FEMS Microbiol Lett. Author manuscript; available in PMC 2011 February 08.

Published in final edited form as:
FEMS Microbiol Lett. 2009 February ; 291(1): 1–16. doi:10.1111/j.1574-6968.2008.01420.x.

Noncellulosomal cohesin- and dockerin-like modules in the three domains of life

Ayelet Peer[1], **Steven P. Smith**[2], **Edward A. Bayer**[3], **Raphael Lamed**[1], and **Ilya Borovok**[1]

[1] Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel

[2] Department of Biochemistry, Queen's University, Kingston, ON, Canada

[3] Department of Biological Sciences, Weizmann Institute of Science, Rehovot, Israel

Abstract

The high-affinity cohesin–dockerin interaction was originally discovered as modular components, which mediate the assembly of the various subunits of the multienzyme cellulosome complex that characterizes some cellulolytic bacteria. Until recently, the presence of cohesins and dockerins within a bacterial proteome was considered a definitive signature of a cellulosome-producing bacterium. Widespread genome sequencing has since revealed a wealth of putative cohesin-and dockerin-containing proteins in Bacteria, Archaea, and in primitive eukaryotes. The newly identified modules appear to serve diverse functions that are clearly distinct from the classical cellulosome archetype, and the vast majority of parent proteins are not predicted glycoside hydrolases. In most cases, only a few such genes have been identified in a given microorganism, which encode proteins containing but a single cohesin and/or dockerin. In some cases, one or the other module appears to be missing from a given species, and in other cases both modules occur within the same protein. This review provides a bioinformatics-based survey of the current status of cohesin- and dockerin-like sequences in species from the Bacteria, Archaea, and Eukarya. Surprisingly, many identified modules and their parent proteins are clearly unrelated to cellulosomes. The cellulosome paradigm may thus be the exception rather than the rule for bacterial, archaeal, and eukaryotic employment of cohesin and dockerin modules.

Keywords

bioinformatics; cellulosome; enzyme complex; protein-protein interaction

Introduction

The cellulosome was initially defined in the early 1980s as a discrete multienzyme complex responsible for the binding and degradation of the most common and abundant polysaccharide in nature – cellulose (Bayer *et al.*, 1983; Lamed *et al.*, 1983a, b). This

Correspondence: Edward A. Bayer, Department of Biological Sciences, Weizmann Institute of Science, Rehovot 76100, Israel. Tel.: +972 8 934 2373; fax: +972 8 946 8256; ed.bayer@weizmann.ac.il.

Additional Supporting Information may be found in the online version of this article.

original classification was based on studies of the cellulase system in the anaerobic, thermophilic, cellulolytic bacterium, *Clostridium thermocellum*. With additional descriptions of cellulosome systems in other bacteria (mainly from Gram-positive anaerobes) over the past quarter century (Table 1), this definition has been broadened to include the degradation of various other insoluble plant cell-wall complex polysaccharides besides cellulose (Shoham *et al.*, 1999; Schwarz, 2001; Bayer *et al.*, 2004; Doi & Kosugi, 2004; Demain *et al.*, 2005). Indeed, many of the known bacterial cellulosomes comprise different types of hemicellulases, such as xylanases, mannanases, arabinofuranosidases, lichenases, and pectate lyases, in addition to the various cellulases. Moreover, additional types of enzymes (e.g. peptidases), serpins, and putative structural proteins also appear to be components of cellulosomes. However, the defining quality of the cellulosome, which enables its efficient hydrolysis activity, remains its assembly into an organized single entity of intricate structure and architecture, and not a simple random mixture of enzymes in the free state.

The bacterial cellulosome

The classical bacterial cellulosome can be characterized by its principal subunit, the scaffoldin, which is a large non-catalytic protein that functions as an integrator of the various enzymes and other cellulosomal components into a single functional entity. The scaffoldin comprises a substrate-targeting cellulose-binding module and cohesin modules, usually found as numerous tandem repeats, whereas the enzymes contain a catalytic domain and a single dockerin module complementary in specificity to the cohesin modules. Divergent affinities and specificities between the cohesin and the dockerin modules mediate the coordinated integration of the enzymes and other dockerin-bearing components into the scaffoldin and create the intricate cellulosome architecture. Some cohesin–dockerin pairings are among the highest known affinity interactions ($K_a > 10^{11}$ M^{-1}) between two proteins and can be perceived as a kind of plug-and-socket arrangement, whereby the dockerin module plugs into the cohesin module (Pagès *et al.*, 1997; Fierobe *et al.*, 1999; Mechaly *et al.*, 2001; Schaeffer *et al.*, 2002; Carvalho *et al.*, 2003, 2007; Jindou *et al.*, 2004a, b; Adams *et al.*, 2006; Pinheiro *et al.*, 2008). Other types of cohesin-containing scaffoldins serve supplementary functional roles in cellulosome structure, including anchoring or adaptor components. This array of scaffoldins is assembled by divergent types of cohesin–dockerin interactions, which govern overall cellulosome architecture and function. The differences among the cellulosomal components in the various cellulosome-producing species reflect the enormous and surprising diversity of cellulosome architecture in nature.

Cohesin and dockerin modules

Cohesin modules commonly comprise *c.* 140 amino-acid residues and are typically deficient in tryptophan, tyrosine, and cysteine residues. They display substantial sequence variability among cellulosome-producing bacteria, and even within a single bacterium. Several types of cohesin modules have been distinguished based on their primary structure, tertiary structures, and binding specificity (Salamitou *et al.*, 1992; Leibovitz & Béguin, 1996). The type-I cohesin module is an elongated, conical molecule, which folds into a nine-stranded β-sandwich that exhibits a jellyroll topology (Shimon *et al.*, 1997; Tavares *et al.*, 1997;

Carvalho *et al.*, 2003). The structure of the type-II cohesin module has an overall fold similar to that of type-I, but includes distinctive additions: two 'β-flaps' interrupting strands 4 and 8 and an α-helix at the crown of the protein module (Carvalho *et al.*, 2005; Noach *et al.*, 2005; Adams *et al.*, 2006). The discovery of cellulosomes in *Ruminococcus flavefaciens* has contributed additional diverse types of cohesin modules to the overall repertoire (Ding *et al.*, 2001; Rincon *et al.*, 2003, 2004, 2005). However, the nine-stranded β-sandwich and jellyroll topology appears to be a definitive structural characteristic of the cohesin fold (Alber *et al.*, 2008).

Dockerin modules typically comprise 60–70 amino-acid residues and are classified into types according to the cohesin with which they interact. The dockerin sequences usually contain two duplicated *c.* 22-residue segments, frequently separated by a linker of 9–18 residues (Tokatlidis *et al.*, 1991; Volkman *et al.*, 2004). The initial 12 residues of these duplicated sequences bear striking similarity to the consensus sequence of the calcium-binding loop in the EF-hand motif (Chauvaux *et al.*, 1990), where residues at the calcium-coordinating positions 1, 3, 5, 9, and 12 are highly conserved (usually Asp and Asn), as is the glycine residue at the hinge position 6 (Fig. 1) (Giallo *et al.*, 1983). This similarity is, in effect, restricted to the two calcium-binding loops and their respective exiting F helix, thereby suggesting an 'F-hand motif' (Pagès *et al.*, 1997; Lytle *et al.*, 2001), which would in part distinguish the dockerin module from the classical EF-hand motif (Rigden & Galperin, 2004). The type-III dockerin modules from *R. flavefaciens* cellulosomal components are much more divergent, particularly in their second segment, where, in some cases, the identity of the calcium-binding loop is not immediately recognizable (Rincon *et al.*, 2005, 2007). Experimental evidence has shown that dockerin modules do bind calcium, which is required for the structural stability of the module in the unbound state and for cohesin recognition (Yaron *et al.*, 1995; Lytle *et al.*, 2000; Adams *et al.*, 2005).

Sequence alignment analyses of cellulolytic type-I dockerin modules have suggested positions 10, 11, 17, 18, and 22 of each F-hand motif as the major cohesin-recognition residues, which has been corroborated recently by mutagenesis studies and the crystal structures of cohesin–dockerin complexes (Pagès *et al.*, 1997; Mechaly *et al.*, 2000; Schaeffer *et al.*, 2002; Carvalho *et al.*, 2004, 2007; Pinheiro *et al.*, 2008). These residues enable both species-specific and function-linked recognition (type-I vs. type-II) of the cohesin modules, and therefore represent residue positions that are more variable among the different classes of dockerin modules.

Cohesin–dockerin interactions are mediated by extensive hydrogen-bonding networks and hydrophobic interactions (Mechaly *et al.*, 2001; Carvalho *et al.*, 2003; Handelsman *et al.*, 2004; Nakar *et al.*, 2004). The type-I dockerin module associates with the cohesin module in an asymmetric manner, despite the internal twofold symmetry of the dockerin sequence (Carvalho *et al.*, 2003, 2005; Bayer *et al.*, 2004; Gilbert, 2007; Pinheiro *et al.*, 2008), whereby one face of the cohesin module (8-3-6-5 face) interacts predominantly with only one of the helices of the dockerin module. The twofold symmetry provides a second mode of interaction, whereby the dockerin module is rotated 180° with respect to its cohesin partner. Mutational and structural studies have illustrated how this dual mode of interaction allows for plasticity in cohesin recognition, which may thereby provide the structural flexibility

considered necessary either for the assembly of the various enzymes onto the scaffoldin and/or as a 'conformational switch' that would enhance synergistic action among cellulosomal enzymes (Carvalho *et al.*, 2004, 2007; Pinheiro *et al.*, 2008). In contrast, the type-II dockerin module in *C. thermocellum* does not display obvious sequential symmetry, and consequently interacts with its cognate cohesin module via a much more extensive hydrophobic surface involving the entire length of both helices (Adams *et al.*, 2005, 2006). Likewise, conspicuous sequence asymmetry in the type-III *R. flavefaciens* dockerins would argue against a dual mode of binding to their respective cohesin counterparts.

Global search for noncellulosomal cohesin and dockerin modules

The assembly of complementary catalytic subunits into the cellulosome enhances the synergistic interactions among enzymes, ultimately leading to highly efficient polysaccharide degradation. It was thus surprising to us that traditional cellulosomal signature components, namely cohesin- and dockerin-like sequences (Bayer *et al.*, 1998), have also been detected in noncellulolytic microorganisms. On the other hand, why would nature not exploit such a strong and specific protein–protein interaction for other purposes? The current explosion of newly detected cohesin- and dockerin-like modules are direct byproducts of genome sequencing programs, with the first such example being the discovery of cohesin- and dockerin-like modules in the archaeon *Archaeoglobus fulgidus* (Bayer *et al.*, 1999).

Two tandem genes identified in *A. fulgidus* encode proteins containing either a putative cohesin module or both cohesin-and a dockerin-like sequences (Bayer *et al.*, 1999). Their functioning as *bona fide* cohesin and dockerin modules has been confirmed biochemically; both cohesin modules recognize the lone dockerin (Haimovitz *et al.*, 2008). The apparent lack of cellulases and hemicellulases in this microorganism was, at the time, puzzling, because we had assumed that cohesin and dockerin sequences constituted 'cellulosome-signature sequences.' These observations suggest a much broader spectrum of cellulosomal components in microbial systems and a variety of corresponding noncellulosome-related functions.

Here, we explore the likelihood that the classical cellulosome is but one specific example of cohesin and dockerin use, and that newly detected modules imply novel functions for cohesin- and dockerin-bearing components. In this context, we present bioinformatics analyses designed to discover putative noncellulosomal cohesin and dockerin modules in a broad range of prokaryotic species, and surprisingly in a small number of eukaryotic species. For this purpose, in the present study, we have excluded the few known anaerobic bacteria that have already been shown to produce classical cellulosomes (Table 1), which are involved in polysaccharide degradation and exhibit cohesin-containing scaffoldin(s) and dockerin-containing carbohydrate-active enzymes.

Widespread distribution of noncellulosomal cohesin and dockerin modules

Bioinformatics-based analyses of the various databases indicate that putative cohesin and dockerin modules are widely distributed among different taxonomic divisions over and

above the classic cellulosome-producing *Firmicutes* group (summarized in Table 2). A comprehensive detailed list of bacteria, archaeons, and primitive eukaryotes that possess putative cohesin and dockerin sequences in their genome is presented in Supporting Information, Table S1, complete with the designated genes, their products, the location within the protein of the cohesin- or the dockerin-like sequence, and the actual sequence.

Nearly 40% of the known archaeal genomes and 14% of the known bacterial genomes contain either a putative cohesin and/or a dockerin module. Interestingly, the majority of these microbial genomes encode for only one or a few of these modules. Although both cohesin- and dockerin-like sequences appeared in many organisms, one or the other module was apparently lacking in about a quarter of the Archaea and 60% of the Bacteria. It is currently unknown whether this apparent deficiency in cohesin and dockerin pairs in the designated microbial genomes reflects the difficulty in identifying particularly divergent cohesin or dockerin sequences or whether the counterpart modules are indeed absent. In lieu of a modular partner, the lone cohesin- or dockerin-like sequence detected in a given microorganism may bind to a different type of as-yet-unidentified protein component. Alternatively, cohesin and dockerin modules produced by different organisms in the same ecosystem may be important for binding interactions among their components or even for interspecies cell–cell adhesion. On the other hand, a recognized cohesin or dockerin module may not participate in a binding interaction at all and may play a different type of role that has evolved in the given microorganism.

Cellulosome footprints in the Archaea

Until the identification of cellulosome-like signature sequences in the Archaea (Bayer *et al.*, 1999), cohesin and dockerin modules were associated exclusively with anaerobic bacteria, mainly Gram-positive bacteria, such as clostridia and ruminococci. However, following the discovery of cellulosome signature sequences in *A. fulgidus*, similar cohesin- and dockerin-like sequences were discovered in many of the sequenced genomes of other archaeal species. Intriguingly, the latter species were confined to Euryarchaeota; no cohesin-and dockerin-like sequences were detected in other archaeal phyla. Thus, 18 of the 31 species of the known euryarchaeotal genomes sequenced to date carry putative cohesins and/or dockerins (Table 2).

In several cases, notably the methanogens, the genes encoding putative cohesin and dockerin modules are organized in clusters on the chromosome. In a few genomes, an ORF that contains only a single copy of either a cohesin or a dockerin module is present (Table 2). In isolated cases, such as *Haloarcula marismortui* and *Haloquadratum walsbyi*, a single cohesin and several dockerin-encoding genes are evident. The genome of *Methanosarcina acetivorans* is particularly notable, because it includes a gene that encodes a cohesin-containing protein, 26 dockerin-containing genes, and a gene encoding a protein that contains both a cohesin and a dockerin module. This latter gene is reminiscent of the bifunctional gene in *A. fulgidus*. Curiously, a related but different species of the same genus, *Methanosarcina mazei*, contains a comparable set of genes that encode for similar parent proteins that are devoid of cohesin- and dockerin-like sequences. Interestingly, the genome of *M. mazei* does include a different cohesin-containing gene that codes for a hypothetical

protein as well as a gene that codes for both a single cohesin and a single dockerin module. The *Methanococcoides burtonii* genome also contains an orthologous gene for a cohesin/dockerin-bearing protein as well as two cohesin-containing genes.

Although the functional consequences of these observations are currently unknown, it is clear that the arrangement of cohesin and dockerin modules in the Archaea follows a paradigm that differs from the cellulosome mode, because no multiple cohesin-bearing, scaffoldin-like encoding genes are present in any of the archaeal genomes. Moreover, none of the predicted dockerin-containing proteins are consistent with common cellulosomal enzyme components (i.e. cellulases and hemicellulases).

Cellulosome footprints in the Bacteria

The concept of numerous cohesin modules arranged in a scaffoldin subunit for integration of catalytic and/or other types of subunits into a grand multienzyme complex seems to be unique to the cellulosomes of the anaerobic cellulolytic and hemicellulolytic bacteria. Thus far, like those of the archaeal genomes, in the other known bacterial genomes, the cohesin-like sequences do not appear to be organized into scaffoldin-like subunits.

In most cases, annotation of the deduced cohesin- and dockerin-containing proteins is not particularly informative, with most being hypothetical proteins or proteins of unknown function, and the role such proteins may play in the given bacterium remains obscure. With few exceptions, most of the proteins that contain these modules are not glycoside hydrolases. One notable exception is *Clostridium perfringens*, a human pathogen in which numerous cohesin- and dockerin-like sequences are integral parts of glycoside hydrolases, as will be discussed below in greater detail.

A more common theme related to glycoside hydrolases is the appearance of family-31 glycoside hydrolases (GH31) in several bacterial genomes with an associated cohesin and/or dockerin module – frequently both, and often with no other detectable cohesin- or dockerin-like sequences in the same bacterial species. This modular arrangement is observed in members of various divisions and classes of bacteria, and mainly includes species associated with the normal human gut microbiota, for example, *Eubacterium dolichum, Enterococcus faecalis, Lactobacillus casei, Ruminococcus torques, Akkermansia muciniphila*, as well as numerous *Bacteroides* species and clostridia – notably, in some cases, human pathogens (see Cohesin and dockerin modules in the bacterial pathogen *C. perfringens*). In addition to the cohesin and dockerin modules, the GH31 enzymes also commonly contain one or more copies of a family-32 carbohydrate-binding module (CBM32). The CAZY database (Coutinho & Henrissat, 1999) indicates that this particular family of enzymes acts on α-glucosidic bonds, whereas the binding specificities of the CBM32s show a preference for galactose and lactose residues. However, the precise function(s) of the GH31 enzymes in the various bacteria and their relationship with the CBM32s are currently unclear and await detailed biochemical characterization.

Additional examples of bacterial genes encoding cohesin and dockerin modules in the same protein exist, particularly in the order *Bacillales*. In this case, a cohesin module and a

dockerin module are cocomponents of putative serine proteases. The logic of having cohesin and dockerin modules together in the same protein is unclear, but nature often discounts human logic.

In a few isolated bacterial species, their genomes contain genes coding for cohesin and/or dockerin modules that appear in numerous predicted proteins. None of these examples, however, approach the extensive usage observed for cellulosomes, and none of the cohesin modules are organized together into a common multicopy scaffoldin. Nevertheless, within the context of the present article, some species deserve special comment.

Paenibacillus sp. JDR-2 has been described as an aggressively xylanolytic bacterium (St. John *et al.*, 2006; Chow *et al.*, 2007). The draft genome contains more than 130 ORFs predicted to encode various glycoside hydrolases. Although no multiple cohesin-containing scaffoldin has been identified, 10 different genes that encode a single cohesin-like module together with a cell surface-binding S-layer homology (SLH) module have been sequenced. Three genes code for both a cohesin and a dockerin but lack an SLH module. One of these has been annotated as a putative GH3 xylan 1,4-β-xylosidase. Analysis of the genome revealed only one dockerin-containing glycoside hydrolase (also a putative GH3 enzyme), which would presumably be attached to the cell surface via the SLH module of the aforementioned cohesin-bearing proteins. Four other dockerin-containing genes have been identified. Direct attachment to the surface appears to emphasize the apparent importance of these particular proteins (especially the enzymes) to cell function.

The genomes of three species of the novel bacterial phylum Planctomycetes contain genes encoding putative proteins that comprise putative cohesin and/or dockerin components. These unusual bacteria reproduce by budding, exhibit a complex life cycle, produce a holdfast stalk for attachment during budding, and possess very intricate internal membrane-separated compartments. In the case of the *Planctomyces maris* draft genome, numerous dockerin-bearing proteins were identified as zinc-dependent proteases, and yet no cohesin-like sequences were detected. In contrast, numerous genes that contain both cohesin-and dockerin-coding sequences were detected in the draft genome of *Blastopirellula marina* and the completed genome of *Rhodopirellula baltica*. Among the putative dockerin-containing proteins, *B. marina* codes for two alkaline phosphatases and a serine protease. The complement of putative cohesin- and dockerin-containing proteins in *R. baltica* includes the latter types of enzyme as well as peroxidases, peptidases, and a xylanase. In addition to the cellulosome-like modules, these proteins all contain a conserved N-terminal Planctomycete extracellular motif, thus implying their location outside of the cell. Some of the deduced proteins are very large and are predicted to reside on the cell surface.

The genomes of several other bacterial species also show a multiplicity of genes that encode cohesin- or dockerin-bearing proteins. These include *Hahella chejuensis, Pseudoalteromonas atlantica, Desulfatibacillum alkenivorans*, and various species of *Geobacter* and the cyanobacterium *Gloeobacter violaceus*. Again, the identification (or, in many cases, the lack of identification) of the parent proteins in each case lends little insight into the overall relationship with function in the above bacteria.

Cohesin and dockerin modules in the bacterial pathogen *C. perfringens*

The conservation of cohesin and dockerin modules among cellulolytic clostridia and the critical role they play in the carbohydrate-degrading properties of these microorganisms make it highly plausible that their presence would extend to other clostridial species. This is indeed the case for the *C. perfringens*, a ubiquitous anaerobe found in the soil and as a commensal member of the human and animal gastrointestinal microbial communities (Smith & Gardner, 1949; Songer, 1997). *Clostridium perfringens* is also an opportunistic pathogen that is responsible for the third highest number of food-borne illness cases (Rood & Lyras, 2006), and is the primary causative agent of gas gangrene in humans and enterotoxemia in animals (Songer, 1997; Rood & Lyras, 2006).

Searches of the annotated genomic sequences of the myonecrotic *C. perfringens* strains (ATCC 13124 and strain 13) (Shimizu *et al.*, 2002; Myers *et al.*, 2006) revealed dockerin-like sequences encoded by four ORFs in strain ATCC 13124 and by the orthologous genes in strain 13. A fifth putative dockerin module was also present in an additional gene product (CPF_2130) from ATCC 13124 (Table 2). Seven putative cohesin modules encoded by orthologous genes from *C. perfringens* stains ATCC 13124 and 13 were also identified, as was an additional strain 13-derived cohesin-containing gene product (CPE1523; Table 2, Table S1). Similar to the cellulolytic bacteria, all of the cohesin- and dockerin-containing gene products are predicted glycoside hydrolases (GH2, GH3, GH20, GH31, GH33, GH84, GH89, and GH95), two of which (GH33 and GH84) are identified toxins. However, their predicted specificities were not for biomass- and dietary-based polysaccharides, as might be expected, due to the presence of *C. perfringens* in the gastrointestinal tract of humans and animals, but rather for mammalian polysaccharides, including those found in the mucosal layer of the human gut, glycosaminoglycans, and other cellular glycans (Ficko-Blean & Boraston, 2006). Furthermore, no multiple cohesin-bearing scaffoldins were identified in the genomes of the *C. perfringens* strains analyzed. Rather, only single copies of cohesin-like sequences were detected, with the identified GH31 containing both a cohesin and a dockerin module, similar to that seen in several other human gut microorganisms. These observations suggest that in most cases this bacterium only has the capacity to form enzyme pairs, with GH31 able to form more elaborate carbohydrate-active enzyme complexes.

Our recent biochemical and biophysical studies indicate that at least three of the five *C. perfringens* dockerin modules interact with at least three of the seven cohesin modules (Adams *et al.*, 2008). Nuclear magnetic resonance and X-ray crystallographic structural studies illustrated that the putative cohesin modules from CPF_1442 and CPE1234 and the dockerin module for CPE0191 adopt the typical cohesin and dockerin folds, respectively (Fig. 2). Moreover, like their genuine cohesin and dockerin homologs, they interact at a similar intermolecular interface with ultrahigh affinity, due to extensive hydrogen-bonding and hydrophobic contacts (Adams *et al.*, 2008; Chitayat *et al.*, 2008a, b).

The symbiotic relationship between *C. perfringens* and the human host would suggest that the activities of these putative enzymes and the cohesin–dockerin-mediated complexes would be benign under most conditions. However, the pathogenic properties *C. perfringens* arise from the activity of numerous secreted toxins, including the carbohydrate-active μ-

toxin and the large sialidase, which correspond to the dockerin-containing CPE0191 and cohesin-containing CPF_1442/CPE1234, respectively (Roggentin *et al.*, 1988; Canard *et al.*, 1994; Shimizu *et al.*, 2002; Myers *et al.*, 2006). These toxins are proposed to degrade polysaccharide components of the extracellular matrix and cellular glycans, allowing for the spread of the invading bacteria and to enhance activities of the cell-surface-directed toxins (Rood, 1998). The recent observation that the μ-toxin and large sialidase have the ability to associate through an ultrahigh affinity cohesin–dockerin interaction illustrates a new noncellulosomal architecture and function: the formation of bimolecular glycoside hydrolase/toxin complexes (Adams *et al.*, 2008). Through such an association, these toxins could exert their pathogenic effects via a synergistic mechanism, whereby they can degrade common complex carbohydrates.

Cellulosome footprints in primitive Eukaryotes

Using the approach described in Table S1, we were also able to identify cohesin- and dockerin-like sequences in recently sequenced genomes of primitive eukaryotes. *Monosiga brevicollis* is a choanoflagellate, considered among the closest unicellular relatives of metazoans (Lang *et al.*, 2002). All of the relevant genes in this species were particularly large, coding for proteins several thousand amino-acid residues in length (two of these proteins comprise 10 000 residues); all contained both a predicted cohesin module and a dockerin module, and in three cases a pair of both modules. Furthermore, the cohesin and dockerin comprised tandem modular pairs separated by a short linker segment in each case. The annotations are not particularly informative, and the reasons for this remarkable distribution of cohesin and dockerin modules in these extraordinary predicted proteins await further study.

Analysis of two other recently sequenced genomes of primitive eukaryotes has revealed the existence of but a single dockerin module in each and the absence of any cohesin modules. The placozoan *Trichoplax adhaerens*, considered the simplest known ‘animal,’ and the model eukaryote *Tetrahymena thermophila* both contain a dockerin-like sequence in an uncharacterized protein. It is unknown what role these residual dockerin modules may play in these organisms. No other cohesin or dockerin modules have yet been detected in the currently accumulated genomes of other eukaryotic species. It is as if the utility of the cohesin–dockerin theme has been exhausted early on in the eukaryotes, in evolutionary terms. Perhaps, the dockerin module has been replaced by the related EF-hand motif of calcium-binding proteins, although the precise relationship is currently unknown. If such an evolutionary relationship exists, perhaps the dockerin module has abandoned its cohesin partner and has thus evolved to serve alternative functions in the higher eukaryotes.

One lingering question relates to the possible presence of cellulosomes and dockerin and/or cohesin components in anaerobic fungi. In this context, high-molecular-weight cellulosome-like complexes have been described in *Piromyces, Orpinomyces*, and *Neocalimastix* (Ali *et al.*, 1995; Li *et al.*, 1997; Fujino *et al.*, 1998). The catalytic components of these complexes were shown to contain a 40-amino-acid cysteine-rich, noncatalytic docking ‘domain’ (Steenbakkers *et al.*, 2001), frequently in two copies, interspaced by short linkers, and this domain or module has, in the past, been referred to as a fungal dockerin. However, the

fungal docking modules show no sequence homology to the bacterial dockerins, and recent solution structures of single and double dockerin modules from *Piromyces equi* indicate no coordinated calcium ions and no structural similarity to the bacterial dockerins (Raghothama *et al.*, 2001; Nagy *et al.*, 2007). In fact, the multienzyme cellulolytic complexes in the anaerobic fungi clearly fail to follow the conventional cellulosome paradigm, because recent evidence suggests that the dockerin module in *P. equi* recognizes and binds to the major glycosylated β-glucosidase via its oligosaccharide components (Nagy *et al.*, 2007). Thus, it appears that cohesin modules *per se* are not involved in the multienzyme complex of anaerobic fungi, and their 'dockerin' modules can actually be considered a new family of CBM, unrelated to the dockerin-like modules discussed in this article.

Phylogenetic relationships among the cohesins and dockerins

The sequences of cohesin and dockerin modules from representative archaeal, bacterial, and eukaryotic species were subjected to multiple sequence alignment analysis, and the evolutionary relationships among the different modules were viewed on the relevant phylograms (Fig. 3). It is clear from both the cohesin and the dockerin phylogenetic trees that these modules, associated with proteins from the Archaea and the Eukarya, are distributed among the different branches together in apparent disregard of conventional evolutionary theory and common lines of descent. The relative positions of specified archaeal and primitive eukaryotic cohesin and dockerin modules are thus intermixed with those of the bacteria and with each other.

These findings provide strong evidence for horizontal gene transfer of cohesin and dockerin modules from bacterial genes to both archaeal and eukaryotic genomes. Indeed, horizontal gene transfer between bacteria and *Methanosarcina* has been reported previously (Deppenmeier *et al.*, 2002); in view of the numerous cohesin- and (particularly) dockerin-coding sequences in genes of this archaeon, it is logical to propose that this process would be responsible for the appearance of these modular elements in archaeal proteins. Likewise, genes of apparent bacterial and archaeal origin have been documented in eukaryotic genomes (Morrison *et al.*, 2007; Gladyshev *et al.*, 2008). It is therefore appealing to consider that the process of horizontal gene transfer would also account for the observed cohesin and dockerin modules in the genomes of primitive eukaryotes.

Conclusions and future prospects

The field of cellulosome research has progressed enormously since the initial discovery of the multienzyme complex in *C. thermocellum*. The 'definitive' cellulosomal-like components – the cohesin and dockerin modules – have now been discovered in a wide variety of Bacteria, Archaea, and primitive Eukaryota that clearly fail to produce cellulosomes. Consequently, they would serve as modular protein appendages in roles uncharacteristic of the conventional polysaccharide-degrading cellulosome components. In this context, dockerin modules are attached to a variety of parent protein types that are inconsistent with cellulosome action. Moreover, the arrangement of the putative cohesin modules in the newly described parent proteins is inconsistent with that of the cellulosomal scaffoldins. These findings suggest the apparently broad involvement of cohesin and

dockerin modules and/or the cohesin–dockerin interaction in a wide variety of different biological and/or cellular processes that are unrelated to classic cellulosome function.

The wealth of the accumulated putative cohesin and dockerin sequences thus underscores the haste in their previous description as 'cellulosome-signature sequences,' as first implied by their initial discovery in the *A. fulgidus* genome. The broad distribution of these modules in nature, their purported noncellulosomal mode of action, and their possible appearance as single entities (cohesins without dockerins and vice versa) serve to raise the very basic question as to what are cohesin and dockerin modules, what functional role(s) they play, and whether they actually have to interact solely with one another. In truth, it might be premature at this stage to redefine these terms. We are now in a state of flux; the newly discovered modules are based mainly on bioinformatics analysis of the available genomic databases, which will undoubtedly be expanded greatly in the coming years. To date, the choice of organisms for genome sequencing projects has clearly been anthropocentric. The bioinformatics-based data presented in this communication can therefore be viewed as an initial compilation of cohesin and dockerin modules among the genomes from all three domains of life. In future work, the simple presence of these modules in a given protein should be verified experimentally by concrete biochemical analysis of the individual cohesin and dockerin modules, their functional interactions in the relevant species, and their contribution(s) to the lifestyle of the organism and its place in nature.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Parts of the research described in this article were supported by the Israel Science Foundation (Grant nos. 422/05 and 159/07), a Grant from the Alternative Energy Research Initiative (Weizmann Institute), and by Grants from the United States –Israel Binational Science Foundation (BSF), Jerusalem, Israel, and by a Canadian Institutes of Health Research operating Grant (MOP-77776). S.P.S. is a Canadian Institutes of Health Research New Investigator. E.A.B. is the incumbent of The Maynard I. and Elaine Wishner Chair of Bio-organic Chemistry at the Weizmann Institute of Science.

References

Adams JJ, Webb BA, Spencer HL, Smith SP. Structural characterization of type II dockerin module from the cellulosome of *Clostridium thermocellum*: calcium-induced effects on conformation and target recognition. Biochemistry. 2005; 44:2173–2182. [PubMed: 15697243]

Adams JJ, Pal G, Jia Z, Smith SP. Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex. P Natl Acad Sci USA. 2006; 103:305–310.

Adams JJ, Gregg K, Bayer EA, Boraston AB, Smith SP. Structural basis for a novel mode of *Clostridium perfringens* toxin complex formation. P Natl Acad Sci USA. 2008; 105:12194–12199.

Alber O, Noach I, Lamed R, Shimon LJW, Bayer EA, Frolow F. Preliminary X-ray characterization of a novel type of anchoring cohesin from the cellulosome of *Ruminococcus flavefaciens*. Acta Crystallogr. 2008; F64:77–80.

Ali BR, Zhou L, Graves FM, Freedman RB, Black GW, Gilbert HJ, Hazelwood GP. Cellulases and hemicellulases of the anaerobic fungus *Piromyces* constitute a multiprotein cellulose-binding complex and are encoded by multigene families. FEMS Microbiol Lett. 1995; 125:15–21. [PubMed: 7867916]

Bayer EA, Kenig R, Lamed R. Adherence of *Clostridium thermocellum* to cellulose. J Bacteriol. 1983; 156:818–827. [PubMed: 6630152]

Bayer EA, Chanzy H, Lamed R, Shoham Y. Cellulose, cellulases and cellulosomes. Curr Opin Struc Biol. 1998; 8:548–557.

Bayer EA, Coutinho PM, Henrissat B. Cellulosome-like sequences in *Archaeoglobus fulgidus*: an enigmatic vestige of cohesin and dockerin domains. FEBS Lett. 1999; 463:277–280. [PubMed: 10606737]

Bayer EA, Belaich J-P, Shoham Y, Lamed R. The cellulosomes: multi-enzyme machines for degradation of plant cell wall polysaccharides. Annu Rev Microbiol. 2004; 58:521–554. [PubMed: 15487947]

Canard B, Garnier T, Saint-Joanis B, Cole ST. Molecular genetic analysis of the *nagH* gene encoding a hyaluronidase of *Clostridium perfringens*. Mol Gen Genet. 1994; 243:215–224. [PubMed: 8177218]

Carvalho AL, Dias FM, Prates JA, et al. Cellulosome assembly revealed by the crystal structure of the cohesin–dockerin complex. P Natl Acad Sci USA. 2003; 100:13809–13814.

Carvalho AL, Goyal A, Prates JA, et al. The family 11 carbohydrate-binding module of *Clostridium thermocellum* Lic26A-Cel5E accommodates β-1,4- and β-1,3-1,4-mixed linked glucans at a single binding site. J Biol Chem. 2004; 279:34785–34793. [PubMed: 15192099]

Carvalho AL, Pires VM, Gloster TM, et al. Insights into the structural determinants of cohesin-dockerin specificity revealed by the crystal structure of the type II cohesin from *Clostridium thermocellum* SdbA. J Mol Biol. 2005; 349:909–915. [PubMed: 15913653]

Carvalho AL, Dias FMV, Nagy T, et al. Evidence for a dual binding mode of dockerin modules to cohesins. P Natl Acad Sci USA. 2007; 104:3089–3094.

Chauvaux S, Béguin P, Aubert J-P, Bhat KM, Gow LA, Wood TM, Bairoch A. Calcium-binding affinity and calcium-enhanced activity of *Clostridium thermocellum* endoglucanase D. Biochem J. 1990; 265:261–265. [PubMed: 2302168]

Chitayat S, Adams JJ, Furness HS, Bayer EA, Smith SP. The solution structure of the C-terminal modular pair from *Clostridium perfringens* reveals a non-cellulosomal dockerin module. J Mol Biol. 2008a; 381:1202–1212. [PubMed: 18602403]

Chitayat S, Gregg K, Adams JJ, Ficko-Blean E, Bayer EA, Boraston AB, Smith SP. Three-dimensional structure of a putative non-cellulosomal cohesin module from a *Clostridium perfringens* family 84 glycoside hydrolase. J Mol Biol. 2008b; 375:20–28. [PubMed: 17999932]

Chow V, Nong G, Preston JF. Structure, function, and regulation of the aldouronate utilization gene cluster from *Paenibacillus* sp. strain JDR-2. J Bacteriol. 2007; 189:8863–8870. [PubMed: 17921311]

Coutinho PM, Henrissat B. 1999Carbohydrate-active enZYmes server (CAZy Website). 10 September 2008, last date accessedhttp://afmb.cnrs-mrs.fr/~pedro/CAZY/db.html

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–1190. [PubMed: 15173120]

Demain AL, Newcomb M, Wu JH. Cellulase, clostridia, and ethanol. Microbiol Mol Biol R. 2005; 69:124–154.

Deppenmeier U, Johann A, Hartsch T, et al. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archae. J Mol Microb Biotech. 2002; 4:453–461.

Ding S-Y, Bayer EA, Steiner D, Shoham Y, Lamed R. A novel cellulosomal scaffoldin from *Acetivibrio cellulolyticus* that contains a family-9 glycosyl hydrolase. J Bacteriol. 1999; 181:6720–6729. [PubMed: 10542174]

Ding S-Y, Bayer EA, Steiner D, Shoham Y, Lamed R. A scaffoldin of the *Bacteroides cellulosolvens* cellulosome that contains 11 type II cohesins. J Bacteriol. 2000; 182:4915–4925. [PubMed: 10940036]

Ding S-Y, Rincon MT, Lamed R, et al. Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens*. J Bacteriol. 2001; 183:1945–1953. [PubMed: 11222592]

Doi RH, Kosugi A. Cellulosomes: plant-cell-wall-degrading enzyme complexes. Nat Rev Microbiol. 2004; 2:541–551. [PubMed: 15197390]

Faure E, Belaich A, Bagnara C, Gaudin C, Belaich J-P. Sequence analysis of the *Clostridium cellulolyticum* endoglucanase-A-encoding gene *celCCA*. Gene. 1989; 84:39–46. [PubMed: 2558058]

Ficko-Blean E, Boraston AB. The interaction of carbohydrate-binding module from a clostridium perfringens *N*-acetyl-β-hexosaminidase with its carbohydrate receptor. J Biol Chem. 2006; 281:37748–37757. [PubMed: 16990278]

Fierobe H-P, Pagès S, Belaich A, Champ S, Lexa D, Belaich J-P. Cellulosome from *Clostridium cellulolyticum*: molecular study of the dockerin/cohesin interaction. Biochemistry. 1999; 38:12822–12832. [PubMed: 10504252]

Fujino Y, Ogata K, Nagamine T, Ushida K. Cloning, sequencing, and expression of an endoglucanase gene from the rumen anaerobic fungus *Neocallimastix frontalis* MCH3. Biosci Biotech Bioch. 1998; 62:1795–1798.

Giallo J, Gaudin C, Belaich J-P, Petitdemange E, Caillet-Mangin F. Metabolism of glucose and cellobiose by cellulolytic mesophilic *Clostridium* sp. strain H10. Appl Environ Microb. 1983; 45:843–849.

Giallo J, Gaudin C, Belaich J-P. Metabolism and solubilization of cellulose by *Clostridium cellulolyticum* H10. Appl Environ Microb. 1985; 49:1216–1221.

Gilbert HJ. Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. Mol Microbiol. 2007; 63:1568–1576. [PubMed: 17367380]

Giuliano C, Asther M, Khan AW. Comparative degradation of cellulose and sugar formation by three newly isolated mesophilic anaerobes and *Clostridium thermocellum*. Biotechnol Lett. 1983; 5:395–398.

Gladyshev EA, Meselson M, Arkhipova IR. Massive horizontal gene transfer in bdelloid rotifers. Science. 2008; 320:1210–1213. [PubMed: 18511688]

Haimovitz R, Barak Y, Morag E, Voronov-Goldman M, Lamed R, Bayer EA. Cohesin–dockerin microarray: diverse specificities between two complementary families of interacting protein modules. Proteomics. 2008; 8:968–979. [PubMed: 18219699]

Handelsman T, Barak Y, Nakar D, Mechaly A, Lamed R, Shoham Y, Bayer EA. Cohesin–dockerin interaction in cellulosome assembly: a single Asp-to-Asn mutation disrupts high-affinity cohesin–dockerin binding. FEBS Lett. 2004; 572:195–200. [PubMed: 15304347]

Hungate RE. Microorganisms in the rumen of cattle fed a constant ration. Can J Microbiol. 1957; 3:289–311. [PubMed: 13413736]

Jindou S, Kajino T, Inagaki M, et al. Interaction between a type-II dockerin domain and a type-II cohesin domain from *Clostridium thermocellum* cellulosome. Biosci Biotech Bioch. 2004a; 68:924–926.

Jindou S, Souda A, Karita S, et al. Cohesin/dockerin interactions within and between *Clostridium josui* and *Clostridium thermocellum*: binding selectivity between cognate dockerin and cohesin domains and species specificity. J Biol Chem. 2004b; 279:9867–9874. [PubMed: 14688277]

Kakiuchi M, Isui A, Suzuki K, et al. Cloning and DNA sequencing of the genes encoding *Clostridium josui* scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. J Bacteriol. 1998; 180:4303–4308. [PubMed: 9696784]

Kirby J, Martin JC, Daniel AS, Flint HJ. Dockerin-like sequences in cellulases and xylanases from the rumen cellulolytic bacterium *Ruminococcus flavefaciens*. FEMS Microbiol Lett. 1997; 149:213–219. [PubMed: 9141662]

Lamed R, Setter E, Bayer EA. Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*. J Bacteriol. 1983a; 156:828–836. [PubMed: 6195146]

Lamed R, Setter E, Kenig R, Bayer EA. The cellulosome – a discrete cell surface organelle of *Clostridium thermocellum* which exhibits separate antigenic, cellulose-binding and various cellulolytic activities. Biotech Bioeng Symp. 1983b; 13:163–181.

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G. The closest unicellular relatives of animals. Curr Biol. 2002; 12:1773–1778. [PubMed: 12401173]

Leibovitz E, Béguin P. A new type of cohesin domain that specifically binds the dockerin domain of the *Clostridium thermocellum* cellulosome-integrating protein CipA. J Bacteriol. 1996; 178:3077–3084. [PubMed: 8655483]

Leschine SB, Canale-Parola E. Mesophilic cellulolytic clostridia from freshwater environments. Appl Environ Microb. 1983; 46:728–737.

Li X, Chen H, Ljungdahl L. Two cellulases, CelA and CelC, from the polycentric anaerobic fungus *Orpinomyces* strain PC-2 contain N-terminal docking domains for a cellulase–hemicellulase complex. Appl Environ Microb. 1997; 63:4721–4728.

Lytle B, Volkman BF, Westler WM, Wu JHD. Secondary structure and calcium-induced folding of the *Clostridium thermocellum* dockerin domain determined by NMR spectroscopy. Arch Biochem Biophys. 2000; 379:237–244. [PubMed: 10898940]

Lytle BL, Volkman BF, Westler WM, Heckman MP, Wu JHD. Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain. J Mol Biol. 2001; 307:745–753. [PubMed: 11273698]

McBee RH. The characteristics of *Clostridium thermocellum*. J Bacteriol. 1954; 67:505–506. [PubMed: 13152068]

Mechaly A, Yaron S, Lamed R, et al. Cohesin–dockerin recognition in cellulosome assembly: experiment versus hypothesis. Proteins. 2000; 39:170–177. [PubMed: 10737938]

Mechaly A, Fierobe H-P, Belaich A, Belaich J-P, Lamed R, Shoham Y, Bayer EA. Cohesin–dockerin interaction in cellulosome assembly: a single hydroxyl group of a dockerin domain distinguishes between non-recognition and high-affinity recognition. J Biol Chem. 2001; 276:9883–9888. Erratum 19678. [PubMed: 11148206]

Morrison HG, McArthur AG, Gillin FD, et al. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science. 2007; 317:1921–1926. [PubMed: 17901334]

Myers GS, Rasko DA, Cheung JK, et al. Skewed genomic variability in strains of the toxigenic bacterial pathogen. *Clostridium perfringens*. Genome Res. 2006; 16:1031–1040. [PubMed: 16825665]

Nagy T, Tunnicliffe RB, Higgins LD, Walters C, Gilbert HJ, Williamson MP. Characterization of a double dockerin from the cellulosome of the anaerobic fungus *Piromyces equi*. J Mol Biol. 2007; 373:612–622. [PubMed: 17869267]

Nakar D, Handelsman T, Shoham Y, et al. Pinpoint mapping of recognition residues on the cohesin surface by progressive homologue swapping. J Biol Chem. 2004; 279:42881–42888. [PubMed: 15292269]

Noach I, Frolow F, Jakoby H, Rosenheck S, Shimon LJW, Lamed R, Bayer EA. Crystal structure of a type-II cohesin module from the *Bacteroides cellulosolvens* cellulosome reveals novel and distinctive secondary structural elements. J Mol Biol. 2005; 348:1–12. [PubMed: 15808849]

Nolling J, Breton G, Omelchenko MV, et al. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol. 2001; 183:4823–4838. [PubMed: 11466286]

Ohara H, Karita S, Kimura T, Sakka K, Ohmiya K. Characterization of the cellulolytic complex (cellulosome) from *Ruminococcus albus*. Biosci Biotech Bioch. 2000; 64:254–260.

Pagès S, Belaich A, Belaich J-P, Morag E, Lamed R, Shoham Y, Bayer EA. Species-specificity of the cohesin–dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain. Proteins. 1997; 29:517–527. [PubMed: 9408948]

Patel GB, Khan AW, Agnew BJ, Colvin JR. Isolation and characterization of an anaerobic cellulolytic microorganism, *Acetivibrio cellulolyticus*, gen. nov., sp. nov. Int J Syst Bacteriol. 1980; 30:179–185.

Pinheiro BA, Proctor MR, Martinez-Fleites CC, et al. The *Clostridium cellulolyticum* dockerin displays a dual binding mode for its cohesin partner. J Biol Chem. 2008; 283:18422–18430. [PubMed: 18445585]

Pohlschröder M, Canale-Parola E, Leschine SB. Ultrastructural diversity of the cellulase complexes of *Clostridium papyrosolvens* C7. J Bacteriol. 1995; 177:6625–6629. [PubMed: 7592442]

Raghothama S, Eberhardt RY, Simpson P, et al. Characterization of a cellulosome dockerin domain from the anaerobic fungus *Piromyces equi*. Nat Struct Biol. 2001; 8:775–778. [PubMed: 11524680]

Rigden DJ, Galperin MY. The DxDxDG motif for calcium binding: multiple structural contexts and implications for evolution. J Mol Biol. 2004; 343:971–984. [PubMed: 15476814]

Rincon MT, Ding S-Y, McCrae SI, et al. Novel organization and divergent dockerin specificities in the cellulosome system of *Ruminococcus flavefaciens*. J Bacteriol. 2003; 185:703–713. [PubMed: 12533446]

Rincon MT, Martin JC, Aurilia V, et al. ScaC, an adaptor protein carrying a novel cohesin that expands the dockerin-binding repertoire of the *Ruminococcus flavefaciens* 17 cellulosome. J Bacteriol. 2004; 186:2576–2585. [PubMed: 15090497]

Rincon MT, Cepeljnik T, Martin JC, Lamed R, Barak Y, Bayer EA, Flint HJ. Unconventional mode of attachment of the *Ruminococcus flavefaciens* cellulosome to the cell surface. J Bacteriol. 2005; 187:7569–7578. [PubMed: 16267281]

Rincon MT, Cepeljnik T, Martin JC, Barak Y, Lamed R, Bayer EA, Flint HJ. A novel cell surface-anchored cellulose-binding protein encoded by the *sca* gene cluster of *Ruminococcus flavefaciens*. J Bacteriol. 2007; 189:4774–7283. [PubMed: 17468247]

Roggentin P, Rothe B, Lottspeich F, Schauer R. Cloning and sequencing of a *Clostridium perfringens* sialidase gene. FEBS Lett. 1988; 238:31–34. [PubMed: 2901987]

Rood JI. Virulence genes of *Clostridium perfringens*. Annu Rev Microbiol. 1998; 52:333–360. [PubMed: 9891801]

Rood, JI., Lyras, DL. Clostridial genetics. In: Fischetti, VA.Novick, RP.Ferretti, JJ.Portnoy, DA., Rood, JI., editors. Gram-Positive Pathogens. 2. ASM Press; Washington, DC: 2006. p. 672-687.

Sabathe F, Belaich A, Soucaille P. Characterization of the cellulolytic complex (cellulosome) of *Clostridium acetobutylicum*. FEMS Microbiol Lett. 2002; 217:15–22. [PubMed: 12445640]

Salamitou S, Tokatlidis K, Béguin P, Aubert J-P. Involvement of separate domains of the cellulosomal protein S1 of *Clostridium thermocellum* in binding to cellulose and in anchoring of catalytic subunits to the cellulosome. FEBS Lett. 1992; 304:89–92. [PubMed: 1618304]

Schaeffer F, Matuschek M, Guglielmi G, Miras I, Alzari PM, Béguin P. Duplicated dockerin subdomains of *Clostridium thermocellum* endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity. Biochemistry. 2002; 41:2106–2114. [PubMed: 11841200]

Schwarz WH. The cellulosome and cellulose degradation by anaerobic bacteria. Appl Microbiol Biot. 2001; 56:634–649.

Shimizu T, Ohtani K, Hirakawa H, et al. Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. P Natl Acad Sci USA. 2002; 99:996–1001.

Shimon LJW, Bayer EA, Morag E, Lamed R, Yaron S, Shoham Y, Frolow F. A cohesin domain from *Clostridium thermocellum*: the crystal structure provides new insights into cellulosome assembly. Structure. 1997; 5:381–390. [PubMed: 9083107]

Shoham Y, Lamed R, Bayer EA. The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. Trends Microbiol. 1999; 7:275–281. [PubMed: 10390637]

Shoseyov O, Takagi M, Goldstein MA, Doi RH. Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein A. P Natl Acad Sci USA. 1992; 89:3483–3487.

Sijpesteijn AK. On *Ruminococcus flavefaciens*, a cellulose-decomposing bacterium from the rumen of sheep and cattle. J Gen Microbiol. 1951; 5(suppl):869–879. [PubMed: 14908024]

Sleat R, Mah RA, Robinson R. Isolation and characterization of an anaerobic, cellulolytic bacterium, *Clostridium cellulovorans*, sp. nov. Appl Environ Microb. 1984; 48:88–93.

Smith LD, Gardner MV. The occurrence of vegetative cells of *Clostridium perfringens* in soil. J Bacteriol. 1949; 58:407–408.

Songer, JG. Clostridial diseases of animals. In: Rood, JI.McClane, BA.Songer, JG., Titball, RW., editors. The Clostridia: Molecular Biology and Pathogenesis. Academic Press; San Diego: 1997. p. 153-182.

Steenbakkers PJ, Li XL, Ximenes EA, Arts JG, Chen H, Ljungdahl LG, Op Den Camp HJ. Noncatalytic docking domains of cellulosomes of anaerobic fungi. J Bacteriol. 2001; 183:5325–5333. [PubMed: 11514516]

St John FJ, Rice JD, Preston JF. *Paenibacillus* sp. strain JDR-2 and XynA1: a novel system for methylglucuronoxylan utilization. Appl Environ Microb. 2006; 72:1496–1506.

Sukhumavasi J, Ohmiya K, Shimizu S, Ueno K. *Clostridiumn josui* sp. nov., a cellulolytic, moderate thermophilic species from Thai compost. Int J Syst Bacteriol. 1988; 38:179–182.

Tavares GA, Béguin P, Alzari PM. The crystal structure of a type I cohesin domain at 1.7 Å resolution. J Mol Biol. 1997; 273:701–713. [PubMed: 9402065]

Tokatlidis K, Salamitou S, Béguin P, Dhurjati P, Aubert J-P. Interaction of the duplicated segment carried by *Clostridium thermocellum* cellulases with cellulosome components. FEBS Lett. 1991; 291:185–188. [PubMed: 1936262]

Volkman, BF., Lytle, BL., Wu, JHD. Dockerin domains. In: Messerschmidt, A.Bode, W., Cygler, M., editors. Handbook of Metaloproteins. Vol. 3. John Wiley & Sons Ltd; Chichester, UK: 2004. p. 617-628.

Yaron S, Morag E, Bayer EA, Lamed R, Shoham Y. Expression, purification and subunit-binding properties of cohesins 2 and 3 of the *Clostridium thermocellum* cellulosome. FEBS Lett. 1995; 360:121–124. [PubMed: 7875315]

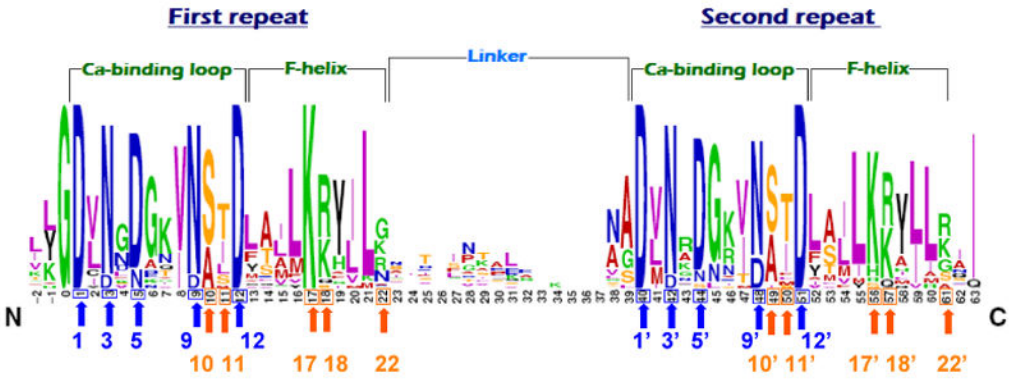

Fig. 1.

Sequence conservation of the dockerin modules from classical cellulosome-producing bacteria: *Clostridium thermocellum, Clostridium cellulolyticum,* and *Clostridium cellulovorans*. The two repeats consist of a calcium-binding loop and an 'F-helix.' Calcium-coordinating residues, in positions 1, 3, 5, 9, and 12 of each repeat, are designated by blue arrows. Putative cohesin-specificity residues, in positions 10, 11, 17, 18, and 22, are designated by orange arrows. The figure was generated using the WEBLOGO application (Crooks *et al.,* 2004).

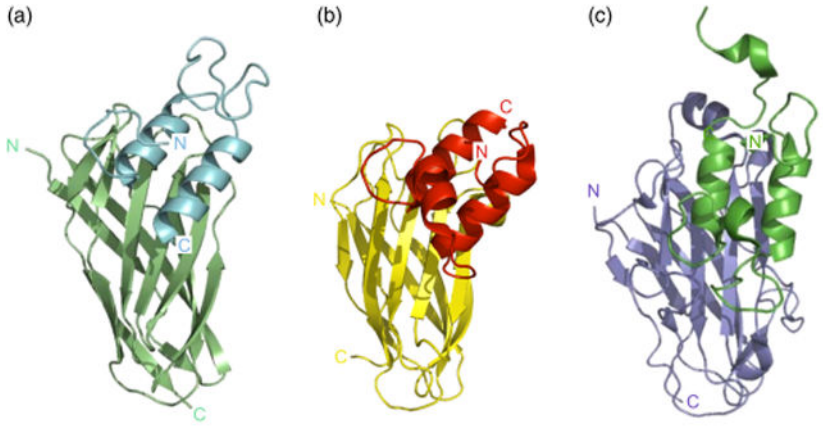

Fig. 2.

Comparison of a noncellulosomal cohesin–dockerin interaction to cellulosomal type-I and type-II cohesin–dockerin interactions. Ribbon representations of (a) *Clostridium perfringens* μ-toxin dockerin (light blue) on *Cp*GH84C cohesin (light green) (Adams *et al.*, 2008); (b) *C. thermocellum* Xyn-10B type-I dockerin (red) on CipA2 type-I cohesin (yellow) (Carvalho *et al.*, 2003); and (c) *Clostridium thermocellum* CipA type-II dockerin (emerald green) on SdbA type-II cohesin (slate blue) (Adams *et al.*, 2006). The N- and C-termini of each module are labeled and colored accordingly.

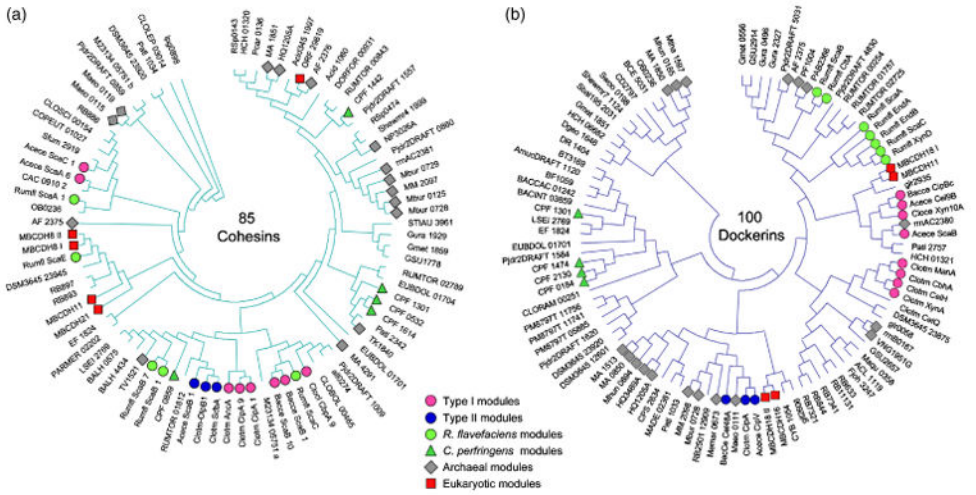

Fig. 3.

Phylogenetic distribution of representative cohesins (a) and dockerins (b) in the three domains of life. In the phylogenetic trees, 85 cohesin-like modules and 100 dockerin-like modules, derived from deduced amino-acid sequences of representative species, were aligned using the CLUSTALW program, which served to create an unrooted tree using the MEGA 4.1 program. Representative archaeal and eukaryotic cohesin and dockerin modules are labeled according to the designated key to facilitate comparison with those of representative cellulosomal modules (i.e. type-I, type-II, and *Ruminococcus flavefaciens* modules) and the recently described noncellulosomal modules of *Clostridium perfringens;* all other bacteria-derived modules are unlabeled on the phylogenetic trees. The terminology for cohesins and dockerins and the accession codes for their parent proteins are given in Tables S2 and S3, respectively.

Table 1

Confirmed cellulosome-producing bacteria

Bacterium	Original source	Optimum temperature*	Genome sequence	References
Acetivibrio cellulolyticus	Sewage sludge	m	No	Patel *et al.* (1980); Ding *et al.* (1999)
Bacteroides cellulosolvens	Sewage sludge	m	No	Giuliano *et al.* (1983); Ding *et al.* (2000)
Clostridium acetobutylicum	Soil	m	Complete	Nolling *et al.* (2001); Sabathe *et al.* (2002)
Clostridium cellulovorans	Woody biomass	m	No	Sleat *et al.* (1984); Shoseyov *et al.* (1992)
Clostridium cellulolyticum	Decayed grass	m	Draft	Giallo *et al.* (1985); Faure *et al.* (1989)
Clostridium josui	Compost	m	No	Sukhumavasi *et al.* (1988); Kakiuchi *et al.* (1998)
Clostridium papyrosolvens	Freshwater swamp	m	No	Leschine & Canale-Parola (1983); Pohlschröder *et al.* (1995)
Clostridium thermocellum	Soil, sewage sludge, horse manure, hot spring	t	Complete	McBee (1954); Lamed *et al.* (1983a, b)
Ruminococcus albus	Rumen	m	Draft	Hungate (1957); Ohara *et al.* (2000)
Ruminococcus flavefaciens	Rumen	m	Draft	Sijpesteijn (1951); Kirby *et al.* (1997)

*Optimal temperature for growth.

t, thermophilic (> 55 °C); m, mesophilic.

Table 2

Summary of prokaryotic genome mining for noncellulosomal dockerin and cohesin modules

ARCHEA	**Summary: 46 individual species genomes of Archaea: 18 genomes encoding cohesins and/or dockerins (c. 39%); the distribution of dockerins and cohesins among only Euryarchaeota (31 genomes) is c. 58%**			
General taxonomy	**Species name***	**Cohesin-containing ORFs†**	**Dockerin-containing ORFs**	**Comments and notes**
Euryarchaeota	***Archaeoglobus fulgidus* DSM 4304**	**2**	**1**	Cohesin+dockerin **ORF; 2-gene cluster. Cohesin and dockerin in single protein (1)**
	***Haloarcula* (*Halobacterium*) *marismortui* ATCC 43049**	**1**	**4**	Subtilisin homolog (**dockerin**), UNK protein (**cohesin**)
	***Halobacterium salinarium* NRC-1/ *Halobacterium halobium* NRC-1**	ND	**1**	Subtilisin homolog
	***Haloquadratum walsbyi* DSM 16790**	**1**	**5**	Cell surface glycoprotein (**cohesin**), subtilisin homolog (**dockerin**), putative cell surface glycoprotein (**dockerin**). **Cohesin and dockerin in single protein** (1)
	***Methanococcoides burtonii* DSM 6242**	**3**	**1**	UNK protein; **cohesin+dockerin** ORF. **Cohesin and dockerin in single protein** (1)
	***Methanococcus aeolicus* Nankai-3**	**3**	**2**	**Dockerin- and cohesin-encoding genes are clustered**. APHP domain protein precursor (**cohesin**); PKD domain containing a protein precursor (**cohesin**), ABC-type Fe^{3+}-hydroxamate transport system periplasmic component-like protein (N-terminal **dockerin**)
	***Methanococcus vannielii* SB** (two other strains)	**2**	**1**	Periplasmic binding protein (N-terminal **dockerin**)
	Methanococcus voltae A3	**2**	**2**	**Dockerin- and cohesin-encoding genes are clustered.** Putative iron transport system substrate-binding protein (N-terminal **dockerin**), APHP domain protein precursor (**cohesin**)
	***Methanoculleus marisnigri* JR1**	**(2)**	**1**	Peptidase M28 precursor (**cohesin**), UNK protein (**dockerin**)
	***Methanosaeta thermophila* PT**	ND	**4**	Periplasmic binding protein (N-terminal **dockerins**)
	***Methanosarcina acetivorans* CA2**	**2**	**27**	**Dockerin- and cohesin-encoding genes are clustered.** Putative cell surface proteins of unknown function (**cohesin, dockerin**), iron(III) ABC transporter, solute-binding protein **(dockerin),** UNK protein **(cohesin). Cohesin and dockerin in a single protein** (1)
	***Methanosarcina mazei* (*Methanosarcina frisia*) Go1**	**2**	**1**	**Cohesin+dockerin ORF. Cohesin and dockerin in a single protein (1)**
	***Methanospirillum hungatei* JF-1**	ND	**7**	Periplasmic binding protein precursor (N-terminal **dockerin**), peptidase, periplasmic copper-binding protein, UNK proteins
	***Natronomonas pharaonis* DSM 2160**	**1**	ND	UNK protein (a **cohesin** module is similar to that in *Geobacter* spp. and Acidobacteria)

ARCHEA	**Summary: 46 individual species genomes of Archaea: 18 genomes encoding cohesins and/or dockerins (c. 39%); the distribution of dockerins and cohesins among only Euryarchaeota (31 genomes) is c. 58%**			
General taxonomy	**Species name***	**Cohesin-containing ORFs†**	**Dockerin-containing ORFs**	**Comments and notes**
	***Pyrococcus abyssi* Orsay/GE5**	ND	**1**	Putative alkaline phosphatase IV
	***Pyrococcus furiosus* Vc1/ATCC 43587**	**(1)**	**1**	Alkaline phosphatase IV **(dockerin),** UNK protein **(cohesin)**
	***Pyrococcus kodakaraensis* (*Thermococcus kodakaraensis*) KOD1**	**(1)**	ND	Cobalt-activating carboxypeptidase, M32 family
	***Thermoplasma volcanium* ATCC 51530**	**(1)**	ND	Putative peptidase (cohesin is similar to *A. cellulolyticus* ScaA cohesins)
Uncultured archaeons	At least three independent genomes	**1**	**4**	Chitosanase–glucanase like **(3× CBM32, dockerin)**; putative peptidase **(cohesin)**

BACTERIA	***c*. 500* individual species complete and incomplete genomes: 74 genomes encoding cohesins and/or dockerins (*c*. 14%)** **• Genomes encoding 'classic' cellulosome proteins are not included (e.g. *C. thermocellum, C. cellulolyticum, C. cellulovorans, R. flavefaciens, R. albus*, etc.)**			
General taxonomy	**Species name***	**Cohesin-containing ORFs†**	**Dockerin-containing ORFs**	**Comments and notes**
Firmicutes: Mollicutes	***Acholeplasma laidlawii* PG-8A**	ND	**1**	UNK protein
	Eubacterium dolichum DSM 3991	**2**	**1**	**GH31 (3×CBM32, cohesin, dockerin**), **GH31 (3× CBM32, cohesin). Cohesin and dockerin in a single protein** (1)
Firmicutes: Bacillales	***Bacillus cereus* G9241** (and other strains)	**2**	**2**	Putative peptidases (**cohesin, dockerin**). **Cohesin and dockerin in the same protein** (2)
	***Bacillus thuringiensis* Al Hakam**	**2**	**2**	Putative proteases (**cohesin, dockerin**). **Cohesin and dockerin in a single protein** (2)
	***Bacillus* sp. B14905**	ND	**1**	Possible microbial serine proteinase
	***Bacillus* sp. GL1**	**1**	ND	Gellan lyase precursor
	***Enterococcus faecalis* V583**	**1**	**1**	**GH31 (CBM32, cohesin, dockerin). Cohesin and dockerin in a single protein** (1)
	***Lactobacillus casei* ATCC 334**	**1**	**1**	**GH31 (CBM32, cohesin, dockerin). Cohesin and dockerin in a single protein** (1)
	***Lactobacillus johnsonii* NCC 533**	**(1)**	ND	Putative anchoring protein (contains a sortase cleavage site)
	***Oceanobacillus iheyensis* HTE831**	**2**	**2**	Putative serine proteinase (**cohesin, dockerin**), UNK proteins. **Cohesin and dockerin in a single protein** (1)
	Paenibacillus sp. JDR-2	**15**	**8**	**20 ORFs**, **3× GH3 (CBM6, dockerin), GH6 (cohesin);** amidase **(cohesin, dockerin)**; SLH proteins (**cohesins**); some genes are organized in putative operons. **Cohesin and dockerin in a single protein** (2)

BACTERIA	***c.* 500* individual species complete and incomplete genomes: 74 genomes encoding cohesins and/or dockerins (*c.* 14%)** **• Genomes encoding 'classic' cellulosome proteins are not included (e.g. *C. thermocellum, C. cellulolyticum, C. cellulovorans, R. flavefaciens, R. albus*, etc.)**			
General taxonomy	**Species name***	**Cohesin-containing ORFs†**	**Dockerin-containing ORFs**	**Comments and notes**
Firmicutes: Clostridia	*Clostridium* sp. L2-50	**2**	**1**	Beta-N-acetylglucosaminidase **(dockerin)**, extracellular UNK protein **(cohesin), operon**
	Clostridium bolteae ATCC BAA-613	**1**	ND	UNK protein (**cohesin**)
	Clostridium butyricum 5521	ND	**1**	UNK protein (**dockerin**)
	***Clostridium difficile* 630** (and at least 5 other strains)	ND+**(1)**	**1**	UNK protein (**dockerin**), cell surface protein or S-layer protein (**cohesin**)
	***Clostridium leptum* DSM 753**	**1**	**1**	**Genes encoding dockerin and cohesin are in an operon**
	***Clostridium perfringens*†**	**7+(1)**	**5**	**GH families: 3, 20, 29, 31 (CBM32, dockerin, cohesin), 33, 84, 89, and 95. Putative cohesin and dockerin in a single protein (1)**
	***Clostridium phytofermentans* ISDg**	**1**	ND	UNK extracellular protein
	Clostridium ramosum DSM 1402	ND	**1**	UNK extracellular protein
	Clostridium scindens ATCC 35704	**1**	ND	UNK extracellular protein
	***Clostridium spiroforme* DSM 1552**	ND	**1**	**GH31** (**2× CBM32, dockerin**)
	Coprococcus eutactus ATCC 27759	**1**	ND	UNK extracellular protein
	***Dorea formicigenerans* ATCC 27755**	**1**	ND	UNK extracellular protein
	***Dorea longicatena* DSM 13814**	**1**	ND	UNK extracellular protein
	Epulopiscium sp. 'N.t. morphotype B'	ND	**1**	UNK extracellular protein
	Eubacterium ventriosum ATCC 27560	ND	**1**	Putative peptidase
	Ruminococcus gnavus ATCC 29149	**1**	ND	UNK function
	Ruminococcus torques ATCC 27756	**3**	**7**	**GH31 protein** (**cohesin**, **2× CBM32**, FN3, etc.); LRR repeat ORFs (**dockerins**)
Deinococcus-Thermus group Bacteroidetes	***Deinococcus radiodurans* R1**	ND	**1**	UNK protein
	***Deinococcus geothermalis* DSM 11300**	ND	**1**	UNK protein
	***Bacteroides caccae* ATCC 43185**	**2**	**2**	**Two GH31 proteins** (**CBM32, cohesin, dockerin**). **Cohesin and dockerin in a single protein** (2)
	***Bacteroides fragilis* NCTC 9343** (and strain YCH46)	**1+(1)**	**1**	**GH31 protein** (**CBM32, cohesin, dockerin**). **Cohesin and dockerin in single protein** (1)
	Bacteroides intestinalis DSM 17393	**1**	**1**	**GH31 protein** (**CBM32, cohesin, dockerin**). **Cohesin and dockerin in a single protein** (1)
	***Bacteroides thetaiotaomicron* VPI-5482**	**1**	**1**	**GH31 protein** (**CBM32, cohesin, dockerin**)

BACTERIA	*c.* 500* individual species complete and incomplete genomes: 74 genomes encoding cohesins and/or dockerins (*c.* 14%) • Genomes encoding 'classic' cellulosome proteins are not included (e.g. *C. thermocellum, C. cellulolyticum, C. cellulovorans, R. flavefaciens, R. albus*, etc.)			
General taxonomy	**Species name***	**Cohesin-containing ORFs†**	**Dockerin-containing ORFs**	**Comments and notes**
	***Flavobacterium johnsoniae* UW101**	**(1)**	**1**	LRR repeat protein (**dockerin**), UNK protein **(cohesin). Cohesin and dockerin in a single protein (1)**
	Microscilla marina ATCC 23134	**2(2× 2)**	ND	**Each putative protein carries 2 cohesin-like modules**; a putative strictosidine synthase
	***Parabacteroides distasonis* ATCC 8503**	ND	**1**	LRR repeat protein
	Parabacteroides merdae ATCC 43184	**1**	**1**	**GH31 protein** (**CBM32, cohesin, dockerin**). **Cohesin and dockerin in a single protein**
	Robiginitalea biformata HTCC2501	ND	**1**	Ring canal kelch motif protein (COG3055)
Planctomycetes	*Blastopirellula marina* DSM 3645	**1+(1)**	**7**	Surface-associated protein CshA precursor (**dockerin, cohesin**), putative α-galactosidase **(dockerin)**, two alkaline phosphatases (each contains a **dockerin**), subtilisin-like serine protease, etc. **Cohesin and dockerin in a single protein** (1)
	Planctomyces maris DSM 8797	ND	**12**	Almost all ORFs contain a 'motif' of Zn-dependent protease (PS00142; zinc protease)
	***Rhodopirellula baltica* SH1**	**2+(1)**	**23**	Putative **xylanase (dockerin)**, vanadium chloroperoxidase **(dockerin),** alkaline phosphatase **(dockerin),** extracellular nuclease, peptidase, fat protein-possibly involved in cell–cell attachment **(cohesin),** probable fibrinogen-binding protein homolog **(cohesin)**
Verrucomicrobia Cyanobacteria	*Akkermansia muciniphila* ATCC BAA-835	ND	**1**	**GH31** (**CBM32, dockerin**)
	***Anabaena* sp. strain PCC 7120**	**(1)**	ND	UNK protein
	***Gloeobacter violaceus* PCC 7421**	ND	**7**	UNK proteins
	***Synechococcus* sp. JA-2-3B′a (2–13)**	ND	**1**	UNK protein
	***Synechococcus* sp. strain JA-3-3Ab**	ND	**1**	UNK protein
Acidobacteria	*Solibacter usitatus* Ellin6076	**1**	ND	Types II and III secretion system protein
	Acidobacteria bacterium strain Ellin345	**1**	**1**	Types II and III secretion system protein (cohesin), integrin-like domains (dockerin)
BETA DELTA	***Ralstonia solanacearum* GMI1000**	**2**	ND	Putative GSPD-related protein (*A. cellulolyticus* ScaA-like **cohesin**)
	Bdellovibrio bacteriovorus HD100	**1**	ND	Cell wall surface anchor family protein precursor
	Desulfatibacillum alkenivorans AK-01	ND	**13**	**Some genes encoding dockerins are clustered**. Conserved UNK protein
	Geobacter bemidjiensis Bem	**1**	**2**	Conserved UNK proteins

BACTERIA	***c.* 500* individual species complete and incomplete genomes: 74 genomes encoding cohesins and/or dockerins (*c.* 14%)** **• Genomes encoding 'classic' cellulosome proteins are not included (e.g. *C. thermocellum, C. cellulolyticum, C. cellulovorans, R. flavefaciens, R. albus*, etc.)**			
General taxonomy	**Species name***	**Cohesin-containing ORFs†**	**Dockerin-containing ORFs**	**Comments and notes**
	Geobacter lovleyi SZ		**2**	ND Types II and III secretion system proteins, UNK protein
	***Geobacter metallireducens* GS-15**	**3**	**3**	Types II and III secretion system protein, NHL repeat domain protein
	***Geobacter uraniireducens* Rf4**	**2**	**9**	NHL repeat containing proteins, UNK proteins (**dockerins**)
	***Geobacter sulfurreducens* PCA**	**2**	**6**	Putative multicopper oxidase (**dockerin**), type II secretion system protein (**dockerin**), Fibronectin type III domain protein (**cohesin**)
	Geobacter sp. strain FRC-32	**1**	**2**	Putative multicopper oxidase, type II secretion system protein
	***Myxococcus xanthus DK* 1622**	**1**	ND	Cysteine-rich repeat protein, 1× PKD
	***Pelobacter carbinolicus* DSM 2380**	**1**	ND	Putative type II secretion system protein
	***Pelobacter propionicus* DSM 2379**	**1**	ND	Types II and III secretion system protein
	Stigmatella aurantiaca DW4/3-1	**1**	ND	F5/8 type C domain (**CBM32**) protein
GAMMA	***Syntrophobacter fumaroxidans* MPOB**	**2**	ND	Peptidase S8 or S53, hypothetical protein
	Alteromonas agarilytica GJ1B	ND	**1**	α-Agarase (**3× CBM6, dockerin**)
	Alteromonas macleodii 'Deep ecotype'	ND	**1**	Extracellular ribonuclease/nuclease fusion protein
	***Colwellia psychrerythraea* 34H**	ND	**1**	Extracellular ribonuclease/nuclease fusion protein
	***Hahella chejuensis* KCTC 2396**	**(1)**	**5**	Extracellular nuclease (**dockerin**), phosphohydrolase (**dockerin**)
	Legionella pneumophila (3 genomes)	**1**	ND	UNK protein (intracellularly induced gene)
	***Marinobacter aquaeolei* VT8**	ND	**1**	Predicted extracellular nuclease
	***Pseudoalteromonas atlantica* T6c**	**3**	**5**	Agarase (GH5**, dockerin**), endonuclease I (**dockerin**), peptidase (**dockerin**)
	***Shewanella baltica* OS195**	ND	**1**	YD repeat protein
	***Shewanella woodyi* ATCC 51908**	ND	**1**	Peptidase S8 and S53 (**cohesin**), large YD repeat protein (**dockerin**)
	***Shewanella* sp. strain MR-7**	ND	**3**	YD repeat protein (**dockerin**)
candidate division WWE1	**Candidatus Cloacamonas**	ND	**1**	Putative zinc-carboxypeptidase D (metallocarboxypeptidase D), putative CBM

EUKARYOTA

General taxonomy	Species name*	Cohesin-containing ORFs†	Dockerin-containing ORFs	Comments and notes
Fungi/Metazoa group, Choanoflagellida; Codonosigidae	**Choanoflagellate *Monosiga brevicollis* MX1/ATCC 50154**	**14**	**14**	Huge putative **cell adhesion proteins** (2775–10 110 aa) carrying tandemly positioned cohesin and dockerin, sometimes such 'tandem' repeated. Most of these proteins also carry other domains/motifs repeated many times: Cadherin (PF00028), TNFR/NGFR cysteine-rich region (PF00020), Ankyrin repeat (PF00023), etc.
Fungi/Metazoa group, Placozoa; Trichoplax	***Trichoplax adhaerens*,** a placozoan; the simplest known animal	ND	**2**	UNK protein; two dockerin-like sequences in the same protein
Alveolata; Ciliophora	***Tetrahymena thermophila* SB210**		**1**	UNK protein

Predictions of dockerin modules are based mainly on Pfam: PF00404 and InterPro: IPR002105 consensuses; Predictions of cohesin modules are based mainly on Pfam: PF00963 and InterPro: IPR002102 motifs. The detailed analysis, upon which this summary table is based, can be found in Table S1, complete with the full list of specific genes and their cohesin and/or dockerin sequences. The date of the final search of the database, used for construction of the table, was 23 September 2008.

*Complete genomes are shown in bold.

†A number of 'weakly' predicted cohesin motifs are shown in parentheses.

ND, not detected. UNK protein, uncharacterized, putative, or hypothetical protein of unknown function.

Abbreviations AKHD, NHL, PKD, etc. are those defined in the Pfam database (http://pfam.sanger.ac.uk/).