependent binding sites for integrins α1β3, α1β1, or α1β3, [95, 96, 98, 100]. On denatured collagens it is interesting to note that only some of the RGD sequences are active and recognized by integrins. In particular, among the 13 RGD sequences contributed by the human collagen VI chains only one has activity [100]. Similarly, none of the 3 RGD sequences present on collagen IV α chains seems to be active. This indicates that integrin recognition requires specific folds of the RGD motif, a way to control the number of active cell adhesion sites. Similarly, assembly of collagen monomers into fibrils may result in a local heterogeneity modulating accessibility of the cell binding sites [101].

Many more adhesion molecules are present in the ECM and contribute to the control of cell behavior. Furthermore, other molecules devoid of direct cell binding activity probably modulate the activity of adhesion molecules. This is the case with decorin, which binds to Fn or thrombospondin and interferes with their cell adhesion promoting activity [58, 59]. Another indirect way to control cellular functions is used by proteoglycans. In general, proteoglycans control the hydration of the ECM and the spacing between collagen fibrils and network which could facilitate cell migration. They also regulate cell function by controlling growth factor activity as has been well illustrated for decorin, biglycan, and fibromodulin, which bind to all isoforms of transforming growth factor β [102, 103], and heparan sulfate proteoglycans, which bind and store basic fibroblast growth factor [104].

Our understanding of the molecular basis for the formation of ECM polymers and of their functions is expanding rapidly and is both completed and confirmed by genetic and molecular studies of human diseases and of transgenic mice with deleted, mutated, or truncated genes. As will be developed in another contribution, there is now ample evidence showing that several human diseases are due to alterations or absence of the interactions between ECM components or cells.

Acknowledgements The authors are supported by the Centre National de la Recherche Scientifique and the Association pour la Recherche sur le Cancer.

References

5. Chu ML, Pan TC, Conway D, Kuo HJ, Glanville RW, Timpl R, Mann K, Deutzmann R (1989) Sequence analysis of α1(VI) and α2(VI) chains of human type VI collagen reveals internal tripletization of globular domains similar to the A domains of von Willebrand factor and two α(VI) chain variants that differ in the carboxy terminus. EMBO J 8:1939–1946


71. Bruckner-Tuderman L, Bruckner P (this issue)
84. Leathy DJ, Aukhil I, Erickson HP (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
99. Eble JA, Golbik R, Mann K, Kühn K (1993) The $\alpha_1\beta_1$ integrin recognition site of the basement membrane collagen molecule $[\alpha_1(IV)]2\alpha_2(IV)$. EMBO J 12:4795–4802


