Supporting Information

The structure of the *Clostridium thermocellum* RsgI9 ectodomain provides insight into the mechanism of biomass sensing

Brendan J. Mahoney[1-2], Allen Takayesu[1], Anqi Zhou[1], Duilio Cascio[2], Robert T. Clubb[1-3] *

[1]Department of Chemistry and Biochemistry, [2]UCLA-DOE Institute of Genomics and Proteomics, and the [3]Molecular Biology Institute, University of California, Los Angeles, 611 Charles E. Young Drive East, Los Angeles, CA 90095, USA.

*To whom correspondence should be addressed:

Prof. Robert T. Clubb, Department of Chemistry and Biochemistry, University of California, Los Angeles, 602 Boyer Hall, Los Angeles, CA 90095.

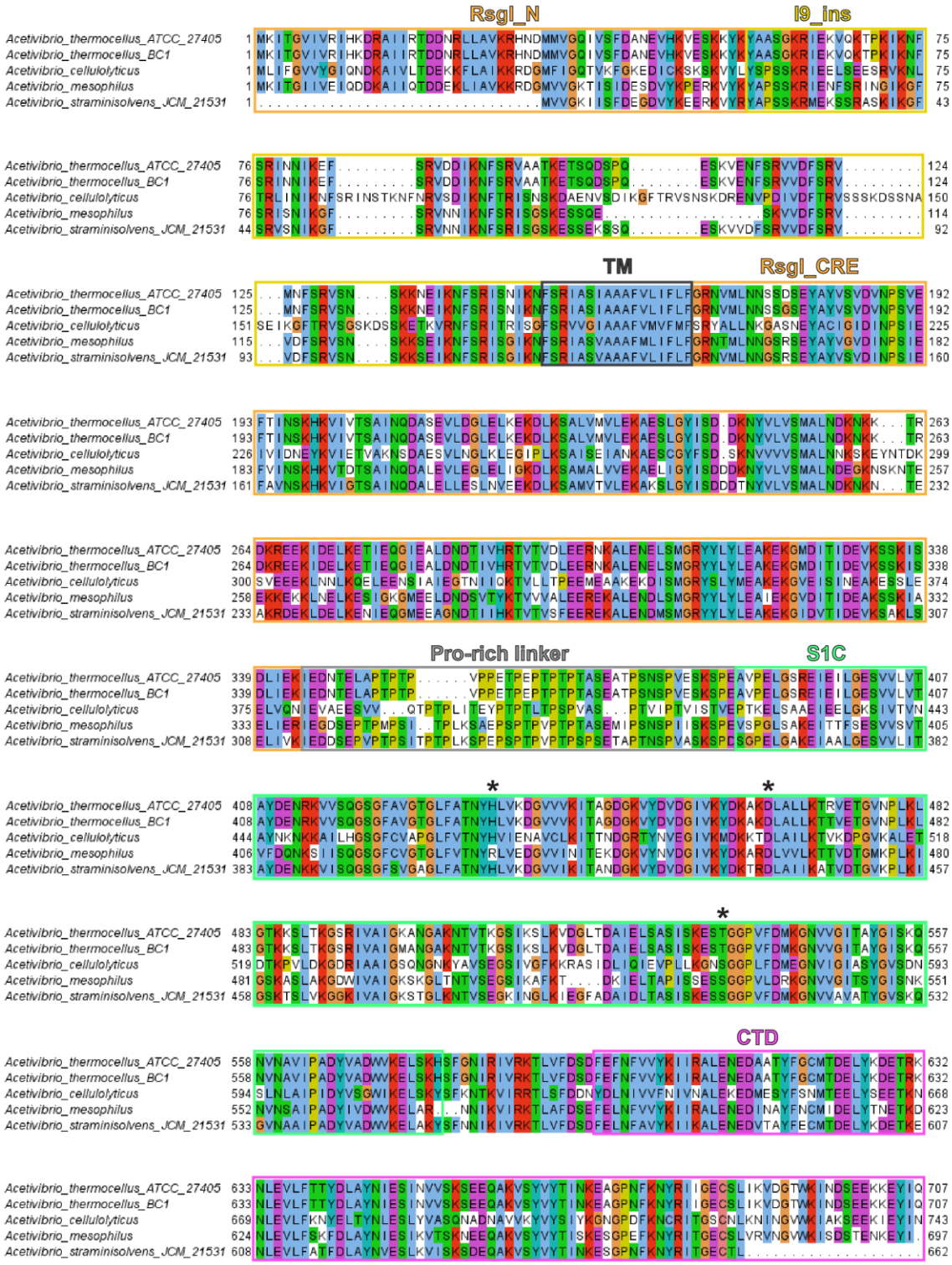

Figure S1. Protein sequence alignment of RsgI9 sequences in *Clostridia*. Alignment of the amino acid sequences of RsgI9 in *Clostidium thermocellum* (*Acetivibrio thermocellus*) and related species using Clustal Omega[1]. The domains are boxed and labeled in the same colors, and the catalytic triad residues in the S1C peptidase domain are indicated by the asterisks.

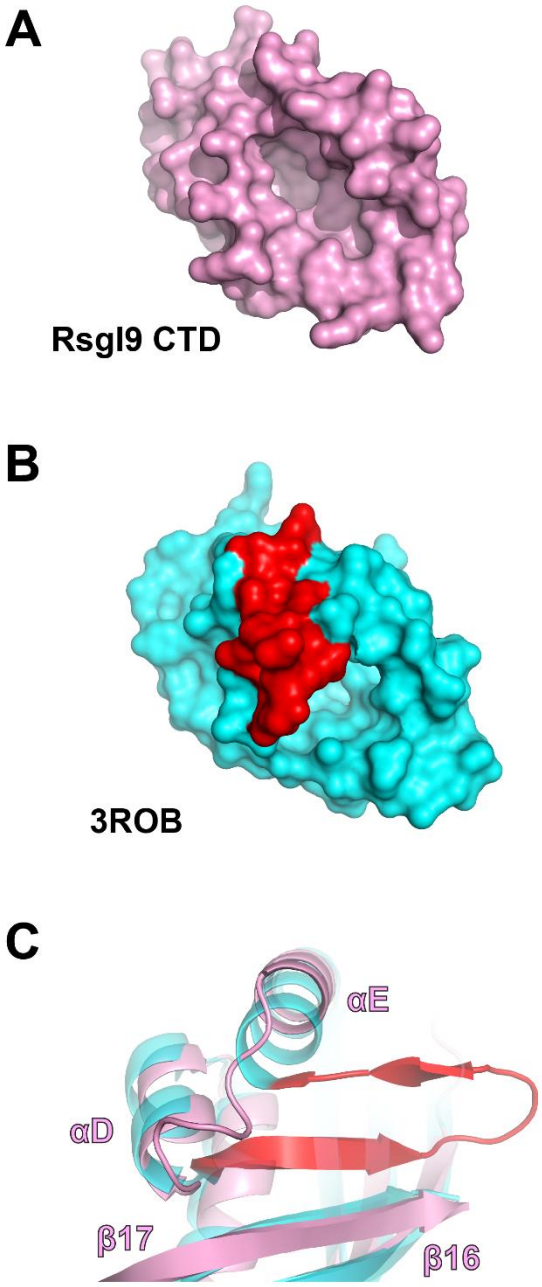

Figure S2. Comparison of RsgI9's CTD with a structurally similar NTF2-like protein. (A) Surface representation of RsgI9 C-terminal domain (CTD) (pink), and (B) a NTF2-like structural homolog shown in the same orientation (cyan) (PDB accession code 3ROB). The structures are similar and have a DALI Z-score = 13.7 and PDBeFold Z-score = 7.8. (C) Cartoon figure showing the extended β-sheet found in other NTF2-like proteins (red) that is missing in RsgI9 (pink). RsgI9 secondary structural elements are labeled. The positioning of the structures is the same in each panel.

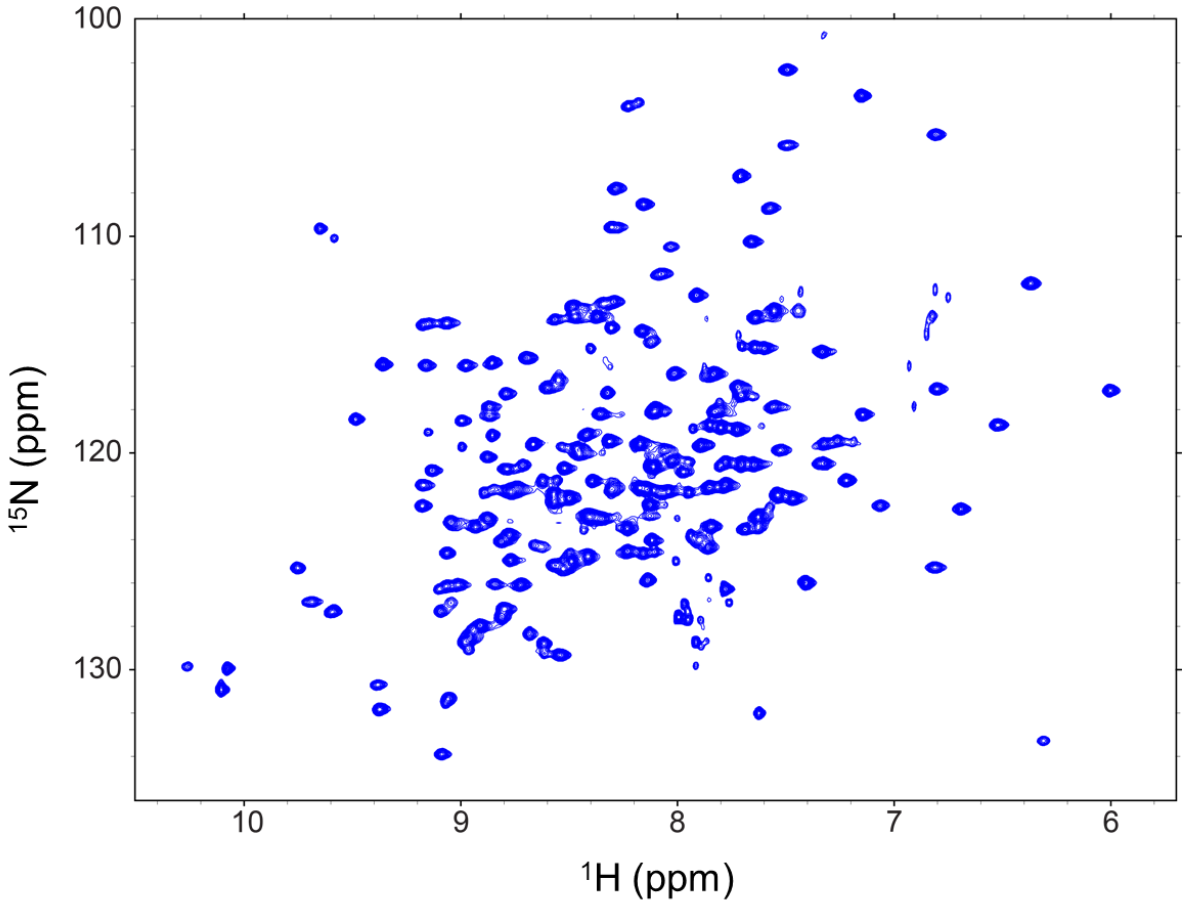

Figure S3. 2-D NMR spectrum of RsgI9S1C-CTD. A 2-D TROSY-HSQC NMR spectrum of ^{15}N-labeled RsgI9$^{S1C\text{-}CTD}$ was acquired prior to ^{15}N-TRACT correlation time determination. The spectrum indicates a well-folded protein with backbone amide signals well-dispersed.

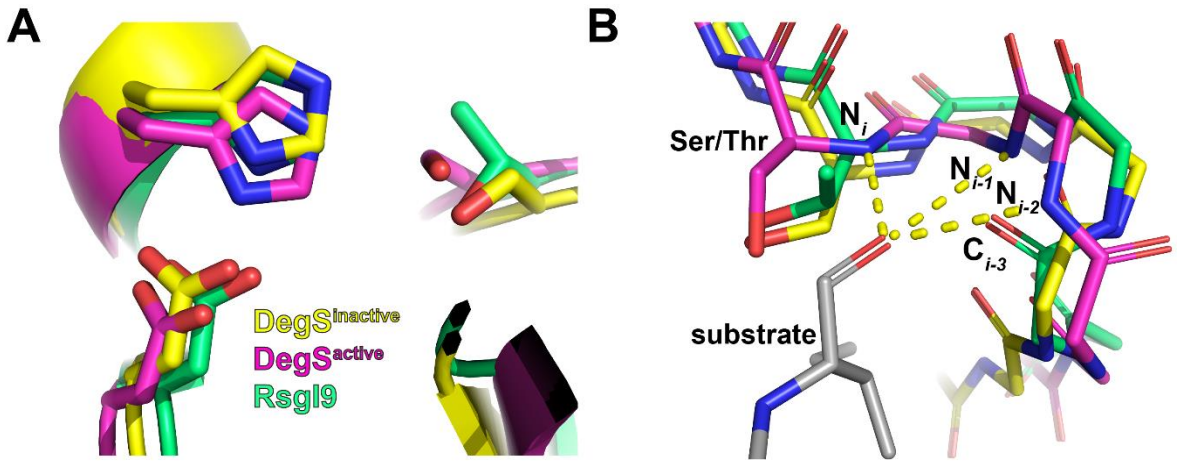

Figure S4. RsgI9's catalytic triad resembles inactive form of homolog DegS. (A) Alignment of the catalytic triad residues found in RsgI9 with those in the active (magenta, PDB: 4RQZ) and inactive (yellow, PDB: 4RQY) forms of the *E. coli* DegS protease. RsgI9's configuration more closely resembles the inactive form, as the distance between the His Nε2 and the Ser/Thr hydroxyl group is too far to make a productive hydrogen bond. (B) Alignment of the oxyanion hole residues adjacent to the nucleophilic Ser/Thr, with the same structures and coloring as in panel A. Both RsgI9 and the inactive form of DegS presumably fail to stabilize the tetrahedral intermediate during proteolysis due to a deformation relative to the active conformation. Notably, the amide N_{i-2} is flipped out while the carbonyl C_{i-3} points inward.

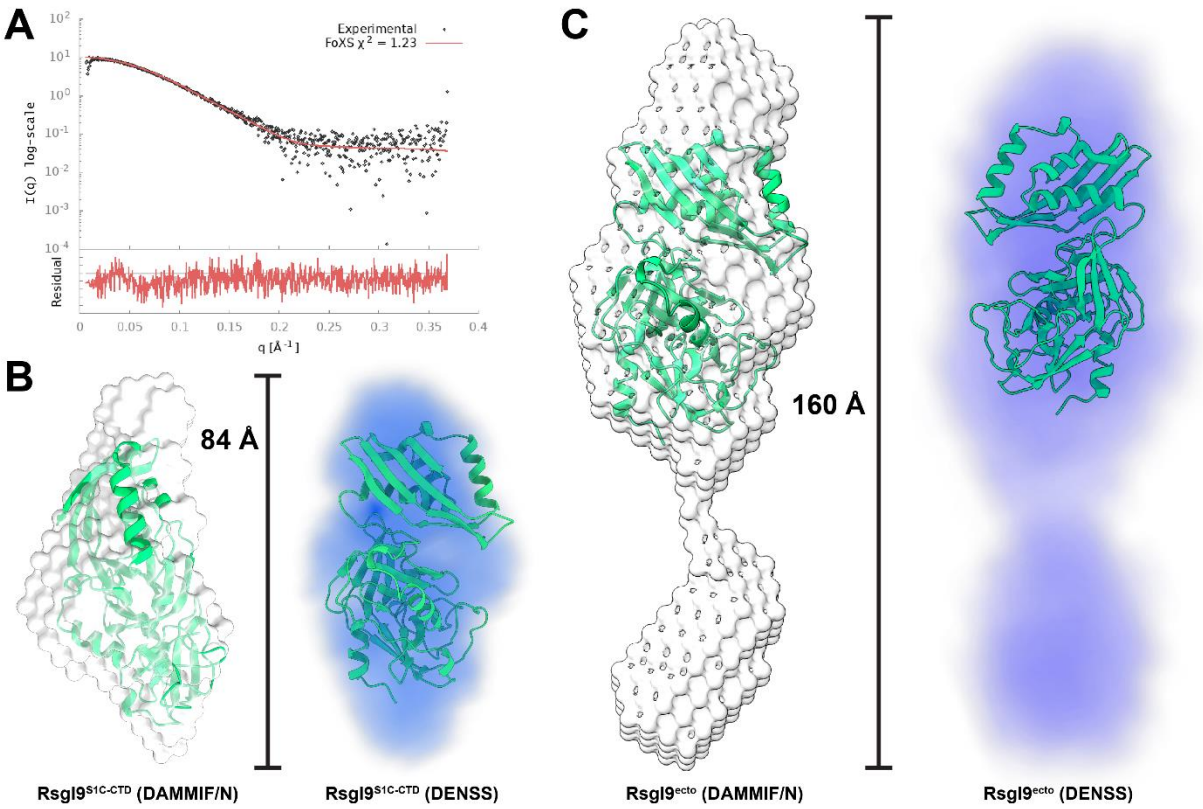

Figure S5. Fit of SAXS data and *ab initio* reconstructions. (A) Fit of experimental scattering curve to that predicted for the RsgI9$^{S1C\text{-}CTD}$ crystal structure by the FoXS server (χ^2 value of 1.23). (B) *Ab initio* bead model and electron density reconstructions of RsgI9$^{S1C\text{-}CTD}$, using DAMMIF/N and DENSS, respectively. The maximum dimension of the model (84 Å) is consistent with the D_{max} observed from the data (82 Å). (C) *Ab initio* models for intact RsgI9^{ecto}. The length of the model is in agreement with the value of D_{max} derived from the SAXS data (160 Å).

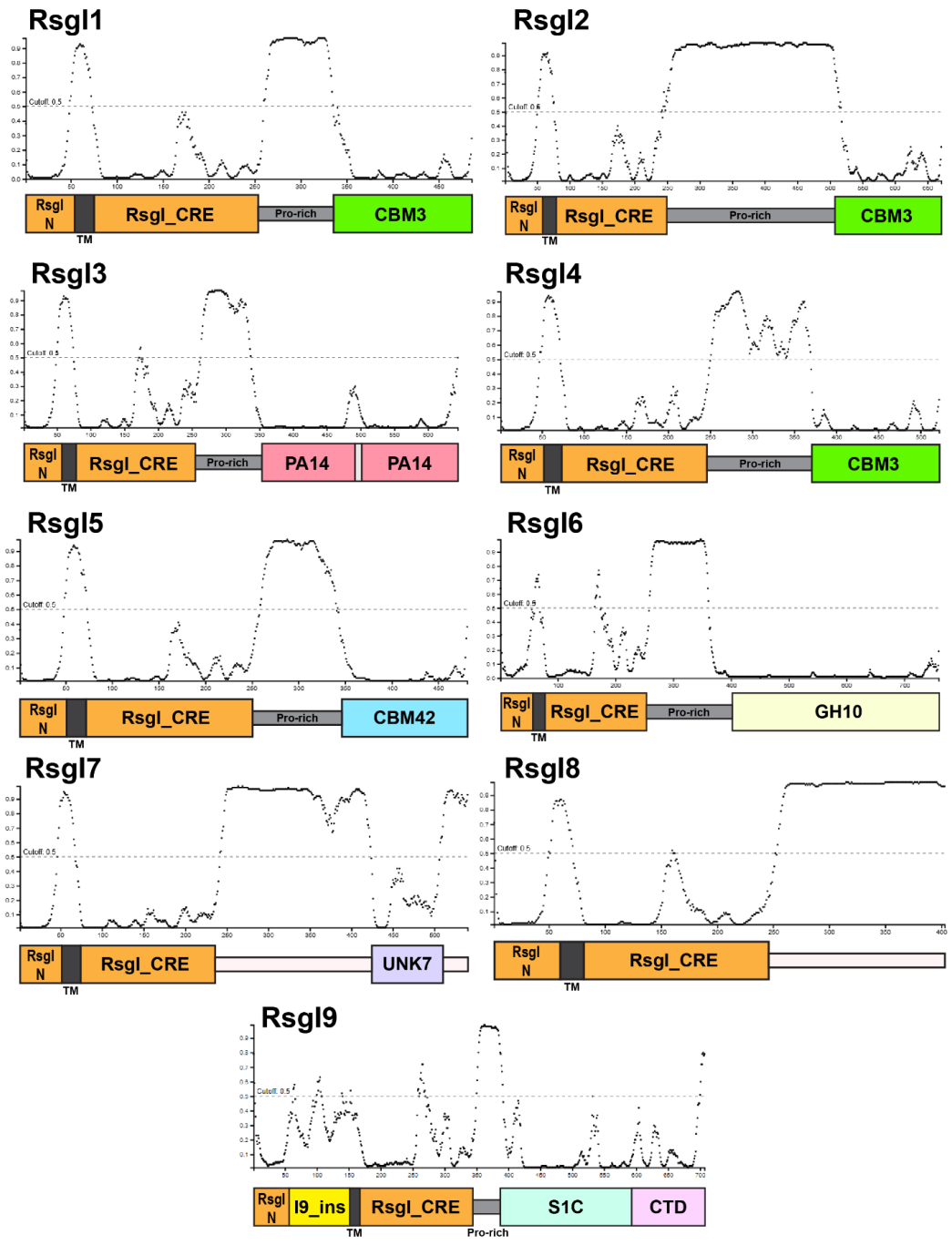

Figure S6. *C. thermocellum* RsgI architecture vs. DISOPRED3 disorder predictions. The primary sequences of the nine RsgI proteins from *Clostridium thermocellum* were analyzed using DISOPRED3[2]. Values above 0.5 are predicted to be disordered, while those below are likely structured. The domain architecture of each RsgI is shown below each plot based on previous structures, UniProtKB annotation, and this work[3-6]. A good agreement is seen between annotated domains and predicted structured regions.

REFERENCES

1. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology.* 2011;7(1):539.
2. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics.* 2015;31(6):857-863.
3. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. In: Humana Press; 2007:89-112.
4. Yaniv O, Fichman G, Borovok I, et al. Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of Clostridium thermocellum anti-sigmaI factors. *Acta Crystallogr D Biol Crystallogr.* 2014;70(Pt 2):522-534.
5. Grinberg IR, Yaniv O, Ora LO, et al. Distinctive ligand-binding specificities of tandem PA14 biomass-sensory elements from Clostridium thermocellum and Clostridium clariflavum. *Proteins: Structure, Function, and Bioinformatics.* 2019;87(11):917-930.
6. Bahari L, Gilad Y, Borovok I, et al. Glycoside hydrolases as components of putative carbohydrate biosensor proteins in Clostridium thermocellum. *Journal of Industrial Microbiology & Biotechnology.* 2011;38(7):825-832.