Received: 27 January 2019 | Revised: 26 May 2019 | Accepted: 30 May 2019
DOI: 10.1002/prot.25753

RESEARCH ARTICLE

PROTEINS WILEY

Distinctive ligand-binding specificities of tandem PA14 biomass-sensory elements from *Clostridium thermocellum* and *Clostridium clariflavum*

Inna Rozman Grinberg[1,2] | Oren Yaniv[1] | Lizett Ortiz de Ora[1,3] | Iván Muñoz-Gutiérrez[4,5] | Almog Hershko[1] | Oded Livnah[6] | Edward A. Bayer[4] | Ilya Borovok[1] | Felix Frolow[1†] | Raphael Lamed[1] | Milana Voronov-Goldman[1]

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

[2]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

[3]Department of Chemistry, University of California, Irvine, California

[4]Department of Biomolecular Sciences, The Weizmann Institute of Science, Rehovot, Israel

[5]Outreach Research Training and Minority Science Programs, School of Biological Sciences, University of California, Irvine, California

[6]Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Wolfson Centre for Applied Structural Biology, The Hebrew University of Jerusalem, The Edmond J. Safra Campus, Jerusalem, Israel

Correspondence
Milana Voronov-Goldman, Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv, Israel.
Email: mvoron@gmail.com

Funding information
Israel Science Foundation, Grant/Award Numbers: 1349/13, 24/11, 2566/16, 293/08; National Natural Science Foundation of China, Grant/Award Number: 31661143023; Consejo Nacional de Ciencia y Tecnología - México, Grant/Award Number: 440354

Abstract

Cellulolytic clostridia use a highly efficient cellulosome system to degrade polysaccharides. To regulate genes encoding enzymes of the multi-enzyme cellulosome complex, certain clostridia contain alternative sigma I (σ^I) factors that have cognate membrane-associated anti-σ^I factors (RsgIs) which act as polysaccharide sensors. In this work, we analyzed the structure-function relationship of the extracellular sensory elements of *Clostridium* (*Ruminiclostridium) thermocellum* and *Clostridium clariflavum* (RsgI3 and RsgI4, respectively). These elements were selected for comparison, as each comprised two tandem PA14-superfamily motifs. The X-ray structures of the PA14 modular dyads from the two bacterial species were determined, both of which showed a high degree of structural and sequence similarity, although their binding preferences differed. Bioinformatic approaches indicated that the DNA sequence of promoter of *sigI/rsgI* operons represents a strong signature, which helps to differentiate binding specificity of the structurally similar modules. The σ^{I4}-dependent *C. clariflavum* promoter sequence correlates with binding of RsgI4_PA14 to xylan and was identified in genes encoding xylanases, whereas the σ^{I3}-dependent *C. thermocellum* promoter sequence correlates with RsgI3_PA14 binding to pectin and regulates pectin degradation-related genes. Structural similarity between clostridial PA14 dyads to PA14-containing proteins in yeast helped identify another crucial signature element: the calcium-binding loop 2 (CBL2), which governs binding specificity. Variations in the five amino acids that constitute this loop distinguish the pectin vs xylan specificities. We propose that the first module ($PA14^A$) is dominant in directing the binding to the ligand in both bacteria. The two X-ray structures of the different PA14 dyads represent the first reported structures of tandem PA14 modules.

KEYWORDS
anti-sigma factors, biomass sensing, cellulosome, crystallography, RsgI, sigI, sigma factors

Abbreviations: CBL, calcium-binding loop; CBM, carbohydrate-binding module; GH, glycoside hydrolase; RsgI, anti-σ^I factor; sigI, sigma I (σ^I) factor.

†Deceased, 29 August 2014

1 | INTRODUCTION

Cellulolytic clostridia, notably *Clostridium* (*Ruminiclostridium*) *thermocellum*, have been extensively studied for their remarkable ability to ferment plant-derived polysaccharides.[1,2] The biomass degradation is performed by a highly efficient multi-enzyme system called cellulosome.[3] Genomes of cellulosome-producing bacteria encode a large variety of saccharolytic enzymes, including cellulases, hemicellulases, and pectin-degrading enzymes.[3,4] During the saccharolytic process, the type(s) of enzymes that are incorporated into the cellulosome are adjusted to suit the type(s) of polysaccharide present in the biomass.[5-7]

In an effort to better understand the mechanisms that govern the adjustment of the enzymatic composition of the cellulosome, our research group discovered in *C. thermocellum* a collection of eight RNA polymerase alternative sigI (σ^I) factors (Figure 1).[8,9] Six of these σ^Is (σ^{I1} to σ^{I6}) have cognate membrane-associated anti-σ^I factors (RsgI1-RsgI6) that contain C-terminal sensory elements, such as carbohydrate-binding modules (eg, CBM3 and CBM2), sugar-binding elements (eg, PA14 motif) or a glycoside hydrolase (eg, GH10 and GH5 families). The RsgI-borne sensory elements are displayed on the cell surface, suggesting that RsgIs may act as polysaccharide sensors.[8] Furthermore, the expression of σ^{I1} to σ^{I6} was shown to be affected by the presence of different polysaccharides in the growth medium (eg, cellulose and xylan).[9] Recent publications have shown that these sensing systems are also present in other cellulosome-producing bacteria, namely, *Clostridium* (*Ruminiclostridium*) *clariflavum* (Figure 1), *Clostridium* (*Ruminiclostridium*) *straminisolvens*, *Clostridium* sp. Bc-iso-3, *Acetivibrio cellulolyticus*, and *Bacteroides* (*Pseudobacteroides*) *cellulosolvens*.[10,11]

Alternative σ factors are highly promoter specific.[12-14] In this regard, a recent study showed that the *C. thermocellum* σ^{I3} factor is involved in the regulation of genes encoding pectin-degrading enzymes by using highly stringent promoter sequences.[10] These observations correlate with the binding capacities of the sensory element of the anti-σ^{I3} factor, which comprises two tandem PA14 motifs that bind pectin.[8]

The PA14 motif is a conserved protein module, possessing a core β-barrel topology, shared by a wide variety of bacterial and eukaryotic proteins.[15-19] PA14 occurs either as a single module, a doublet, or even a triplet in different proteins. However, the biological significance of these duplications has not been clear.[20] PA14 modules appear as components of carbohydrate-active enzymes,[17,21,22] epithelial adhesins,[15,16,18,19] flocculins,[18] and bacterial toxins, such as the anthrax protective antigen (PA), which is also a component of the anthrax toxin complex of known 3D structure.[23] According to Rigden et al,[20] PA14 was predicted to be a putative CBM. Only the structures of single PA14 modules have been determined to date.

In the present work, we combined biochemical analysis, computational biology and structural methods to better understand the relationship between the comparative binding specificities of RsgI-based PA14 modular dyads and the respective regulons of their cognate σ^I factors in two different clostridial species. The term, dyad, as used in this article, differs from the crystallographic meaning as dimers. In this article, dyad refers to a modular element comprising two parts and is consistent with the term modular dyad as used previously.[24-27] By using bioinformatic approaches, we showed that the σ^{I4} of *C. clariflavum* appears to be involved in the regulation of genes encoding mainly xylanases, as opposed to σ^{I3} of *C. thermocellum* that was previously shown to be involved in the regulation of genes encoding mainly pectinases. These results correlate well with the binding capacities of the sensory elements that comprise PA14 dyads in their cognate RsgIs.

The crystal structures of the PA14 modular dyads in both *C. thermocellum* and *C. clariflavum* showed a high degree of similarity

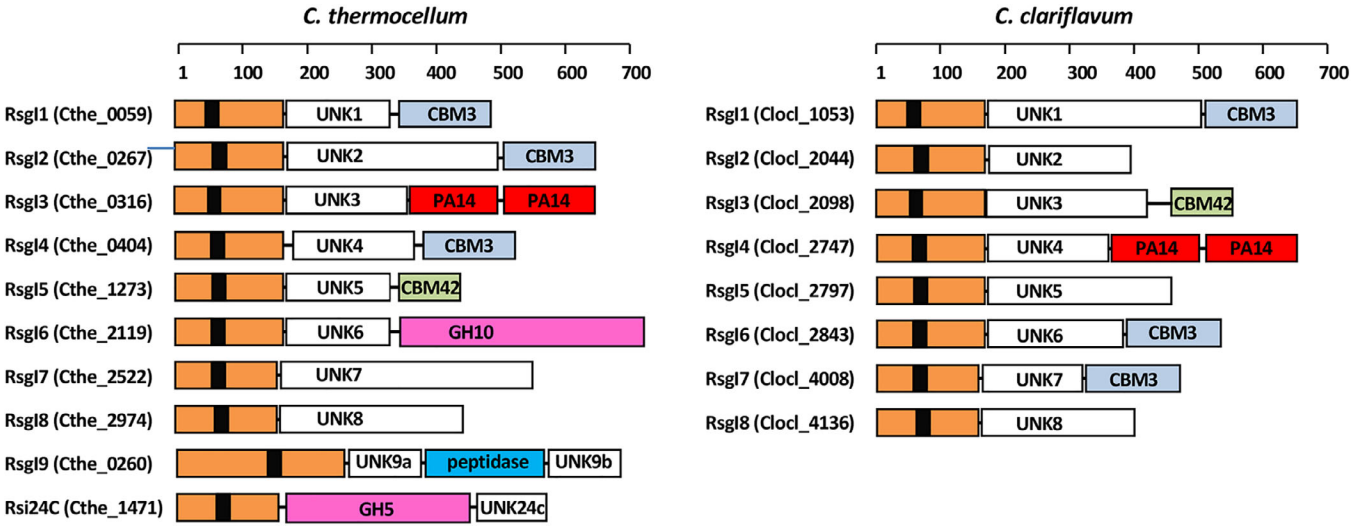

FIGURE 1 Schematic representation of the modular organization of RsgI proteins in *C. thermocellum and C. clariflavum.* Each multi-modular protein has a RsgI domain (orange) with an N-terminal transmembrane region (black). Additional modules are marked in different colors as follows: PA14, red; GH10, magenta; GH5, magenta; CBM3, light blue; CBM42, mint; UNK, divergent domains of unknown function/s having no sequence identity to other proteins, white; peptidase, blue. See main text for details. The linkers between the modules are indicated as lines. The ruler above indicates the length of the proteins (number of amino acids) [Color figure can be viewed at wileyonlinelibrary.com]

to the single PA14 modules from yeast that bind either galactosyl or mannosyl carbohydrates.[18,28] The rare D*cis*D calcium-binding motif was detected in the RsgI_PA14. The potential carbohydrate-binding mechanism of RsgI_PA14 was proposed, based on the comparison between the RsgI_PA14 modules and PA14 modules derived from yeast. Differences in primary structures of the calcium-binding loop 2 (CBL2) of the RsgI_PA14s appear to play an important role in the recognition of their target polysaccharide.

2 | MATERIALS AND METHODS

2.1 | Cloning

DNA fragments encoding the PA14 dyad ($PA14^{AB}$) of RsgI4 (Clocl_2747) (GenBank accession No. WP_014255852) were amplified by PCR from *C. clariflavum* DSM 19732 genomic DNA. DNA fragments encoding RsgI3 (Cthe_0316) PA14 dyad ($PA14^{AB}$; GenBank accession No. KM504391), were amplified by PCR from *C. thermocellum* ATCC 27405 genomic DNA. Genomic DNA was isolated as described by Murray and Thompson.[29] For the cloning of *C. clariflavum* PA14A: PA14A_C.cla_For (NdeI) 5′-GGAATTCCATATGGGATTAAGAGGCGATTA-3′ and PA14A_C.cla_Rev (XhoI) 5′-CCGCTCGAGTCACGCCGGATACAGGCATTCA-3′ were used; for $PA14^{B}$: PA14B_C.cla_For (NdeI) 5′-GGAATTCCATATGGGCTTGTTCTATGAGTA-3′ and PA14B_C.cla_Rev (XhoI) 5′-CCGCTCGAGTTATCTTGGATATAAACAGCT-3′; for the cloning of the PA14AB: PA14A_C.cla_For and PA14B_C.cla_Rev. For the cloning of *C. thermocellum* PA14A: PA14A_C.the_For (NcoI) 5′-CCATGGCGCCAACCGTCAGAAACGG-3′ and PA14A_C.the_Rev (NotI) 5′-GCGGCCGCGGAAGGATACAGTTGACTTG-3′ were used; for PA14B: PA14B_C.the_For (NcoI) 5′-CCATGGACGGCCCGCTGCCTCAG-3′ and PA14B_C.the_Rev (NotI) 5′-GCGGCCGCATCTGCAAACAAATTTTTTGAAG-3′; for the cloning of the PA14AB: PA14A_C.the_For and PA14B_C. the_Rev. To obtain constructs with N′ terminal hexahistidine (His) tag, primers PA14AB_C.the_For (NdeI) 5′-CATATGGGACTTAGGGGAGAGTATTAC-3′ and PA14AB_C.the_Rev (XhoI) 5′-CTCGAGATCTGCAAACAAATTTTTTGAAGG-3′ were used for the cloning of *C. thermocellum* PA14AB. The PCR products were purified, cloned into pGEM-T easy vector according to the manufacturer's instructions (Promega, Madison, WI), cleaved with the indicated restriction enzymes, and inserted into the pET-28a(+) expression vector (Novagen, Madison, WI). The obtained constructs, that is, pET-PA14AB_C.cla, containing the PA14 dyad, pET-PA14A_C.cla, and pET-PA14B_C.cla, containing both single PA14 modules, contained an N-terminal His tag and thrombin cleavage site. Constructs pET-PA14AB_C.the_C′, pET-PA14AB_C.the_N′, containing the PA14 dyad and pET-PA14A_C.the and pET-PA14B_C.the, contained either C-terminal or N-terminal His tags as indicated.

2.2 | Protein expression

Overnight cultures of *E. coli* BL21(DE3)/pET28a(+), bearing pET-PA14AB_*C.cla*, pET-PA14A_*C.cla*, pET-PA14B_*C.cla*, pET-PA14AB_*C.the_C′*, pET-PA14AB_*C.the_N′*, pET-PA14A_*C.the*, and pET-PA14B_*C.the* were diluted to an absorbance at 600 nm of 0.1 in LB broth (Lennox, Difco, BD Diagnostics, Sparks, MD), containing kanamycin (50 μg/mL), and shaken vigorously at 37°C. When cultures reached an absorbance at 600 nm, isopropyl-β-D-thiogalactopyranoside (IPTG; Sigma, St. Louis, MO) was added to a final concentration of 0.5 mM. The cells were grown overnight at ~20°C and harvested by centrifugation.

2.3 | Protein purification

The cell pellet was resuspended in sonication buffer (50 mM sodium phosphate buffer pH 7.6 containing 300 mM NaCl, glycerol 10%, and imidazole 10-20 mM), and the suspension was sonicated in an ultrasonic processor (Misonics, Seoul, Korea) until the suspension clarified. The sonicate was centrifuged at 5000*g* for 45 min at 4°C. The recombinant His-tagged protein was first isolated by metal-chelate affinity chromatography, whereby the supernatant fluids were loaded onto a HisTrap FF Ni Sepharose column (GE Healthcare Bio-Sciences AB, Uppsala, Sweden), equilibrated with the sonication buffer, washed thoroughly with buffer, and eluted with buffer containing 500 mM imidazole. Further purification was accomplished by fast protein liquid chromatography on a Superdex 75 16/60 column using an ÄKTA prime system (GE Healthcare, Chicago, IL), equilibrated with a solution of 50 mM Tris-HCl at pH 7.5 or 8.4, 150 mM NaCl, and sodium azide 0.05%. The eluted protein (the major peak eluted as monomer) was collected.

For crystallization, *C. clariflavum* $PA14^{AB}$ was subsequently cleaved by thrombin (Novagen) according to manufacturer's instructions, in buffer containing 20 mM Tris-HCl pH 8.4, 150 mM NaCl, 2.5 mM $CaCl_2$, incubated at 15°C overnight. The reaction mixture was dialyzed against buffer containing 50 mM Tris-HCl, pH 7.6, concentrated, and subjected to anion exchange chromatography using a Mono Q anion exchange chromatography column (GE Healthcare) in the same buffer. The protein failed to bind to the column and eluted at the void volume. The thrombin-cleaved protein contained three additional residues (GSH) that originated from the cleavage site of the enzyme at its N-terminus. Protein purity was evaluated by SDS-PAGE (15%). *C. clariflavum* $PA14^{AB}$ was concentrated to 15 mg/mL in buffer containing 50 mM Tris-HCl, pH 7.6, using Centriprep YM-10 centrifugal filter devices (Amicon Bioseparation, Millipore, Billerica, MA), and *C. thermocellum* $PA14^{AB}$ was concentrated to 18 mg/mL in buffer containing 50 mM Tris-HCl, pH 8.4, 150 mM NaCl, and the resultant proteins were subjected to crystallization screening.

While 2.5 mM $CaCl_2$ was included in the thrombin cleavage buffer for *C. clariflavum* $PA14^{AB}$, no calcium was added to *C. thermocellum* proteins during any of the purification steps.

For carbohydrate-binding assays, the uncleaved *C. clariflavum* $PA14^{AB}$ protein was used. *C. thermocellum* $PA14^{AB}$ proteins containing both N′ and C′ terminal His tags were used with similar results. Protein concentration was determined by measuring the UV absorbance at 280 nm, using the calculated extinction coefficient of the protein, as determined by the ProtParam tool (www.expasy.org/tools/protparam.html).

2.4 | SDS-PAGE-based carbohydrate-binding assay

Qualitative carbohydrate-binding assays were performed as described previously.[30-32] In brief, protein samples (50 μg in a total volume of 200 μL in buffer containing 50 mM Tris-HCl, pH 8.4, and 150 mM NaCl) were mixed with various polysaccharides (5 mg; Sigma Chem. Co., St Louis, MO): microcrystalline cellulose (Avicel), 1% amorphous cellulose.[33] The respective contents of the xylans from different sources were as follows (Sigma): Oat spelt xylan (10% arabinose, 15% glucose, and 70% xylose), birchwood xylan (90% xylose), and beechwood xylan (90% xylose). Pectin from apple (70-75% esterification), polygalacturonic acid, lichenan, and starch were also purchased from Sigma. $CaCl_2$ (10 mM) was included in assays containing pectin and polygalacturonic acid in order to precipitate the polysaccharides, since both are soluble in the absence of calcium. In specified assays, including xylan, 10 mM EDTA was included in the mixture. The mixtures were maintained at room temperature for 60 min with gentle rotation, then centrifuged at 12 000*g* for 5 min to sediment the polysaccharide and bound proteins. The supernatant (containing unbound proteins) was recovered and applied to SDS-PAGE gels. The polysaccharide precipitates were also washed four times with 1-mL aliquots of buffer. After centrifugation, the polysaccharides were resuspended into 200 μL of the same buffer and placed in a boiling water bath for 10 min. After centrifugation, the supernatant was recovered, and the proteins were subjected to SDS-PAGE. Each assay was repeated at least three times.

2.5 | Crystallization and data collection

Crystals were grown at 293 K by the hanging-drop vapor-diffusion method {McPherson, 1982 #89}. Crystals of *C. thermocellum* $PA14^A$ appeared in condition number 35 of Crystal screen 1 (0.1M HEPES pH 7.5, 0.8M sodium phosphate monobasic monohydrate, and 0.8M potassium phosphate monobasic). The crystals of $PA14^B$, appeared in condition number 10 of Crystal Screen 1 (0.2M ammonium acetate, 0.1M sodium acetate pH 4.6, and 30% PEG 4000). Finally, the crystals of $PA14^{AB}$ appeared in condition B7 from Wizard 3 (0.2M ammonium nitrate and 20% PEG 3350). Crystals of *C. clariflavum* $PA14^A$ appeared after 2 days in condition No. 2 of PEG ION 2 kit from Hampton Research (0.2M sodium malonate pH 4 and 20% w/v polyethylene glycol 3350). The crystals of $PA14^B$ and $PA14^{AB}$ appeared after a month in condition No. 7 of PEG ION 1(0.2M calcium chloride dihydrate and 20% PEG 3350) and in condition No. 22 of Crystal Screen 2 (0.1M MES pH 6.5, 12% PEG 20000), respectively.

The crystals were harvested from the crystallization drop using a MiTeGen micromount (http://www.mitegen.com) made of polyimide and transferred for several seconds into cryostabilization solution, composed of equal volumes of a twofold-concentrated solution of the crystallization components and a solution consisting of 18% (w/v) sucrose, 16% (w/v) glycerol, 16% (w/v) ethylene glycol, and 4% (w/v) glucose. For data collection, crystals were mounted on the MiTeGen micromount, plunged into liquid nitrogen, and placed in pucks for mounting at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). Diffraction data were collected on beamline BM14 for $PA14^A$ and $PA14^B$ from *C. thermocellum* and *C. clariflavum*, respectively, beamline ID23-1 for $PA14^B$, and beamline ID29 for $PA14^A$, $PA14^{AB}$ from *C. thermocellum* and *C. clariflavum*, respectively. X-ray data were collected on a Dectris PILATUS 6M pixel detector at 100 K.

2.6 | Structure determination and refinement

The crystal structure of the *C. thermocellum* $PA14^A$ was determined by molecular replacement using Molrep implemented in the CCP4i suite,[34,35] using the atomic coordinates from the anthrax protective antigen, PDB entry code 1ACC, as a search model. Although there is only 25.6% sequence identity between the two proteins, the structure of $PA14^A$ was successfully resolved. The model of $PA14^A$ was refined using REFMAC[36] as implemented in CCP4. The iterative model building and correlation to the electron density maps were conducted via COOT.[37] Water molecules were added by ARP/wARP.[38] The structure of the $PA14^B$ was determined by molecular replacement using $PA14^A$ as a search model. The structure of the $PA14^{AB}$ was determined by two-step molecular replacement (MR), using Molrep[34] from the CCP4 program suite.[35] In the first step, the coordinates of *C. thermocellum* $PA14^A$ were used as a search model for molecular replacement. Consequently, the $PA14^A$ coordinates were used as a fixed model, and the coordinates of $PA14^B$ were used as a search model. The initial coordinates that were obtained (including both $PA14^A$ and $PA14^B$) were subjected to ARP/wARP[38] for model building, followed by further refinement with CCP4, and manually rebuilt using Coot, until convergence (Table 1). The structure of the *C. clariflavum* $PA14^{AB}$ was determined and refined via a similar procedure as that used for *C. thermocellum* $PA14^A$. However, anisotropic temperature factors were refined due to the high resolution (Table 1). The refined structures and the corresponding amplitudes of the structure factors have been deposited in the PDB with accession codes 6QDI and 6QE7.

2.7 | Bioinformatics

Protein structure databases were searched for similar structures with the $PA14^A$ and $PA14^B$ modules as queries.[39] Phylogenetic analysis of σ^Is was conducted using the MEGA7 program[40] (http://www.megasoftware.net/) and multi-sequence analysis of σ^Is was performed using the MUSCLE[41] algorithm, implemented by MEGA7. The evolutionary history of 66 nucleotide sequences was inferred using the Neighbor-Joining method[42] with 2000 bootstrap replicates.[43] The evolutionary distances were computed using the Maximum Composite Likelihood method.[44] The phylogenetic tree was visualized and annotated with the online tool Interactive Tree Of Life (iTOL; available at: http://itol.embl.de/).[45]

DNA promoter motif searchers were carried out using the program Pattern Locator[46] (http://www.cmbl.uga.edu/software/patloc.html) and analyzed with the Jalview software.[47] Multiple DNA sequence alignments were performed using the T-Coffee algorithm[48] provided by Jalview. DNA sequence logos were generated with the program WebLogo[49] (http://weblogo.berkeley.edu/logo.cgi). The sequences used to build the phylogenetic tree and their accession numbers are given in Supporting Information.

TABLE 1 Crystallographic data collection and refinement statistics

	PA14ccl_AB	PA14ct_AB
ESRF beamline	ID29	ID29
Wavelength (Å)	0.98	0.972
PDB entry	6QDI	6QE7
Space group	C2	$P2_1$
Cell unit	*a* = 104.60	*a* = 65.34
	b = 41.43	*b* = 121.67
	c = 73.31	*c* = 65.94
	β = 100.35	β = 106.39
Resolution range (Last Shell) Å	45.97-1.13 (1.159-1.13)	63.26-2.06 (2.113-2.06)
Unique reflections	111 927 (7515)	60 326 (4500)
Redundancy	2.6 (2.5)	3.0 (3.0)
$R_{sym}(I)$	5.1 (42.7)	3.7 (36.4)
Completeness	96.67 (92.54)	95.73 (97.49)
I/σ	9.9 (1.9)	13.5 (2.5)
CC(1/2)	99.9 (81.0)	99.9 (89.7)
Number of protein atoms	2436	7117
Number of Ca atoms	4	6
Number of ethylene glycol (EDO) atoms	8	—
Number of solvent atoms	388	463
R factor, %	13.4 (23.2)	18.7 (30.1)
R_{free}, %	16.5 (26.7)	25.5 (35.2)
Avg. *B* factor, $Å^2$		
Protein	11.53	43.36
Ca^{+2}	7.30	34.86
EDO	11.92	—
Solvent	23.26	45.47
MolProbity Score (percentile)	1.69 (49th)	1.94 (84th)
rms deviation from ideality		
Bond length, Å	0.018	0.015
Bond angle, °	1.85	2.01
Ramachandran Plot		
Favored, %	95.9	93.7
Allowed, %	4.1	5.9
Outlier, %	0	0.5

Note: Quantities in the parentheses are defined for the last shell.

3 | RESULTS AND DISCUSSION

3.1 | Binding properties of $PA14^{AB}$ dyads from *C. thermocellum* and *C. clariflavum*

The alternative anti-sigma factors, *C. thermocellum* RsgI3 (Cthe_0316) and *C. clariflavum* RsgI4 (Clocl_2747), each possesses a C-terminal polysaccharide sensory element, composed of two tandem PA14-superfamily motifs, that were herein named $PA14^A$ (N-terminal) and $PA14^B$ (C-terminal). The PA14 modular dyads share high sequence similarity of 66% with an identity of 38.6%. However, the respective $PA14^A$-$PA14^B$ dyads ($PA14^{AB}$) from *C. thermocellum* RsgI3 and *C. clariflavum* RsgI4 appear to have different binding preferences. In a previous study, the *C. thermocellum* $PA14^{AB}$ dyad was examined for its binding to various polysaccharides and was found to bind strongly to pectin as well as to polygalacturonic acid[8] (Figure 2A). During the present work, we reproduced the previous approach with the *C. thermocellum* dyad to determine which of the subunits is/are responsible for the binding. Indeed, we verified that $PA14^{AB}$ binds pectin. Moreover, the single $PA14^A$ module (without $PA14^B$) also bound pectin, whereas $PA14^B$ exhibited negligible levels of binding to pectin and polygalacturonic acid and no binding to xylan (Figure 2A).

Similar to the RsgI-borne PA14 motifs in *C. thermocellum*, we also examined the *C. clariflavum* PA14 modules for binding to the same insoluble polysaccharides. The results (Figure 2B) show that *C. clariflavum* $PA14^{AB}$ binds xylan and to a much lesser extent pectin derivatives. The individual *C. clariflavum* PA14 modules, $PA14^A$, $PA14^B$ (or $PA14^{AB}$) were able to bind all three types of xylan tested in this work, that is, oat spelt xylan, birchwood xylan, and beechwood xylan (Figure 2B), with a preference for oat spelt xylan and less for birchwood and then beechwood xylans. Binding of xylan by $PA14^{AB}$ was abolished by addition of a metal chelator EDTA to the reaction mixture (Figure 2C; for details see section 3.6).

3.2 | The conservation of σ^I sequences represents a strong signature for differentiating the binding specificity of PA14 modules

To reveal the consequences of the different binding preferences of the two very similar PA14 dyads of the respective anti-σ^I factor (RsgI), we performed a phylogenetic analysis of the respective σ^I genes from available complex cellulosome-producing bacteria. It can be observed that *C. thermocellum* σ^{I3} clusters together with *B. cellulosolvens* σ^{I11}, *A. cellulolyticus* σ^{I6}, *Clostridium sp. Bc-iso-3*, and *C. straminisolvens* σ^{I3} factors (light blue-highlighted clade in Figure 3). *C. clariflavum* σ^{I4} and similar σ^I factors from other cellulolytic clostridia, namely, *A. cellulolyticus* σ^{I3} (WP_010247803.1) and *B. cellulosolvens* σ^{I1} (Bccel_0204), cluster together (light green-highlighted clade in Figure 3) at a branch relatively remote from that of the *C. thermocellum* σ^{I3}-like proteins. The C-terminal sensory elements of the *C. thermocellum* RsgI3-like proteins would presumably sense pectin as demonstrated by the binding assays performed with $PA14^{AB}$ of *C. thermocellum* (Figure 2A) and the PA14-CBM35 of *B. cellulosolvens* RsgI11 (data not shown). The C-terminal sensory elements of the *C. clariflavum* RsgI4-like proteins presumably sense xylan, as demonstrated by the binding assays performed with the $PA14^{AB}$ of *C. clariflavum* (Figure 2B) and the $PA14^{AB}$ dyad of *B. cellulosolvens* RsgI1 (data not shown). These results suggest that sequence conservation of the σ^I factors represents a signature that reflects the specificities of the PA14 sensory elements, located on their cognate RsgI.

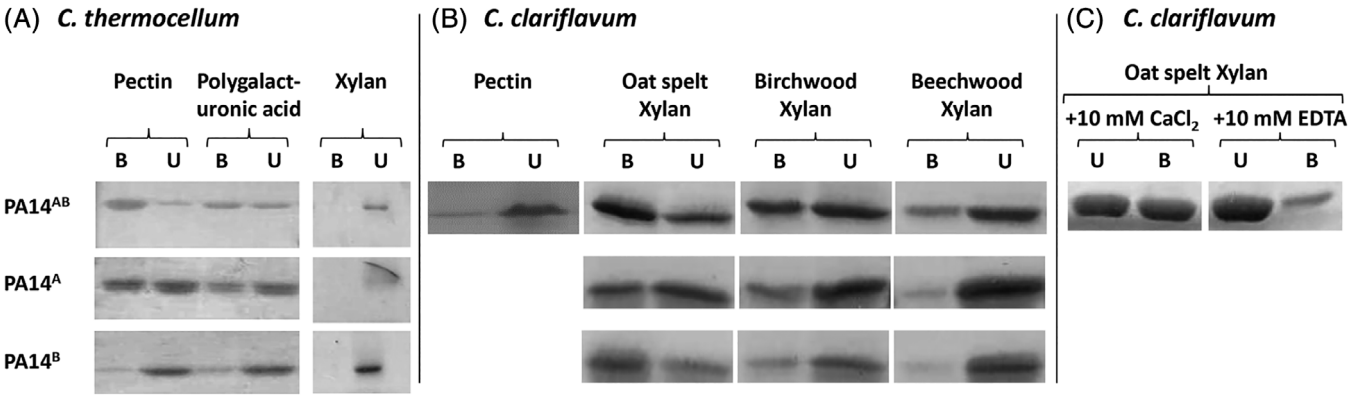

FIGURE 2 Binding of RsgI-borne PA14 modules from (A) *C. thermocellum* and (B) *C. clariflavum* to complex polysaccharides. Proteins were incubated with the insoluble polysaccharide, followed by centrifugation to separate pellet and supernatant, both of which were subjected to SDS-PAGE. The bound protein fraction (B) settled with the polysaccharide-containing pellet, whereas the unbound, noninteracting fraction (U) remained in the supernatant fluids. (C) Binding of *C. clariflavum* PA14^{AB} to oat spelt xylan in the presence of calcium or EDTA

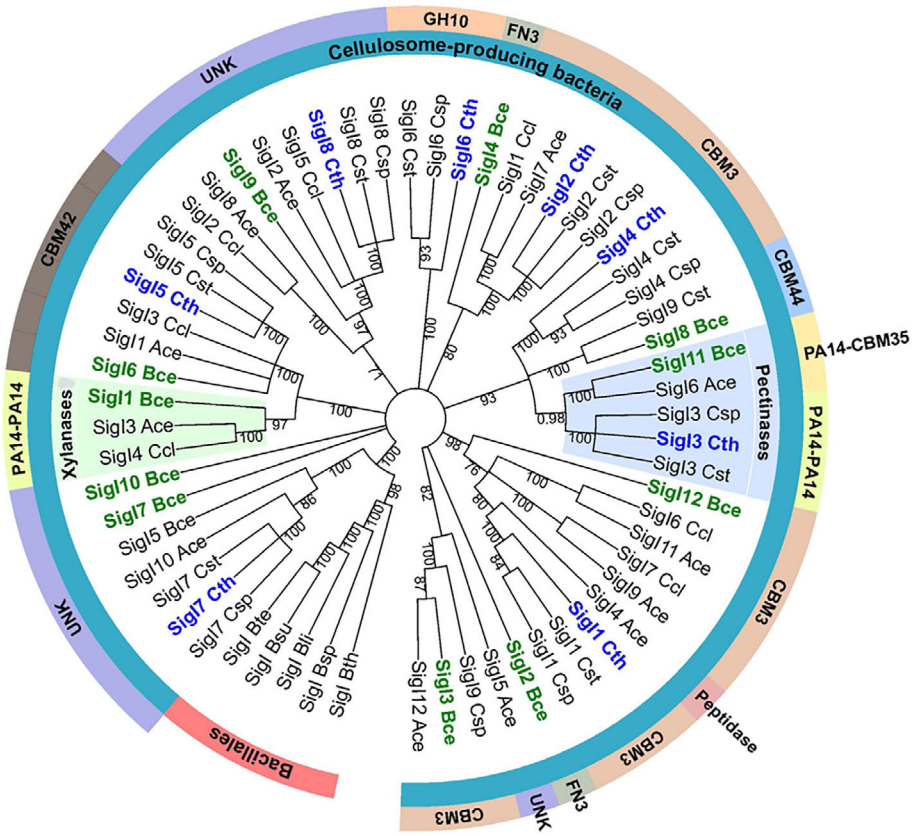

FIGURE 3 Evolutionary relationships of σ^{I} factors, derived from cellulosome-producing clostridia that harbor multiple σ^{I}s. Branches corresponding to partitions reproduced in <50% bootstrap replicates are collapsed. Corresponding putative sensory elements of cognate RsgIs are indicated by the outside labels. The clade of σ^{I} factors, containing the *B. cellulosolvens* σ^{I11} and *C. thermocellum* σ^{I3}, involved in regulating pectinase genes, is highlighted in light blue. The clade of σ^{I} factors, containing *B. cellulosolvens* σ^{I1}, presumably involved in regulating xylanase genes, is highlighted in light green. Ace, *Acetivibrio cellulolyticus*; Bce, *Bacteroides cellulosolvens*; Bli, *Bacillus licheniformis*; Bsp, *Bacillus* sp. NRRL B14911; Bsu, *Bacillus subtilis*; Bte, *Bacillus tequilensis*; Bth *Bacillus thuringiensis*; Cce, *Clostridium cellobioparum*; Ccl, *Clostridium clariflavum*; ChBD3, chitin-binding domain; Csp, *Clostridium* sp. Bc-iso-3; Cst, *Clostridium straminisolvens*; Cte, *Clostridium termitidis*; Cth, *Clostridium thermocellum*; FN3, fibronectin type 3 domain; GH, glycoside hydrolase; UNK, unknown [Color figure can be viewed at wileyonlinelibrary.com]

3.3 | The promoter sequence of the *sigI/rsgI* operon represents a strong signature for differentiating the binding specificity of the PA14 modules

Previous work indicated that *C. thermocellum* σ^{I3} and σ^{I11} from *B. cellulosolvens* use highly conserved promoter sequences to regulate genes encoding pectin-degrading enzymes.[50]

In order to identify the conserved promoter motif for genes encoding xylan-degrading enzymes, we performed bioinformatics analysis to identify the putative regulon of *C. clariflavum* σ^{I4}. We first compared the promoter regions of *C. clariflavum* sigI4 and *A. cellulolyticus* sigI4 to the previously predicted σ^{I}-dependent promoter of *B. cellulosolvens sigI1*.[50] As shown in Figure 4A, the predicted −35 region of the putative σ^{I} promoter of *C. clariflavum sigI4* has a conserved AAAA tetrad, and a conserved CGA triad three nucleotides upstream of the latter tetrad. A conserved CGAAT pentad can be observed in the predicted −10 region (Figure 4A). The spacer between the −35 and −10 region is 13 nucleotides (Figure 4A). Assuming that *C. clariflavum sigI4* is autoregulated, as has already been shown in

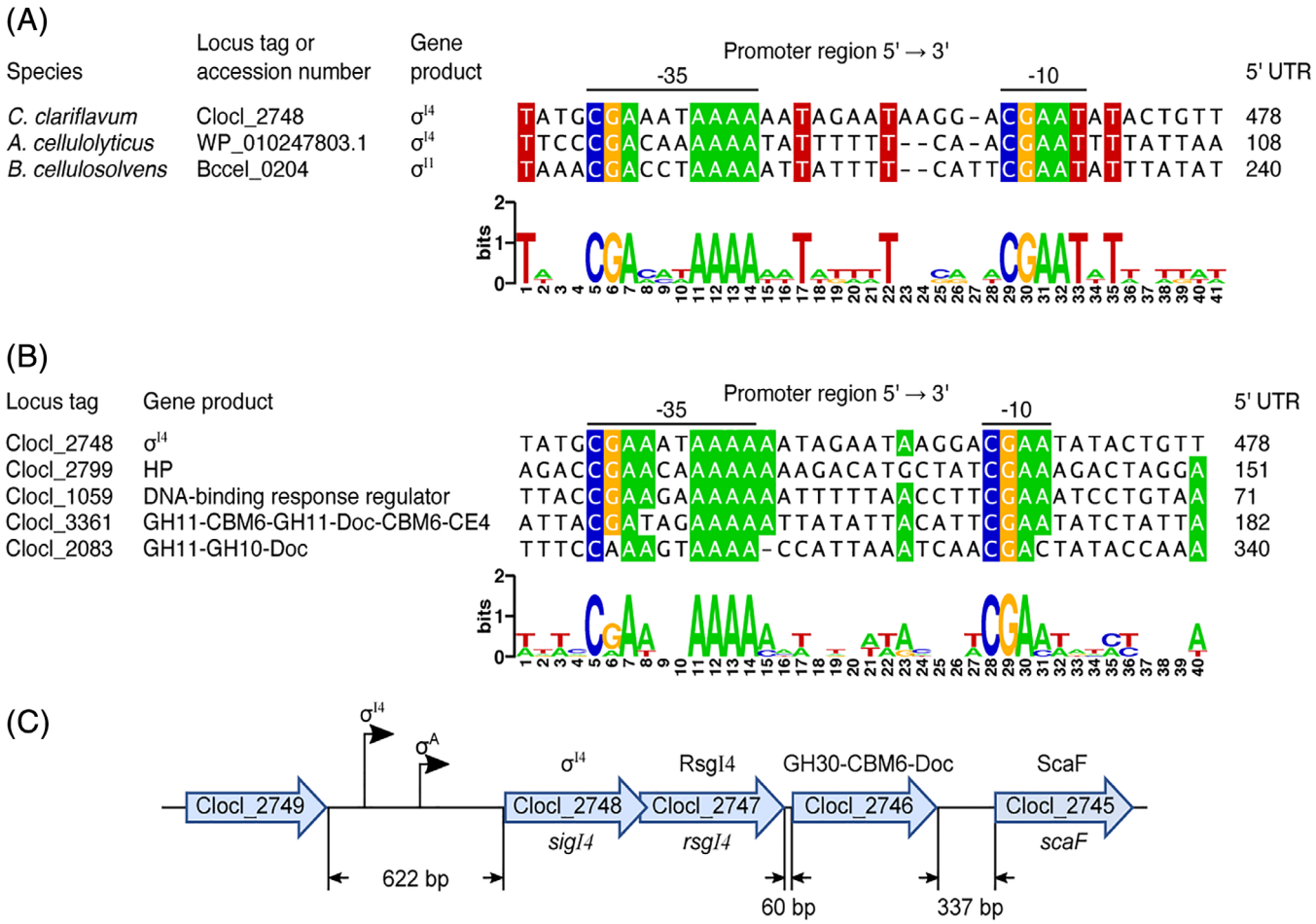

FIGURE 4 Identification of conserved elements of the *C. clariflavum* σ^{I4}-dependent promoter sequences. (A) Alignment of putative σ^{I}-dependent promoters of the *C. clariflavum* σ^{I4}, *A. cellulolyticus* σ^{I4}, and *B. cellulosolvens* σ^{I1} genes. Predicted −35 and −10 promoter elements are indicated. Distances between the promoter region and the first codon of the corresponding genes are shown in the column labeled 5′-UTR (5′-untranslated region). The WebLogo was generated with the sequence shown in the alignment. (B) Alignment of the putative *C. clariflavum* σ^{I4}-dependent promoter sequences. CBM6, family 6 carbohydrate-binding module; CE4, family 4 carbohydrate esterase; Doc, dockerin; GH10, family 10 glycoside hydrolase; GH11, family 11 glycoside hydrolase; HP, hypothetical protein. (C) Organization of the genomic context of the *C. clariflavum sigI4-rsgI4* operon. GH30, family 30 glycoside hydrolase; ScaF, scaffoldin F [Color figure can be viewed at wileyonlinelibrary.com]

other alternative σ^{I} factor genes,[9,50] we searched for putative σ^{I4}-dependent promoters in the genome of *C. clariflavum* by using the conserved motifs of the predicted σ^{I}-dependent promoter of *C. clariflavum sigI4* (Figure 4A). During the search, we allowed a distance of 12-14 nucleotides between the AAAA tetrad and the CGAAT pentad. Additionally, we allowed one mismatch in each of the −35 and −10 promoter sequences. As can be observed in Figure 4B, two of the four putative *C. clariflavum* σ^{I4}-dependent promoters identified in the present work correspond to genes encoding modules of a family 10 glycoside hydrolase (GH10) and/or family 11 glycoside hydrolase (GH11). According to the Carbohydrate-Active enZYmes (CAZy) Database, the GH11 family is composed only of xylanases and members of the GH10 family are mainly xylanases.

In *C. clariflavum*, the cellulosomal component genes are mainly scattered on the chromosome. Interestingly, 60 bp downstream of the *C. clariflavum sigI4-rsgI4* operon, there is a gene (Clocl_2746) encoding a dockerin-containing family 30 glycoside hydrolase (GH30; Figure 4C). We analyzed the intergenic region between *C. clariflavum rsgI4* and Clocl_2746 and the region upstream of *sigI4*, in order to search for putative σ^{I}- and σ^{A}-dependent promoters (σ^{A} is the housekeeping σ factor). During this analysis, we identified a putative σ^{A}-dependent promoter upstream of *sigI4* that has the sequence TTGAAA-17N-TATAAA, but we did not detect any σ^{I}- or σ^{A}-dependent promoter in the intergenic region between *sigI4* and Clocl_2746 (Figure 4C). These results indicate that the *C. clariflavum sigI4-rsgI4* pair of genes forms an operon with Clocl_2746, which is under the control of both σ^{I4} and σ^{A}. Interestingly, a recent report of *C. clariflavum* showed that Clocl_2746 has xylanolytic activity and that this enzyme was highly expressed, as measured by protein abundance in cellulosomes, upon growth of cells on different polysaccharide substrates including acid-pretreated switchgrass.[51]

The results of the above-described bioinformatic analysis indicate that *C. clariflavum* σ^{I4} most likely regulates genes encoding xylan-degrading enzymes, and hence, *C. clariflavum* RsgI4 likely senses xylan.

This correlates well with the binding capacities of the *C. clariflavum* RsgI4 PA14^{AB} dyad. We have thus concluded that the promoter sequence of the *sigI/rsgI* operon represents a strong signature which can be used to identify binding specificities of similar modules.

3.4 | Crystal structures of *C. thermocellum* and *C. clariflavum* PA14^{AB} modules

In order to study the mode of action of the *C. thermocellum* RsgI3 and *C. clariflavum* RsgI4 to the target polysaccharides, we determined the X-ray crystal structures of both dyads from each of the two clostridial species. It is important to mention that the term, dyad, used here does not follow the crystallographic definition of dyad as having a twofold axis that forms a dimer. The *C. thermocellum* PA14^{AB} dyad was crystallized and diffracted to a resolution of 2.06 Å (Figure 5A), and the *C. clariflavum* PA14^{AB} dyad crystal diffracted to 1.13 Å (Figure 5D). The statistics and data collection are given in Table 1. The structures were submitted to the Protein Data Bank (PDB: 6QDI and 6QE7).

The *C. thermocellum* and *C. clariflavum* RsgI-PA14 modules exhibit β-barrel structures, typical to those displayed by other PA14 modules,

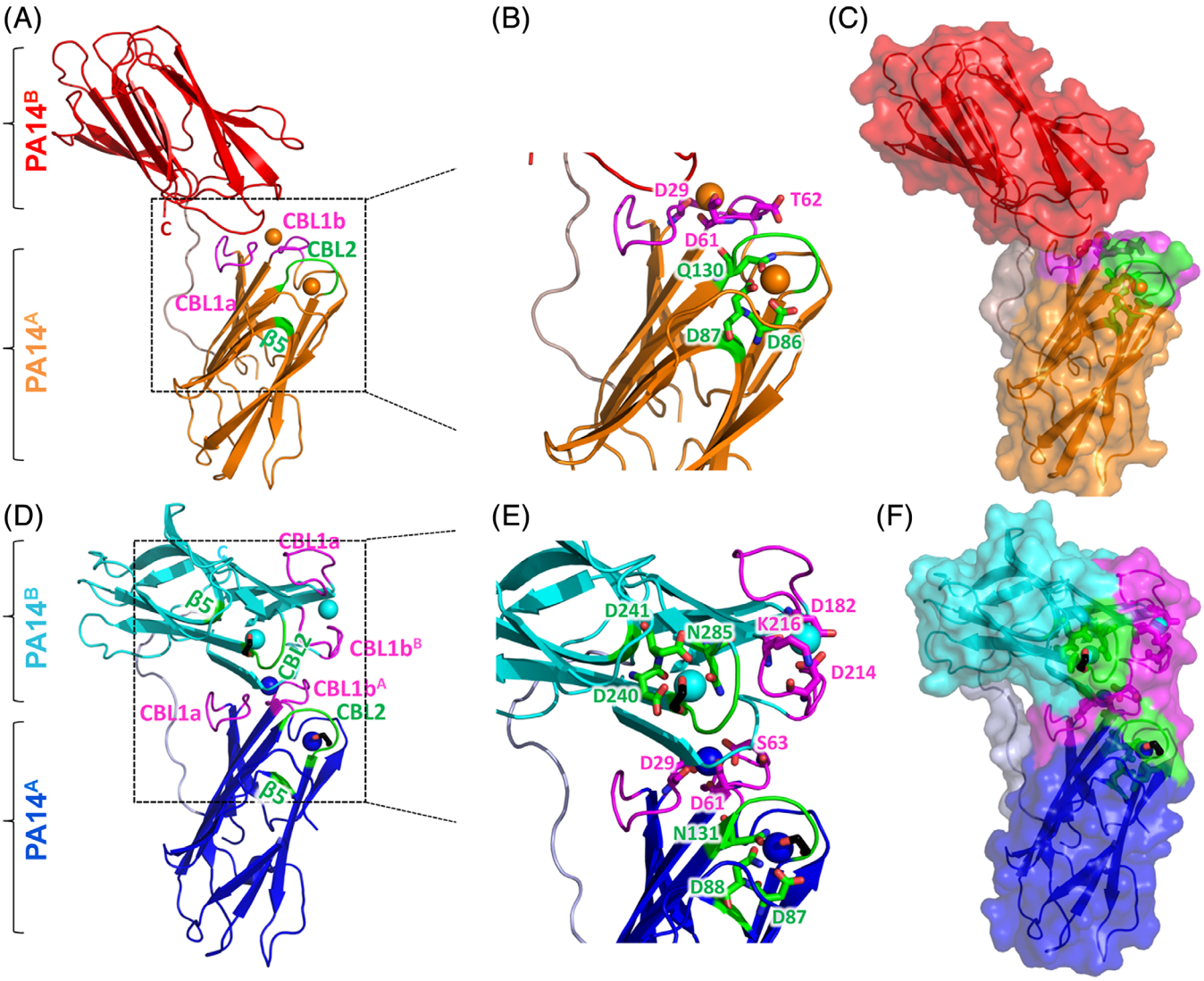

FIGURE 5 Crystal structures of the *C. thermocellum* (A-C) and *C. clariflavum* (D-F) PA14^{AB} modules. (A/D) Cartoon diagrams of the three-dimensional structures of the *C. thermocellum* RsgI PA14^{AB} and *C. clariflavun* RsgI PA14^{AB} dyads. Ca^{2+} atoms are represented as spheres. The first calcium atom of PA14 is coordinated by calcium-binding loop 1a (CBL1a) and CBL1b (magenta). The second Ca^{2+}atom is anchored by CBL2 and β-strand 5 (green). C and N termini are indicated. (A) The PA14^{A} module is in orange and PA14^{B} is in red. Only the PA14^{A} module possesses two calcium atoms, whereas PA14^{B} is void of calcium. (D) The PA14^{A} module is in blue and PA14^{B} is in cyan. Both PA14^{A} and PA14^{B} possess two Ca^{2+}atoms each. (B/E) Close-up view of the calcium-binding loops in *C. thermocellum* PA14^{A} and *C. clariflavum* PA14^{AB}. The calcium-binding residues of CBL1a and CBL1b are presented as magenta sticks, and the calcium-binding residues of CBL2 and β5 are presented as green sticks. (C/F) Transparent surface views of the *C. thermocellum* RsgI PA14^{AB} and *C. clariflavum* RsgI PA14^{AB} dyads. PA14^{A} of the two structures were first superimposed and then the structures were separated for clarity. CBL1a and CBL1b are colored magenta, and CBL2 and β5 are colored green. The Ethylene Glycol molecules are presented as black sticks [Color figure can be viewed at wileyonlinelibrary.com]

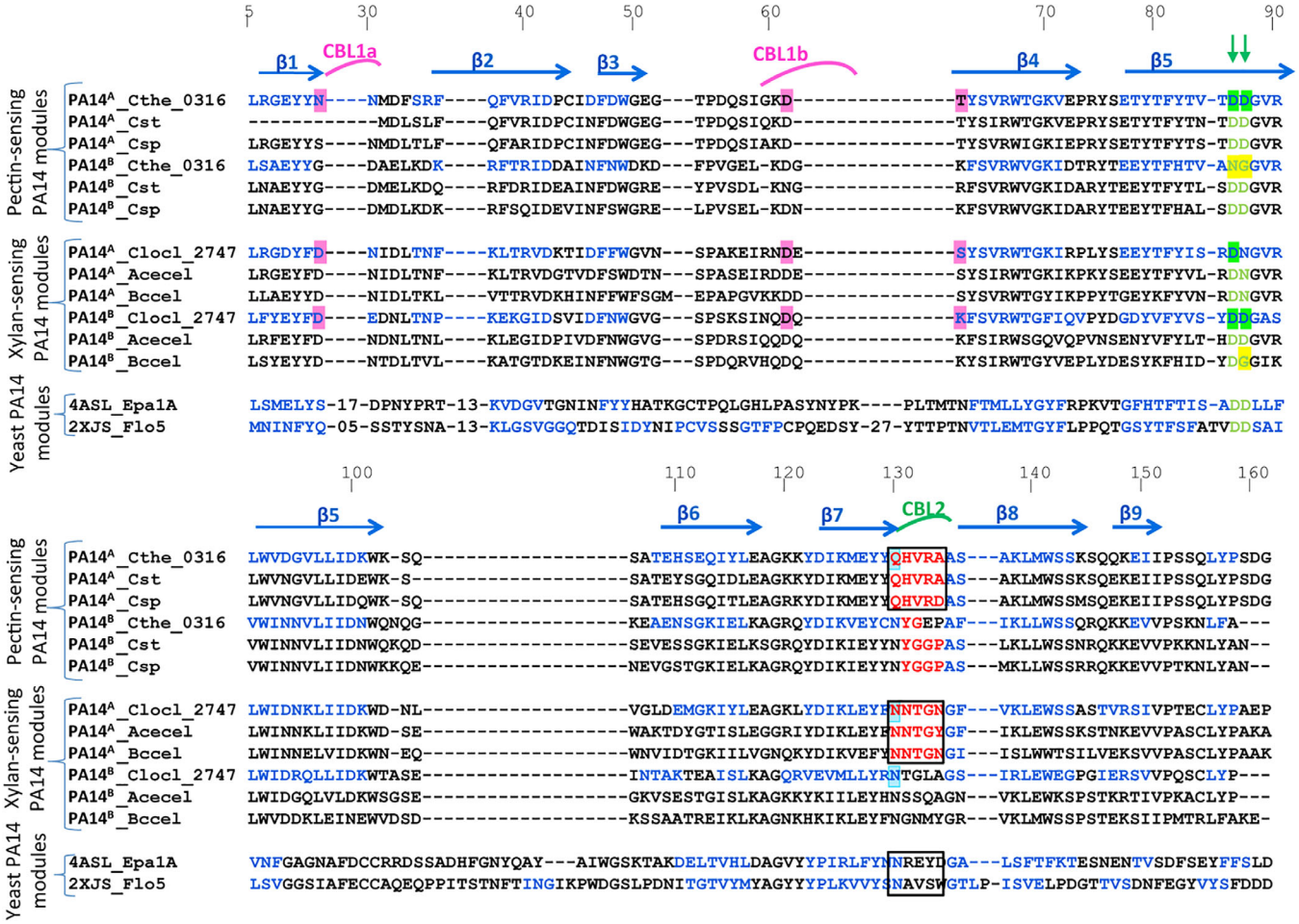

FIGURE 6 Sequence-based alignment of PA14 protein modules. β-Strand residues (arrows) at homologous positions are indicated in blue font. Amino acids responsible for Ca^{2+}binding in CBL1a and CBL1b are highlighted in magenta. Amino acids responsible for Ca^{2+}binding in CBL2 are highlighted in green. Amino acids of CBL2 are colored red, outlined by a black square. For clarity, the positions of selected residues in the sequence are enumerated. *C. clariflavum* $PA14^A$ and $PA14^B$ (Clocl_2747), *C. thermocellum* $PA14^A$ (Cthe_0316), $PA14^B$ (Cthe_0316) vs RsgI_PA14 modules of *C. straminisolvens* (Cst), *Clostridium* sp. Bc-iso-3 (Csp), *Acetivibrio cellulolyticus* (Acecel), *Bacteroides cellulosolvens* (Bccel), and yeast PA14 modules: Epa1A—*Candida glabrata* adhesin and Flo5—*Saccharomyces cerevisiae* flocculin [Color figure can be viewed at wileyonlinelibrary.com]

consisting of a 10-stranded anti-parallel β-sandwich fold. One of the β-sheets of the β-sandwich is formed by strands 1, 2, 4, 9, 6, and 7 while the other sheet comprises anti-parallel strands 3, 10, 5, and 8.

Clostridium thermocellum, $PA14^A$ contains two calcium ions. The first is coordinated by calcium-binding loop 1a (CBL1a) by D29 and calcium-binding loop 1b (CBL1b) by D61 and T62 (Figure 5A–C). CBL1a is positioned between the β strands 1 and 2, while CBL1b is positioned between strands 3 and 4 (Figure 6). The second calcium ion is coordinated by residue Q130 of CBL2 together with residues D86 and D87 of β-strand 5. CBL2 is located between β7 and β8 (Figure 6). Unlike $PA14^A$, *C. thermocellum* $PA14^B$ lacks calcium in its structure. In contrast, in the *C. clariflavum* $PA14^{AB}$ dyad, two calcium ions are present in both modules—one by CBL1a and CBL1b and the second by CBL2 and β5 (Figures 5D,E and 6). In *C. clariflavum* $PA14^A$, the first calcium atom is coordinated by D29 (CBL1a) and D61, S63 (CBL1b), and the second is coordinated by D87, N88 (CBL2), and N131 (β5). In *C. clariflavum* $PA14^B$, the first calcium atom is coordinated by D182 (CBL1a) and D214, K216 (CBL1b), and the second calcium is coordinated by D240, D241, and N285 (Figure 4E). The second calcium ion in each *C. clariflavum* module is bound to an ethylene glycol molecule, which presumably originated from the cryo-protectant used for mounting the crystals (Figure 5D-F, black sticks). Figure 5F clearly represents the ethylene glycol molecule bound to the potential carbohydrate-binding site of $PA14^{AB}$ from *C. clariflavum*.

Superposition of the $PA14^A$ and $PA14^B$ modules resulted in an rmsd of 0.63 Å over 95/95 residues and in 0.563 Å over 115/115 residues for *C. clariflavum* and *C. thermocellum* respectively, as calculated using Cα atoms in PyMOL. Overall, the structures of both modules are very similar, with some minor differences in the loops. The superposition of $PA14^A$ domains from both bacteria revealed that $PA14^B$ is located 150° relative to $PA14^A$ in *C. thermocellum* and about 90° relative to $PA14^A$ in *C. clariflavum* (Figure 5C,F).

The $PA14^A$ and $PA14^B$ modules are separated by a flexible linker (Figure 5A,D, gray). It is known that linkers may adopt alternative conformations, in accordance with their natural environmental conditions.[52] The linkers between the $PA14^A$ and $PA14^B$ modules (residues 160-174 in *C. clariflavum* and 159-170 in *C. thermocellum*) are long and presumably

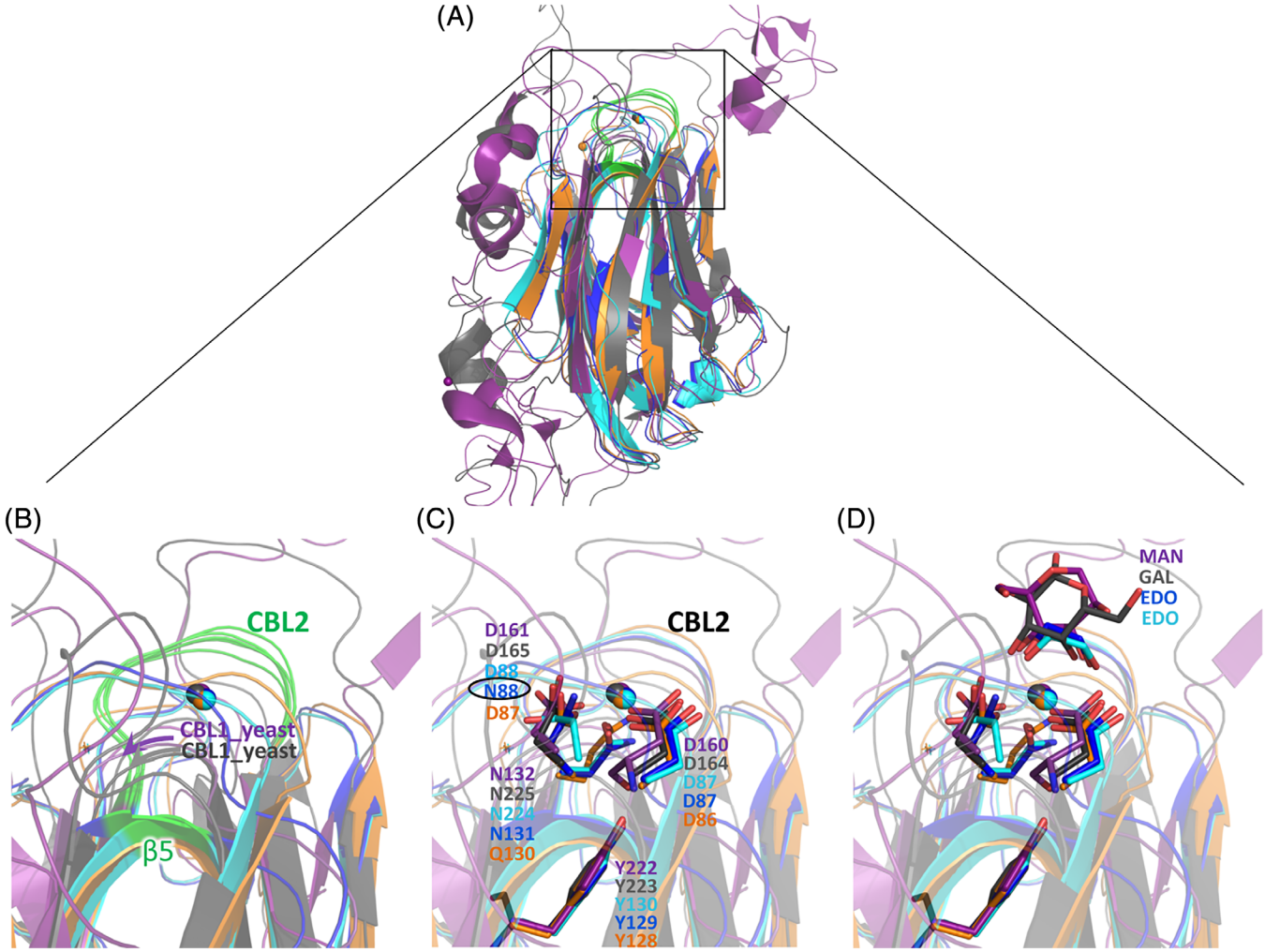

FIGURE 7 Structural comparison of *C. clariflavum* and *C. thermocellum* RsgI_PA14 with PA14 modules from yeast. *C. clariflavum* PA14^{A} (blue), *C. clariflavum* PA14^{B} (cyan), *C. thermocellum* PA14^{A} (orange), Ca. *glabrata* epithelial adhesin (4ASL, gray), and *S. cerevisiae* flocculin (2XJS, purple). (A) Comparison of *C. clariflavum* PA14^{A}, *C. clariflavum* PA14^{B} and *C. thermocellum* PA14^{A} with representative PA14 modules from yeast. Ca^{2+}atoms are presented as spheres. CBL2 and β-strand 5, which coordinate the Ca^{2+}atom, are colored green in all species. (B) Close-up view of CBL2 and β-strand 5 from RsgI_PA14 (green), and CBL1 from yeast (gray and purple), that interact with the calcium atom. (C) Structural superposition of the calcium-binding DD-N triad of members of the PA14 superfamily. N88 of the D*cis*N motif of *C. clariflavum* PA14^{A} is designated by a black circle. The four functionally important tyrosine residues in the various PA14 modules are shown. (D) The ethylene glycol molecules (cyan and blue) superimpose well with the hydroxyls of the sugar rings bound by yeast CBL2

flexible, and the two modules may possess different orientations towards each other. The two different crystal structures presented here may represent only part of the possible conformations, where the modules may be distant from one another (as in *C. thermocellum*) or in close proximity (as in *C. clariflavum*). Their position may depend, for instance, on the presence of an environmental signal, such as their carbohydrate substrate. The presence of the PA14 dyads and their interaction with the substrate may, in fact, play a regulatory role in signal transduction through the cell envelope.

3.5 | Clostridial RsgI_PA14s display similarity to PA14 modules from yeast

In order to better understand the mechanism of action of the RsgI_PA14 modules, we searched for proteins which share structural similarity. The Protein Data Bank was therefore searched with *C. thermocellum* and *C. clariflavum* PA14^{A}/ PA14^{B} (monomers) as queries, using the DALI server,[39] and several similar structures were found. High structural similarity of 2.2 and 2.3 Å was found with *Saccharomyces cerevisiae* flocculin (Flo) proteins (2XJS, 4AIL, and 4LHL-A) and *Candida glabrata* epithelial adhesins (Epa) (4ASL), respectively.[16,18,53-55] The RsgI_PA14 modules and the PA14 modular structures of the flocculin and adhesin proteins superimpose well at the core part of the β-sandwich region, although the adhesin (Figure 7A, gray) and flocculin (Figure 7A, purple) proteins are decorated by additional structural regions, which are absent from the RsgI PA14 modules. In both cases, the latter PA14 modules primarily exist in the single state in the parent proteins and not as modular dyads.

S. cerevisiae flocculins and Ca. *glabrata* adhesins[16,18,53,55] rely on the D*cis*D-N (Asp*cis*Asp-Asn) motif for complexation of the Ca^{2+} ions,

which contribute to their carbohydrate-binding properties as part of sugar-binding elements.[56] The D*cis*D-N calcium-binding signatures are found in over 85% of the 200 Epa adhesin-like domains that are present in the known fungal genome sequences[57] and also in the clostridial RsgI_PA14 modules as described herein. In the case of the clostridial PA14s, the conserved D*cis*D motif is located on β-strand 5, which is equivalent in location to CBL1 from yeast (Figure 7B), while the asparagine of the D*cis*D-N signature is located in CBL2, similar to that of the yeast. Whereas in the yeast PA14, both CBL1 and CBL2 are involved in binding of the same calcium atom, in the clostridial PA14s, CBL1a together with CBL1b bind the first calcium atom and CBL2, together with β5, bind the second. The CBL1a and CBL1b loops of RsgI_PA14 modules comprise a unique feature of biomass-sensing RsgI_PA14 modules and are absent from the structures of other PA14. The biological role of these loops is unclear.

The two aspartate residues (D*cis*D) positioned in *cis* orientation are rare and assume a structurally unfavorable conformation. In proteins for which the D*cis*D motif has been described,[18] the motif is highly conserved. (Figure 6, green arrows). However, in some of the RsgI_PA14 modules, like in *C. clariflavum* PA14^{A}, the second Asp is replaced by Asn (N88), which can be considered a comparatively minor replacement (Figure 6, Figure 7C, black circle). In RsgI_PA14, the CBL2 of both modules (res.132-137) links strands β7 and β8, and residues in β-strand 5 adopt a conserved Asp-*cis*-Asp (Figure 7C, D160, D161, cyan) in PA14^{B} and Asp-*cis*-Asn motif (Figure 7C, D87, N88, blue) in PA14^{A}. Both Asp and Asn interact with the calcium ion via their carbonyl group. The unfavorable *cis*-peptide configuration necessary for calcium binding is stabilized by a hydrogen bond to the hydroxyl group of the highly conserved *C. clariflavum* Y129 (in PA14^{A}) and Y283 (in PA14^{B}), similar to the equivalent tyrosines in the yeast, that is, Y222 in Flo5A (flocculin) and Y223 in Epa1A (adhesin; Figure 7C). Moreover, CBL2 contributes to calcium binding via the side chain of a highly conserved polar side residue: Asn (N131 in PA14^{A} and N285 in PA14^{B}) in *C. clariflavum* and Asn Q130 in PA14^{A} of *C. thermocellum* (Figure 6C). Carbonyl groups of the CBL2 main chain also contribute to calcium binding: T133 and N135 in *C. clariflavum* PA14^{A}, G287 and A289 in PA14^{B}, V132 and A134 in *C. thermocellum* PA14^{A}. Interestingly, in *C. thermocellum* RsgI_PA14, only the PA14^{A} module possesses the D*cis*D calcium-binding motif (Figure 6, indicated in green arrows). In the amino acid sequence of *C. thermocellum* PA14^{B}, however, the first aspartate is substituted by asparagine and the second by glycine (Figure 6, yellow-highlighted), thereby precluding the presence of the D*cis*D motif and preventing calcium binding, as demonstrated by its absence in the crystal structure (Figure 5A). These results are in good agreement with substrate-binding experiments, which show that *C. thermocellum* PA14^{B} is unable to bind pectin on its own, indicating that a calcium ion is essential for the ability to coordinate the polysaccharide.

3.6 | The calcium ion of CBL2 is involved in direct interaction with the ligand in RsgI_PA14 modules

Structural comparison of the clostridial PA14 modules with the yeast PA14 modules revealed that the position of the calcium ion in CBL2 is highly conserved between the Ca. *glabrata* epithelial adhesins, *S. cerevisiae* flocculins, and clostridial RsgI_PA14 modules (Figure 7B). This is in addition to the presence of the conserved D*cis*D motif (Figure 7C). Therefore, we assumed that the substrate-binding site of RsgI_PA14 is similar to that of the PA14 modules of adhesins and flocculins, and the carbohydrates may directly bind to the calcium ion via hydroxyls 3 and 4 of their xylose residues. In support of this hypothesis, the structures of both PA14^{A} and PA14^{B} modules in *C. clariflavum* have ethylene glycol molecules, bound to the calcium ion via hydroxyls (Figure 7D; EDO colored cyan, EDO colored blue), which superimpose perfectly onto the sugar ring hydroxyls, bound to the calcium ions in the structures of both Epa1a and Flo5 (Figure 7D MAN, GAC). Unfortunately, attempts to crystallize the RsgI_PA14 modules in the presence of xylose, xylobiose, and xylohexaose, failed to yield positive results. The ethylene glycol capture of the carbohydrate-binding site may have blocked the binding region under the standard cryoprotective conditions.

To verify the direct involvement of calcium in carbohydrate binding by RsgI_PA14, we tested the ability of *C. clariflavum* PA14^{AB} to bind xylan in the absence of calcium. EDTA was added to the carbohydrate-binding assay reaction mixture (Figure 2C). Chelation of calcium by EDTA completely abolished the binding of PA14^{AB} to xylan, demonstrating the involvement of calcium in carbohydrate binding. Parallel assays were not conducted for *C. thermocellum* PA14 binding to pectin, because calcium is necessary to keep the pectin insoluble for this assay. We propose, however, based on high structural similarity, that the role of calcium in *C. thermocellum* PA14^{AB} is the same as in *C. clariflavum* modules.

Based on structural comparisons in *C. clariflavum*, the presence of ethylene glycol at the carbohydrate-binding site and the results for binding capacities in the presence and absence of calcium, we propose that in the *C. clariflavum* RsgI_PA14 modules, the xylose ring binds directly to the calcium ion at CBL2.

3.7 | The protein sequence of CBL2 represents a signature for differentiation of xylan vs pectin binding by RsgI_PA14

Bioinformatics helped us to understand that the *sigI/rsgI* promoter sequence and the sequence of the sigI factor itself together dictate which kind of polysaccharide will be bound to the cognate RsgI PA14 module. In addition, review of the structural literature for PA14 modules revealed that prior to publishing the first PA14 modular structures, Zupancic et al[19] performed sequence swapping between various epithelial adhesins (Epas) and mapped a five amino acid region important for substrate binding specificity, within the PA14 module. In the 3D structure of Epa1a (PDB 4ASL), this region comprises the CBL2, which was shown to be directly involved in the interaction with the carbohydrate substrate (Figure 5 black square). Mutation of some key amino acids in this loop, specifically E227 and Y228 (first two amino acids of CBL2) affects the specificity of carbohydrate recognition by Epa1a.[28] These results would imply that the CBL2 sequence defines substrate recognition in *B. cellulosolvens* PA14^{A} (Figure 6 PA14A_Bccel) and *Acetivibrio cellulolyticus* PA14^{A} (Figure 6,

PA14^{A}_AceCel). Their CBL2s comprise residues NNTGN and NNTGY respectively, which are very similar to CBL2 from PA14^{A} of *C. clariflavum* (NNTGN; Figure 6, black square). Three of them were predicted to bind xylan according to the promoter sequences of their respective genes and were verified experimentally.[50] The supporting observation that the sequence of CBL2 is conserved within the group of pectin-binding bacteria was also detected in the sequences of CBL2 from *C. straminisolvens* and *C. sp.*, that comprise residues QHVRA and QHVRD, respectively, which are very similar to CBL2 from PA14^{A} of *C. thermocellum* (QHVRA) (Figure 6, black square). The sequences of CBL2 in both pectin-sensing and xylan-sensing RsgI PA14 bacterial groups are very different from each other. This variation, we believe, explains the divergence of carbohydrate-binding specificity between the two species. The sequence conservation of the CBL2 would thus constitute an additional control signature for substrate binding specificity.

From the sequence conservation in the PA14^{A} modules, it follows that PA14^{A} is more responsible for leading RsgI binding to its substrate as opposed to the PA14^{B} modules. Thus, the sequences of the CBL2 of the PA14^{B} modules within the members of the bacterial group that bind xylan are not conserved completely. In contrast, these sequences in members of the pectin-sensing bacterial group are conserved, although it was shown that CBL2 of *C. thermocellum* cannot bind substrate alone. The role of the conservation of CBL2 PA14^{B} sequence within the pectin-binding bacterial group is still unclear.

3.8 | Summary

Since the recombinant PA14 dyad from an RsgI of *C. thermocellum* (Cthe_0316) binds pectin strongly, it was thus initially expected that a similar PA14 dyad, discovered in an RsgI (Clocl_2747) of a related bacterium, *C. clariflavum*, would also bind pectin. Surprisingly, despite the strong sequence and structural similarity to the *C. thermocellum* protein, the *C. clariflavum* RsgI_PA14 was found to bind mainly xylan. Our findings therefore represent a fine-tuned structure-function relationship in sugar binding. The main function of oligosaccharide binding was preserved in the PA14 dyads from both bacteria; however, the type of target sugar differed.

A clearer understanding of this phenomenon became possible after comparing the relevant promoter regions and the protein sequences of the *sigma I* factors of the different species. The promoter and the *sigma I* factor of RsgI_PA14 from *C. thermocellum* possess conserved sequences associated with pectin sensing, whereas RsgI_PA14 from *C. clariflavum* possesses conserved sequences associated with xylan sensing. Structural similarity between the PA14 modules from yeast and those of the RsgI_PA14 served to pinpoint residues on CBL2 that appear responsible for substrate binding. Minor changes in the yeast sequence interfere with the binding capacity. Sequence divergence in this loop of PA14 pectin-binding dyads vs xylan-binding dyads thus implies that this loop is important for binding specificity.

It is important to consider comparison between PA14 and CBM modules. Most CBMs possess calcium atoms in their structures, however calcium generally plays a structural role, and its removal by chelation does not affect its carbohydrate-binding ability.[31,32] However, several CBM families, which, according to classification by Boraston et al,[58] belong to CBM type C, utilize calcium for recognition of their substrate polysaccharides, by either direct interaction of calcium with the hydroxyls of the sugar ring or with carboxylate moieties on a relevant carbohydrate.[59] These families are CBM35, CBM36, CBM60, CBM6, and CBM62.[37,60-63] Despite the fact that selected PA14s clearly bind carbohydrates, and in this particular case, calcium is directly involved in the interaction with the carbohydrate substrate, PA14 has yet to be classified as a bona fide CBM family.

4 | CONCLUSIONS

Redundancy is a common phenomenon in nature. Here, we detected three types of signature that allow differentiation of the function of the different *sigI/rsgI* pairs for targeting pectin vs xylan substrates, using very similar PA14 dyads as biosensors. The first relies on the DNA sequence of promoter region of the respective *sigI/rsgI* operon. This is a strong tool to predict the actual target of apparently similar protein biosensors. The second signature is based on the high similarity of the DNA sequence of related sigma factors, as reflected in their phylogenetic relationship. The third is based on the sequence similarity of specific regions within the CBL2 of the PA14A modules. Therefore, the same binding specificity data is "coded" at several biological levels and not only derived from protein structure. The present study provides a simplified method for predicting sugar binding specificity of newly discovered PA14 modules from *sigI/rsgI* operons.

ACKNOWLEDGEMENTS

The authors thank the staff scientists of ESRF for their remarkable assistance. This research was supported by the Israel Science Foundation (ISF; Grant nos. 293/08 to FF and 1349/13 to EAB). Additional support was obtained by a grant (No. 24/11) issued to RL by the Sidney E. Frank Foundation through the ISF. The authors also acknowledge a research grant from the Israel Science Foundation (ISF) (No. 2566/16) - National Natural Science Foundation of China (NSFC) (No. 31661143023). IM-G is grateful for the award of a Martin Kushner Schnur Post-Doctoral Fellowship at the Weizmann Institute. LOO was supported by the "Consejo Nacional de Ciencia y Tecnología - México" with a PhD scholarship (440354). EAB is the incumbent of The Maynard I. and Elaine Wishner Chair of Bio-organic Chemistry. We thank the staff of ESRF, Grenoble, for their outstanding maintenance and upgrading the facility. We thank Eleanor Gafni for her help with the characterization of *C. thermocellum* RsgI_PA14.

ORCID

Milana Voronov-Goldman https://orcid.org/0000-0002-3336-1967

REFERENCES

1. Akinosho H, Yee K, Close D, Ragauskas A. The emergence of *Clostridium thermocellum* as a high utility candidate for consolidated bioprocessing applications. *Front Chem*. 2014;2:66.

2. Tracy BP, Jones SW, Fast AG, Indurthi DC, Papoutsakis ET. Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. *Curr Opin Biotechnol*. 2012;23(3):364-381.
3. Artzi L, Bayer EA, Morais S. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat Rev Microbiol*. 2017;15(2):83-95.
4. Zverlov VV, Kellermann J, Schwarz WH. Functional subgenomics of *Clostridium thermocellum* cellulosomal genes: identification of the major catalytic components in the extracellular complex and detection of three new enzymes. *Proteomics*. 2005;5(14):3646-3653.
5. Gold ND, Martin VJ. Global view of the *Clostridium thermocellum* cellulosome revealed by quantitative proteomic analysis. *J Bacteriol*. 2007;189(19):6787-6795.
6. Raman B, Pan C, Hurst GB, et al. Impact of pretreated Switchgrass and biomass carbohydrates on *Clostridium thermocellum* ATCC 27405 cellulosome composition: a quantitative proteomic analysis. *PLoS One*. 2009;4(4):e5271.
7. Wei H, Fu Y, Magnusson L, et al. Comparison of transcriptional profiles of *Clostridium thermocellum* grown on cellobiose and pretreated yellow poplar using RNA-Seq. *Front Microbiol*. 2014;5:142.
8. Kahel-Raifer H, Jindou S, Bahari L, et al. The unique set of putative membrane-associated anti-sigma factors in *Clostridium thermocellum* suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. *FEMS Microbiol Lett*. 2010;308(1):84-93.
9. Nataf Y, Bahari L, Kahel-Raifer H, et al. *Clostridium thermocellum* cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. *Proc Natl Acad Sci USA*. 2010;107(43):18646-18651.
10. Munoz-Gutierrez I, Ortiz de Ora L, Rozman Grinberg I, et al. Decoding biomass-sensing regulons of *Clostridium thermocellum* alternative sigma-I factors in a heterologous *Bacillus subtilis* host system. *PLoS One*. 2016;11(1):e0146316.
11. Ortiz de Ora L, Munoz-Gutierrez I, Bayer EA, Shoham Y, Lamed R, Borovok I. Revisiting the regulation of the primary Scaffoldin gene in *Clostridium thermocellum*. *Appl Environ Microbiol*. 2017;83(8):e03088-16.
12. Campagne S, Allain FH, Vorholt JA. Extra cytoplasmic function sigma factors, recent structural insights into promoter recognition and regulation. *Curr Opin Struct Biol*. 2015;30:71-78.
13. Feklistov A, Darst SA. Promoter recognition by bacterial alternative sigma factors: the price of high selectivity? *Genes Dev*. 2009;23(20):2371-2375.
14. Rhodius VA, Segall-Shapiro TH, Sharon BD, et al. Design of orthogonal genetic switches based on a crosstalk map of sigmas, anti-sigmas, and promoters. *Mol Syst Biol*. 2013;9:702.
15. de Groot PW, Klis FM. The conserved PA14 domain of cell wall-associated fungal adhesins governs their glycan-binding specificity. *Mol Microbiol*. 2008;68(3):535-537.
16. Ielasi FS, Decanniere K, Willaert RG. The epithelial adhesin 1 (Epa1p) from the human-pathogenic yeast *Candida glabrata*: structural and functional study of the carbohydrate-binding domain. *Acta Crystallogr D Biol Crystallogr*. 2012;68(Pt 3):210-217.
17. Silipo A, Larsbrink J, Marchetti R, Lanzetta R, Brumer H, Molinaro A. NMR spectroscopic analysis reveals extensive binding interactions of complex xyloglucan oligosaccharides with the *Cellvibrio japonicus* glycoside hydrolase family 31 alpha-xylosidase. *Chemistry*. 2012;18(42):13395-13404.
18. Veelders M, Bruckner S, Ott D, Unverzagt C, Mosch HU, Essen LO. Structural basis of flocculin-mediated social behavior in yeast. *Proc Natl Acad Sci USA*. 2010;107(52):22511-22516.
19. Zupancic ML, Frieman M, Smith D, Alvarez RA, Cummings RD, Cormack BP. Glycan microarray analysis of *Candida glabrata* adhesin ligand specificity. *Mol Microbiol*. 2008;68(3):547-559.
20. Rigden DJ, Mello LV, Galperin MY. The PA14 domain, a conserved all-beta domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends Biochem Sci*. 2004;29(7):335-339.
21. Yoshida E, Hidaka M, Fushinobu S, et al. Role of a PA14 domain in determining substrate specificity of a glycoside hydrolase family 3 beta-glucosidase from *Kluyveromyces marxianus*. *Biochem J*. 2010;431(1):39-49.
22. Zmudka MW, Thoden JB, Holden HM. The structure of DesR from *Streptomyces venezuelae*, a beta-glucosidase involved in macrolide activation. *Protein Sci*. 2013;22(7):883-892.
23. Petosa C, Collier RJ, Klimpel KR, Leppla SH, Liddington RC. Crystal structure of the anthrax toxin protective antigen. *Nature*. 1997;385(6619):833-838.
24. Jindou S, Xu Q, Kenig R, et al. Novel architecture of family-9 glycoside hydrolases identified in cellulosomal enzymes of *Acetivibrio cellulolyticus* and *Clostridium thermocellum*. *FEMS Microbiol Lett*. 2006;254(2):308-316.
25. Levy-Assaraf M, Voronov-Goldman M, Rozman Grinberg I, et al. Crystal structure of an uncommon cellulosome-related protein module from *Ruminococcus flavefaciens* that resembles papain-like cysteine peptidases. *PLoS One*. 2013;8(2):e56138.
26. Noach I, Levy-Assaraf M, Lamed R, Shimon LJ, Frolow F, Bayer EA. Modular arrangement of a cellulosomal scaffoldin subunit revealed from the crystal structure of a cohesin dyad. *J Mol Biol*. 2010;399(2):294-305.
27. Rincon MT, Cepeljnik T, Martin JC, et al. A novel cell surface-anchored cellulose-binding protein encoded by the sca gene cluster of *Ruminococcus flavefaciens*. *J Bacteriol*. 2007;189(13):4774-4783.
28. Ielasi FS, Verhaeghe T, Desmet T, Willaert RG. Engineering the carbohydrate-binding site of Epa1p from *Candida glabrata*: generation of adhesin mutants with different carbohydrate specificity. *Glycobiology*. 2014;24(12):1312-1322.
29. Murray MG, Thompson WF. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res*. 1980;8(19):4321-4325.
30. Xu Q, Morrison M, Nelson KE, Bayer EA, Atamna N, Lamed R. A novel family of carbohydrate-binding modules identified with *Ruminococcus albus* proteins. *FEBS Lett*. 2004;566(1–3):11-16.
31. Yaniv O, Shimon LJ, Bayer EA, Lamed R, Frolow F. Scaffoldin-borne family 3b carbohydrate-binding module from the cellulosome of *Bacteroides cellulosolvens*: structural diversity and significance of calcium for carbohydrate binding. *Acta Crystallogr D Biol Crystallogr*. 2011;67(Pt 6):506-515.
32. Yaniv O, Jindou S, Frolow F, Lamed R, Bayer EA. A simple method for determining specificity of carbohydrate-binding modules for purified and crude insoluble polysaccharide substrates. *Methods Mol Biol*. 2012;908:101-107.
33. Lamed R, Setter E, Bayer EA. Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*. *J Bacteriol*. 1983;156(2):828-836.
34. Vagin A, Teplyakov A. MOLREP: an automated program for molecular replacement. *J Appl Cryst*. 1997;30:1022-1025.
35. Winn MD, Ballard CC, Cowtan KD, et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*. 2011;67(Pt 4):235-242.
36. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr*. 1997;53:240-255.
37. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*. 2010;66(Pt 4):486-501.
38. Perrakis A, Morris R, Lamzin VS. Automated protein model building combined with iterative structure refinement. *Nat Struct Biol*. 1999;6(5):458-463.
39. Holm L, Rosenstrom P. *Dali* server: conservation mapping in 3D. *Nucleic Acids Res*. 2010;38:W545-W549.

40. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016; 33(7):1870-1874.
41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797.
42. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4(4):406-425.
43. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. 1985;39(4):783-791.
44. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA*. 2004;101(30):11030-11035.
45. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 2016;44(W1):W242-W245.
46. Mrazek J, Xie S. Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences. *Bioinformatics*. 2006;22(24): 3099-3100.
47. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189-1191.
48. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 2000;302 (1):205-217.
49. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188-1190.
50. Ortiz de Ora L, Lamed R, Liu YJ, et al. Regulation of biomass degradation by alternative sigma factors in cellulolytic clostridia. *Sci Rep*. 2018;8(1):11036.
51. Artzi L, Morag E, Barak Y, Lamed R, Bayer EA. *Clostridium clariflavum*: key cellulosome players are revealed by proteomic analysis. *MBio*. 2015;6(3):e00411-15.
52. Noach I, Frolow F, Alber O, Lamed R, Shimon LJ, Bayer EA. Intermodular linker flexibility revealed from crystal structures of adjacent cellulosomal cohesins of *Acetivibrio cellulolyticus*. *J Mol Biol*. 2009;391 (1):86-97.
53. Goossens K, Willaert R. Flocculation protein structure and cell-cell adhesion mechanism in *Saccharomyces cerevisiae*. *Biotechnol Lett*. 2010;32(11):1571-1585.
54. Ielasi FS, Goyal P, Sleutel M, Wohlkonig A, Willaert RG. The mannose-specific lectin domains of Flo1p from *Saccharomyces cerevisiae* and Lg-Flo1p from *S. pastorianus*: crystallization and preliminary X-ray diffraction analysis of the adhesin-carbohydrate complexes. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 2013;69(Pt 7):779-782.
55. Maestre-Reyna M, Diderrich R, Veelders MS, et al. Structural basis for promiscuity and specificity during *Candida glabrata* invasion of host epithelia. *Proc Natl Acad Sci USA*. 2012;109(42):16864-16869.
56. Gabius HJ, Andre S, Jimenez-Barbero J, Romero A, Solis D. From lectin structure to functional glycomics: principles of the sugar code. *Trends Biochem Sci*. 2011;36(6):298-313.
57. Diderrich R, Kock M, Maestre-Reyna M, et al. Structural hot spots determine functional diversity of the *Candida glabrata* epithelial Adhesin family. *J Biol Chem*. 2015;290(32):19597-19613.
58. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J*. 2004; 382(Pt 3):769-781.
59. Gilbert HJ, Knox JP, Boraston AB. Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin Struct Biol*. 2013;23(5):669-677.
60. Henshaw J, Horne-Bitschy A, van Bueren AL, et al. Family 6 carbohydrate binding modules in beta-agarases display exquisite selectivity for the non-reducing termini of agarose chains. *J Biol Chem*. 2006;281 (25):17099-17107.
61. Jamal-Talabani S, Boraston AB, Turkenburg JP, Tarbouriech N, Ducros VM, Davies GJ. Ab initio structure determination and functional characterization of CBM36; a new family of calcium-dependent carbohydrate binding modules. *Structure*. 2004;12(7):1177-1187.
62. Montanier C, Flint JE, Bolam DN, et al. Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. *J Biol Chem*. 2010;285(41):31742-31754.
63. Montanier CY, Correia MA, Flint JE, et al. A novel, noncatalytic carbohydrate-binding module displays specificity for galactose-containing polysaccharides through calcium-mediated oligomerization. *J Biol Chem*. 2011;286(25):22499-22509.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Grinberg IR, Yaniv O, de Ora LO, et al. Distinctive ligand-binding specificities of tandem PA14 biomass-sensory elements from *Clostridium thermocellum* and *Clostridium clariflavum*. *Proteins*. 2019;87:917–930. https://doi.org/10.1002/prot.25753