

Crystal Structure of an Uncommon Cellulosome-Related Protein Module from *Ruminococcus flavefaciens* That Resembles Papain-Like Cysteine Peptidases

Maly Levy-Assaraf^{1,2,3}, Milana Voronov-Goldman^{1,2,3}, Inna Rozman Grinberg^{1,2}, Gloria Weiserman¹, Linda J. W. Shimon³, Sadanari Jindou^{1,4}, Ilya Borovok¹, Bryan A. White⁵, Edward A. Bayer⁶, Raphael Lamed^{1,2}, Felix Frolow^{1,2*}

1 Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv, Israel, **2** The Daniella Rich Institute for Structural Biology, Tel Aviv University, Tel Aviv, Israel, **3** Department of Chemical Research Support, The Weizmann Institute of Science, Rehovot, Israel, **4** Faculty of Agriculture, Meijo University, Nagoya, Japan, **5** Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **6** Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot, Israel

Abstract

Background: *Ruminococcus flavefaciens* is one of the predominant fiber-degrading bacteria found in the rumen of herbivores. Bioinformatic analysis of the recently sequenced genome indicated that this bacterium produces one of the most intricate cellulosome systems known to date. A distinct ORF, encoding for a multi-modular protein, RflaF_05439, was discovered during mining of the genome sequence. It is composed of two tandem modules of currently undefined function that share 45% identity and a C-terminal X-dockerin modular dyad. Gaining insight into the diversity, architecture and organization of different types of proteins in the cellulosome system is essential for broadening our understanding of a multi-enzyme complex, considered to be one of the most efficient systems for plant cell wall polysaccharide degradation in nature.

Methodology/Principal Findings: Following bioinformatic analysis, the second tandem module of RflaF_05439 was cloned and its selenium-labeled derivative was expressed and crystallized. The crystals belong to space group $P2_1$ with unit-cell parameters of $a = 65.81$, $b = 60.61$, $c = 66.13$ Å, $\beta = 107.66^\circ$ and contain two protein molecules in the asymmetric unit. The crystal structure was determined at 1.38-Å resolution by X-ray diffraction using the single-wavelength anomalous dispersion (SAD) method and was refined to R_{factor} and R_{free} of 0.127 and 0.152 respectively. The protein molecule mainly comprises a β -sheet flanked by short α -helices, and a globular α -helical domain. The structure was found to be structurally similar to members of the NlpC/P60 superfamily of cysteine peptidases.

Conclusions/Significance: The 3D structure of the second repeat of the RflaF_05439 enabled us to propose a role for the currently undefined function of this protein. Its putative function as a cysteine peptidase is inferred from *in silico* structural homology studies. It is therefore apparent that cellulosomes integrate proteins with other functions in addition to the classic well-defined carbohydrate active enzymes.

Citation: Levy-Assaraf M, Voronov-Goldman M, Rozman Grinberg I, Weiserman G, Shimon LJW, et al. (2013) Crystal Structure of an Uncommon Cellulosome-Related Protein Module from *Ruminococcus flavefaciens* That Resembles Papain-Like Cysteine Peptidases. PLoS ONE 8(2): e56138. doi:10.1371/journal.pone.0056138

Editor: Bostjan Kobe, University of Queensland, Australia

Received: October 27, 2012; **Accepted:** January 5, 2013; **Published:** February 14, 2013

Copyright: © 2013 Levy-Assaraf et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This was supported by regular and energy grants from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel and by the Israel Science Foundation (grants 291/08, 159/07 and 966/09). Additional support was received by a grant (No. 24/11) issued to RL by the Sidney E. Frank Foundation through the Israel Science Foundation (ISF). (<http://www.isf.org.il/>) (<http://www.bsf.org.il/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mbfrolow@post.tau.ac.il

These authors contributed equally to this work.

Introduction

Plant cell wall polysaccharides offer an extraordinary source of carbon and energy that can be utilized by various microorganisms, thus contributing a central component to the carbon cycle. In some cases, free-living microorganisms exploit such polysaccharides from decaying plants; in other cases symbiotic microbes assist both invertebrates (e.g., termites) and higher vertebrates (e.g., ruminants) in converting plant-derived polysaccharides to digest-

ible components. Cellulose is the main structural component of the plant cell wall. It is arranged in highly recalcitrant fibrils which are usually embedded in a colloidal matrix of hemicellulose and lignin [1].

The cellulolytic, fiber-degrading bacterium *Ruminococcus flavefaciens* is one of the critically important inhabitants in the rumen of herbivores, which plays a central role in the degradation of plant cell wall fiber. *R. flavefaciens* FD-1 is a gram-positive cellulosome-

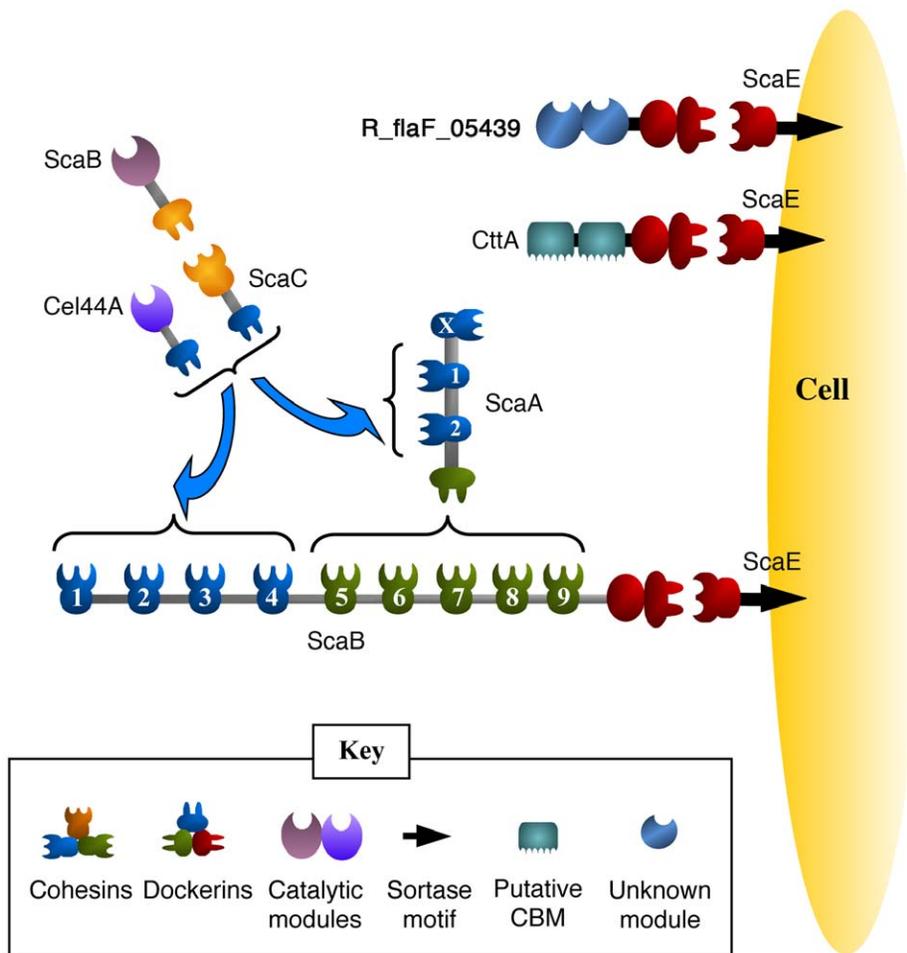


Figure 1. Schematic model of the proposed arrangement of the *R. flavefaciens* FD-1 cellulosomal components. The components of the cellulosome include the various scaffoldins (ScaA, ScaB, ScaC and ScaE) and the dockerin-containing catalytic subunits (represented in the figure by the Ce3B-type and Cel44A-type). ScaE is implanted into the cell surface via a sortase-like signal motif. Like the cellulosomal ScaB, two other cellulosome-related proteins, CttA and RflaF_05439, contain an XDoc and are also attached to the cell surface via the ScaE-Coh. The *R. flavefaciens* FD-1 cellulosome is organized as follows: The ScaA cohesins bind either to the ScaC dockerin or to the various Cel44A-type dockerins, and the lone ScaC cohesin serves as an “adaptor scaffoldin”, which selectively binds to the Ce3B-type dockerins and incorporates their parent proteins into the complex. ScaA is attached via its dockerin into ScaB cohesins 5–9, whereas the specificity of ScaB cohesins 1–4 is similar to those of ScaA. doi:10.1371/journal.pone.0056138.g001

producing anaerobic bacterium [2–6] that possesses the most elaborate cellulosome system thus far discovered in nature. Over 220 different dockerin-containing open reading frames (ORFs) have been identified in this strain [7,8], which is more than triple the number so far detected in the original model bacterium, *Clostridium thermocellum*, from which the cellulosome concept was first described. Only 36% of these ORFs were predicted to be carbohydrate-acting enzymes. Other ORFs were bioinformatically predicted as structural and catalytic protein modules, such as proteases (peptidases), transglutaminases, lipases and leucine-rich repeats (LRRs) [9]. These components have not been known previously to be associated with polysaccharide degradation, and

their presence as part of the cellulosome system is particularly intriguing.

In *R. flavefaciens*, the key cellulosomal scaffoldin components (ScaA, ScaB, ScaC, CttA and ScaE) are organized into a *sca* gene cluster whose presence has been documented in five different strains of this species [10]. The complement of interconnecting cellulosomal components is anchored to the cell surface via interaction of the X-dockerin (XDoc) modular dyad of the ScaB subunit with the cohesin (Coh) of ScaE [11], which is covalently implanted into the cell surface via a sortase-like signal motif (Figure 1). The ScaE-Coh also binds to the same type of XDoc module present on another cellulosome-related protein, CttA. The

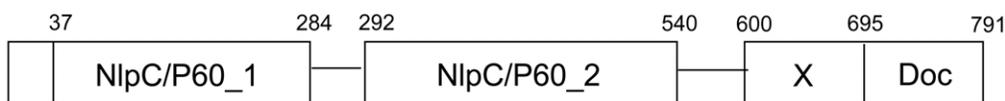


Figure 2. Schematic representation of the modular structure of RflaF_05439. The protein contains a signal peptide followed by two modules, predicted to belong to the NlpC/P60 superfamily (NlpC/P60_1 and NlpC/P60_2) and an X-dockerin (XDoc) modular dyad. doi:10.1371/journal.pone.0056138.g002

modular dyad of the CttA homologue from *R. flavefaciens* 17 was demonstrated to bind to the cellulose substrate [12].

A third XDoc modular dyad was identified by bioinformatics data mining of the *R. flavefaciens* FD-1 genome. The parent protein, RflaF_05439 (accession number: ZP_06142651), consists of a tandem repeat of two conserved modules (Figure 2) at the N terminus and a C-terminal XDoc, which also binds to ScaE-Coh (unpublished results). Protein BLAST [13] searches of the NCBI database using either of the latter repeated segments failed to reveal similar sequences in other species, and functional information was not evident prior to determination of the structure described in this work. In the present communication, we present the crystal structure of the second tandem module of this protein. *In silico* analysis of the structure implied that the protein belongs to an established superfamily of cysteine-like peptidases, termed NlpC/P60, and this domain will hereafter be coined NlpC/P60_2.

Experimental Methods

Cloning, Expression and Purification

The DNA encoding NlpC/P60_2, the second tandem module of the parent protein RflaF_05439, was amplified from genomic DNA of *R. flavefaciens* FD-1 using F- CCGCTAGCATGTA-CAATTCCGACGGCTGGTACA and R- CCGCTCGAGT-TACGTTACCACAGGTTCTGCTTTCT primers. The insert was cloned into the pET28a expression vector (Novagen) with a sequence encoding for a hexa-His tag attached to the 5' end, using *NdeI* and *XhoI* restriction enzymes. The construct was expressed in *Escherichia coli* strain BL21(DE3) RIL. Expression of the seleno-l-methionine-labelled NlpC/P60_2 was conducted according to the method described previously [14]. The expressed His-tagged protein was purified by metal-chelate affinity chromatography using a Ni-IDA resin (Rimom Biotech, Israel) with phosphate buffer (50 mM NaH₂PO₄, pH 6.0, 300 mM NaCl and 10% glycerol). Fast protein liquid chromatography (FPLC) using an AKTA-prime system (Amersham Pharmacia Biotech, Piscataway, NJ) was performed on a Superdex[®] 75 XK 16 (Pharmacia Biotech) pre-packed column in column buffer (50 mM Tris pH 7.5, 0.15 M NaCl, 0.05% sodium azide). The protein was concentrated using Centriprep YM-3 centrifugal filter devices (Amicon Bioseparation, Millipore, Inc., Billerica), yielding 1.5 ml of purified concentrated protein (20 mg ml⁻¹). Protein concentration was determined by measuring the absorbance at a wavelength of 280 nm using the calculated extinction coefficient of the protein (ϵ_{280} [g/l] = 0.92).

Crystallization, Data Collection and Processing

Crystallization was performed using a Hampton Research Index-HTM kit, employing the microbatch crystallization method under 1:1 mixture of silicon and paraffin oil [15] at 293 K. Samples were dispensed using an Oryx 6 Crystallization Robot from Douglas Instruments (<http://www.douglas.co.uk/>). Initial SeMet NlpC/P60_2 multi-layered crystals were obtained after 6 days in a 2- μ l drop containing 1- μ l protein solution and 1 μ l of reservoir solution consisting of 0.2 M ammonium acetate, 0.1 M BIS-TRIS pH 6.5 and 25% w/v polyethylene glycol 3350. Further optimization experiments were performed using the hanging-drop method in a 4- μ l drop consisting of 2- μ l protein solution and 2- μ l reservoir solution equilibrated against 0.4-ml reservoir solution. The streak seeding technique [16] with an eyelash (Ted Pella, Inc; www.tedpella.com) was employed in order to improve clustered crystals. Crystals from the final reservoir solution consist of 0.2 M ammonium acetate, 0.1 M BIS-TRIS pH 6.5 and 23% (w/v) polyethylene glycol 3350. Crystals were harvested from the crystallization drop using a MiTeGen

Table 1. Data-collection, processing and refinement statistics for crystal of SeMet derivative of NlpC/P60_2.

Experimental conditions	
X-ray source	ESRF ID23-1
Wavelength (Å)	0.9796
Temperature (K)	100
Detector	CCD ADSC
Crystal parameters	
Space group	P2 ₁
Unit cell parameters (Å and °):	
a	65.81
b	60.61
c	66.13
β	107.66
Resolution (Å)	32.4–1.38 (1.40–1.38)
Data processing	
No. of reflections	715274
No. of unique reflections #	182049
Completeness (%)	95.9 (83.5)
Mean I/ σ (I)	37.5 (3.5)
R _{merge} ##	0.051 (0.449)
Redundancy	7.3(5.6)
Structure refinement	
Resolution range (Å)	32.39–1.38
R _{work}	0.127 (0.1567)
R _{free}	0.152 (0.1782)
No. of reflections in test set ###	3825
No. of reflections used in refinement ###	178224
Average B factor (Å ²)	
Overall	18.3
Atoms in protein	14.4
Atoms of ligands	39.1
Atoms of solvent	20.9
R.M.S. deviation, bond lengths (Å)	0.009
R.M.S. deviation, bond angles (°)	1.203
Ramachandran statistics	
Permitted (%)	2.9
Favored (%)	97.1

Values for the highest resolution shell are given in parentheses.

$R_{merged} = \frac{\sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle|}{\sum_{hkl} \sum_j I_j(hkl)}$, where \sum_{hkl} denotes the sum over all reflections and \sum_j the sum over all equivalent and symmetry related reflections [44].

##Friedel pairs of reflections were not merged to preserve the anomalous signal.

###For the participation in the test set both reflections comprising a Friedel pair were selected.

doi:10.1371/journal.pone.0056138.t001

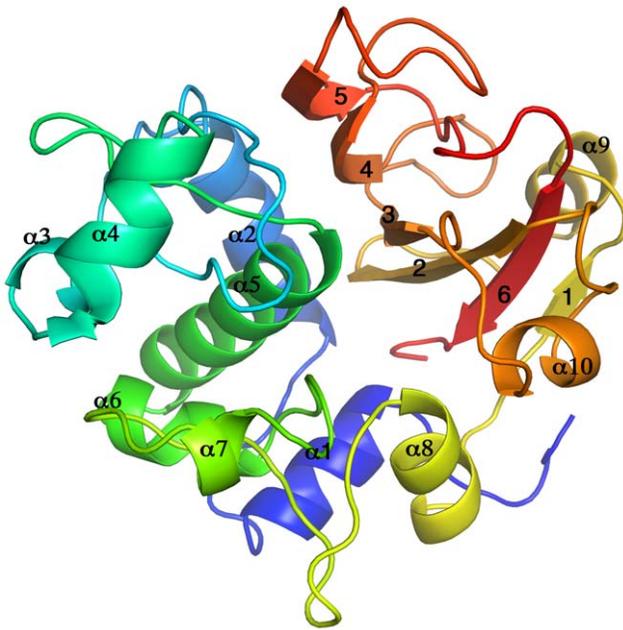


Figure 3. Cartoon diagram of the overall three-dimensional structure of NlpC/P60_2. Rainbow-colored α -helices and β -strands are shown from blue at the N-terminus to red at the C-terminus. The β -sheet (formed by antiparallel β -strands 1-6-2-3-4-5) is flanked by short α -helices (3-4-7-9) and a globular α -helical domain (formed by α -helices 1-2-5-6). All 10 α -helices are located at the N terminus, followed by the six β -strands.

doi:10.1371/journal.pone.0056138.g003

MicroMount (<http://www.mitegen.com>) made of polyimide and transferred for a short time into a cryo-stabilization solution mimicking the mother liquor supplemented with 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 and flash-cooled in a nitrogen stream at a temperature of 100 K produced by an Oxford Cryostream low-temperature generator [17]. Single-wavelength Se-SAD diffraction data were measured to a resolution of 1.38 Å with wavelength tuned to 0.9795 Å to enhance the anomalous signal from the Se atoms. Oscillation range of 0.5° per frame over 360° rotation was implemented. Data were indexed, integrated and scaled using DENZO and SCALEPACK as implemented in HKL2000 [18]. During scaling, Friedel pairs of reflections were kept separate for subsequent use. Details of the data-collection statistics are given in Table 1.

Structure Determination and Refinement

The phases of the structure were determined using *SHELXC/D/E* [19] as employed in the HKL2MAP [20] graphical user interface. Heavy atom substructure containing all ten possible Se-atom sites was identified by *SHELEXD*; primary phasing, phase modification and initial $C\alpha$ tracing were obtained using *SHELXE*. After density modification by DM [21], auto model building was implemented in Arp/WARP [22]. The final structure was refined using PHENIX [23] to a crystallographic R_{work} and R_{free} of 12.91% and 15.32%, respectively. During refinement, Friedel pairs of reflections were kept separated, anomalous dispersion parameters f' and f'' were not refined, and manual corrections and structure validation were performed using *COOT* [24]. A total of 494 amino acid residues and 722 water molecules were present in the final model. Coordinates and structure factor amplitudes for

NlpC/P60_2 have been deposited in the PDB (accession code 4EYZ).

Results and Discussion

Overall Structure of the RflaF_05439 Second Repeat

The X-ray crystal structure of the second module of the tandem repeat of RflaF_05439 from *R. flavefaciens* FD-1 was determined at a resolution of 1.38 Å. There are no significant differences between the two molecules in the asymmetric unit (RMSD of 0.06 Å for all $C\alpha$ atoms). The N-terminal His tag was not modeled into the structure owing to the weak electron density. The overall fold consists of two globular domains that form a tight interface: an N-terminal α -helical domain and a C-terminal domain comprised of 6 antiparallel β -strands covered by short α -helices and loops (Figure 3).

Detection of the NlpC/P60 Superfamily Signature

In an attempt to elucidate a biological function for RflaF_05439, structural homology searches of the RCSB Protein Data Bank queried with the structure of the second module of the tandem repeat were performed using the DALI search engine [25]. Notable similarity was observed to a putative “xylanase” from *Bacteriodes fragilis* (PDB entry 2P1G, Z score of 7.8, unpublished) and to a putative endopeptidase from *Anabaena variabilis* (PDB entry 2HBW, Z score of 7.5, [26]) that belong to the NlpC/P60 superfamily.

NlpC/P60 is an example of a superfamily with members sharing structural similarities that cannot be detected at the sequence level [27]. The NlpC/P60 superfamily of peptidases is characterized by a conserved catalytic triad, which contains cysteine, histidine and a polar residue (i.e., glutamine, asparagine, histidine, glutamate or aspartate) within a catalytic core. Beyond the catalytic core, the structures are highly divergent.

Identification of the second module of the tandem repeat as a candidate member of the NlpC/P60 superfamily according to its 3D structure facilitated further analysis of sequence homology. Interestingly, we found and sequenced a homologous RflaF_05439 gene from the closely related *R. flavefaciens* strain 17. It should be noted that although the two strains (FD-1 and 17) belong to the same species, their cellulosomal organization is, in some of their respective components, dissimilar [10,28]. The homologous gene shows a similar organization of two repeated modules separated by a linker sequence and a C-terminal XDoc. RflaF_05439 of *R. flavefaciens* FD-1, and its homolog of strain 17 shares overall sequence similarity of 39%, with increased similarity of 71% between the respective second NlpC/P60 repeats.

Multiple sequence alignment of the repeated NlpC/P60 modules of RflaF_05439 from the two strains and five other members of the NlpC/P60 superfamily was performed. The alignment emphasizes the conservation of the catalytic triad (Figure 4). The catalytic histidine residue [29] is strictly conserved among all the NlpC/P60 superfamily members presented in the figure, whereas the catalytic polar residue is divergent. The catalytic cysteine is highly conserved with the notable exception of the first repeat of the homologous module from *R. flavefaciens* 17. In the latter module, the conserved catalytic cysteine is substituted by serine. The appearance of a catalytic serine, a catalytic histidine and a catalytic polar residue characterizes serine proteases and lipases [27]. Other elements characterizing the NlpC/P60 superfamily, such as specific polar residues, glycines, hydrophobic residues, and small amino acids [27], are also highly conserved in the RflaF_05439 modules. Owing to this analysis, the two tandem

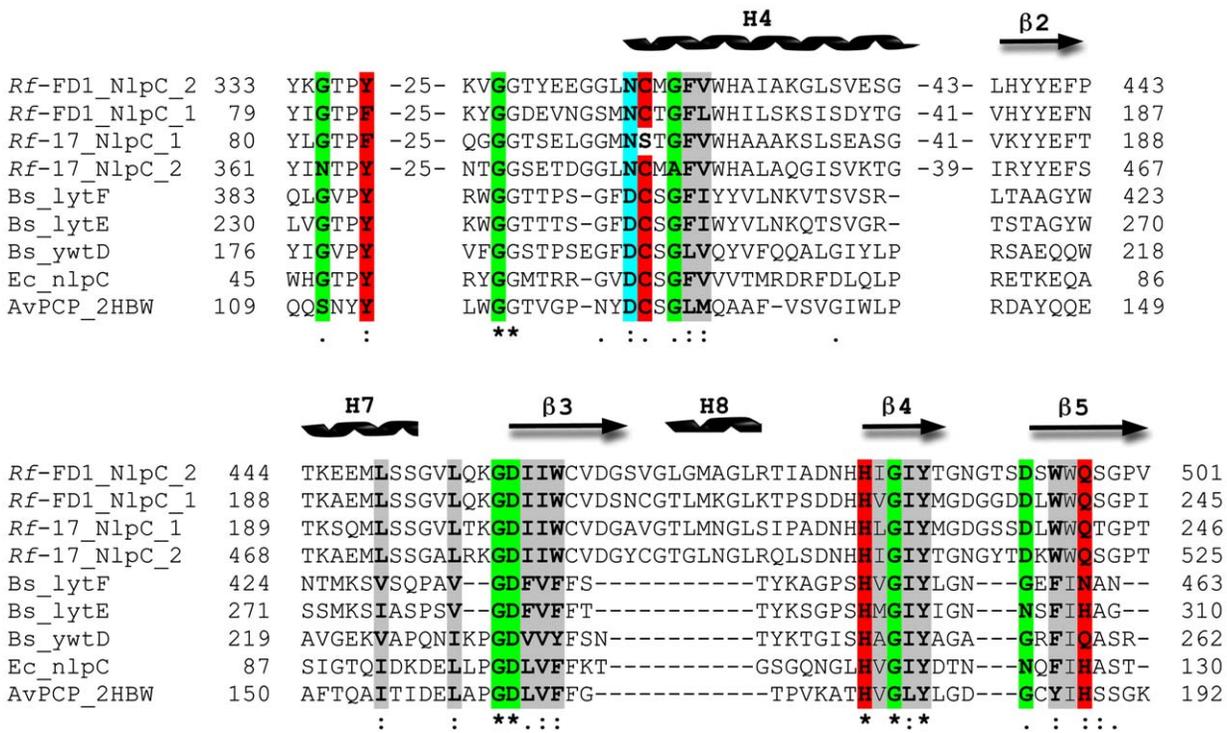


Figure 4. Multiple sequence alignment of selected portions of the RflaF_05439 NlpC/P60 repeats from *R. flavefaciens* strain FD-1 with those of the homologous protein in strain 17 and selected members of the NlpC/P60 superfamily. Multiple sequence alignments were constructed using ClustalW. The numbers within the alignment represent corresponding stretches of non-conserved residues that are not shown. Highly conserved residues in the NlpC/P60 superfamily are highlighted as follows: red – catalytic (including Tyr which is responsible for oxyanion hole formation), gray - hydrophobic, green - small, cyan - polar. Symbols: asterisks - full conservation; colons - strongly similar; periods - weakly similar. The sequences are denoted by their gene name preceded by the species abbreviation: *Rf*, *R. flavefaciens*; *Bs*, *Bacillus subtilis*; *Ec*, *Escherichia coli*. AvPCP represents endopeptidase from *Anabaena variabilis* denoted by its PDB code (2HBW). Secondary structural elements (enumerated β strands and α helices) are indicated.
doi:10.1371/journal.pone.0056138.g004

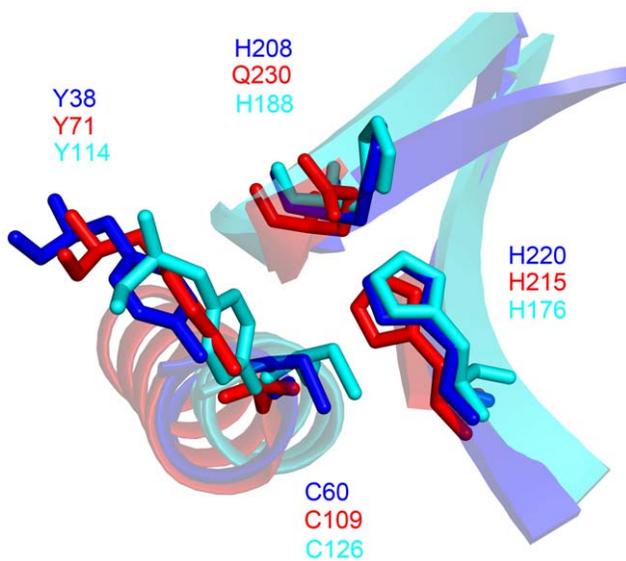


Figure 5. Structural superposition of the catalytic triad of members of the NlpC/P60 superfamily. The catalytic triad of NlpC/P60_2 (blue, PDB code 4EZY), the putative “xylanase” (red, PDB code 2P1G), and AvPCP (cyan, PDB code 2HBW). The fourth functionally important tyrosine residues in NlpC/P60_2 (Tyr71), 2P1G (Tyr38) and 2HBW (Tyr114) are also shown.
doi:10.1371/journal.pone.0056138.g005

repeated modules were hereafter referred to as NlpC/P60_1 and NlpC/P60_2, respectively.

Structural Homology with the NlpC/P60 Superfamily

Structural superposition of the putative catalytic triads of NlpC/P60_2 [30,31] and two other NlpC/P60 superfamily members – a putative “xylanase” from *Bacteriodes fragilis* (PDB code 2P1G, unpublished) and AvPCP, a peptidoglycan cysteine endopeptidase from *Anabaena variabilis* (PDB code 2HBW, [26]) – is presented in Figure 5. The spatial configuration of all catalytic residues (Cys, His and Gln/His/Asn) is strictly conserved, with RMSD values ranging from 0.04 to 0.14 Å. A fourth residue, Tyr71, is also conserved. This residue is important for formation of a putative oxyanion hole, which helps stabilize transitional forms between the peptidase and the substrate [30]. Beyond the catalytic core, the structures are highly divergent through deletions and insertions of structural elements. We therefore suggest that the same structural arrangement of the catalytic residues in NlpC/P60_2 serves to classify the module in the NlpC/P60 superfamily.

In the NlpC/P60_2 structure, extra electron density was discovered contiguous with the side chain of the catalytic cysteine, Cys109 (Figure 6). This extraneous electron density was modeled as acetate. Both oxygen atoms of acetate fit well the electron density map and form hydrogen bonds with neighboring residues and water. The fact that we could not detect protease activity in the purified protein, coupled with the observed derivatization of the presumed active-site cysteine, would suggest that the modifi-

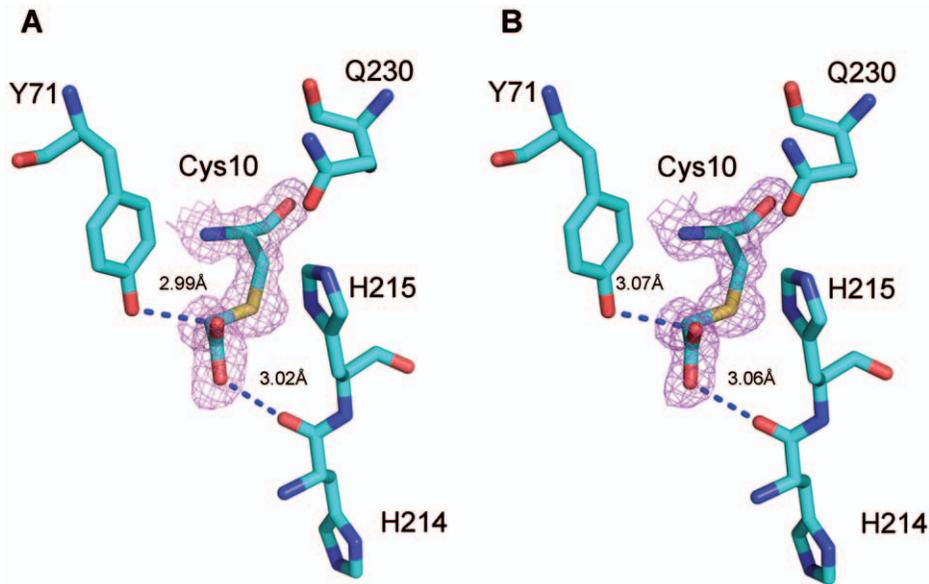


Figure 6. View of the modified active-site cysteine residue shown in both subunits. A.) Subunit A. B.) Subunit B. The 2Fobs – Fcalc electron density map, displayed around the cysteine amino acid residue and rendered at 1.2 RMSD level, is shown in magenta. Various moieties may fit the density map (see text for details), the best of which is an acetate group.
doi:10.1371/journal.pone.0056138.g006

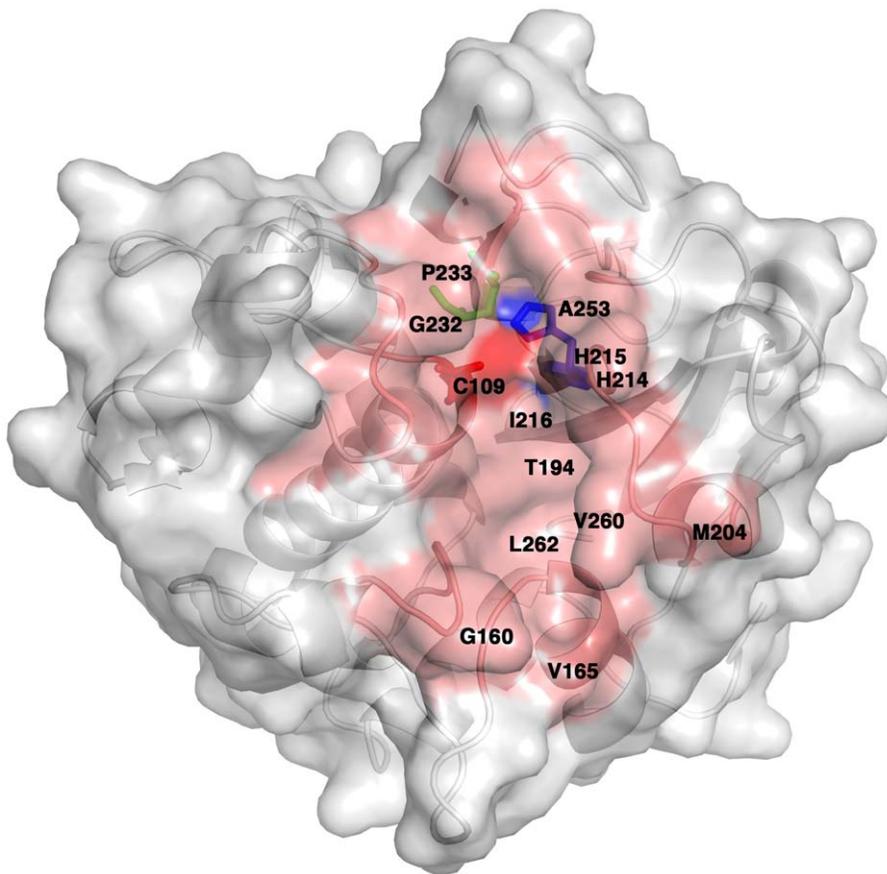


Figure 7. Putative active-site groove of NlpC/P60 2. The residues that form the groove are labeled on the surface representation of the molecule. The cysteine residue (C109) in the catalytic triad is colored red, and the histidine (H215) is colored blue. As in the other known NlpC/P60 structures, the glutamine residue (Q230) is buried. Residues that form the active site are colored salmon red, and indicated by single amino acid code and number.
doi:10.1371/journal.pone.0056138.g007

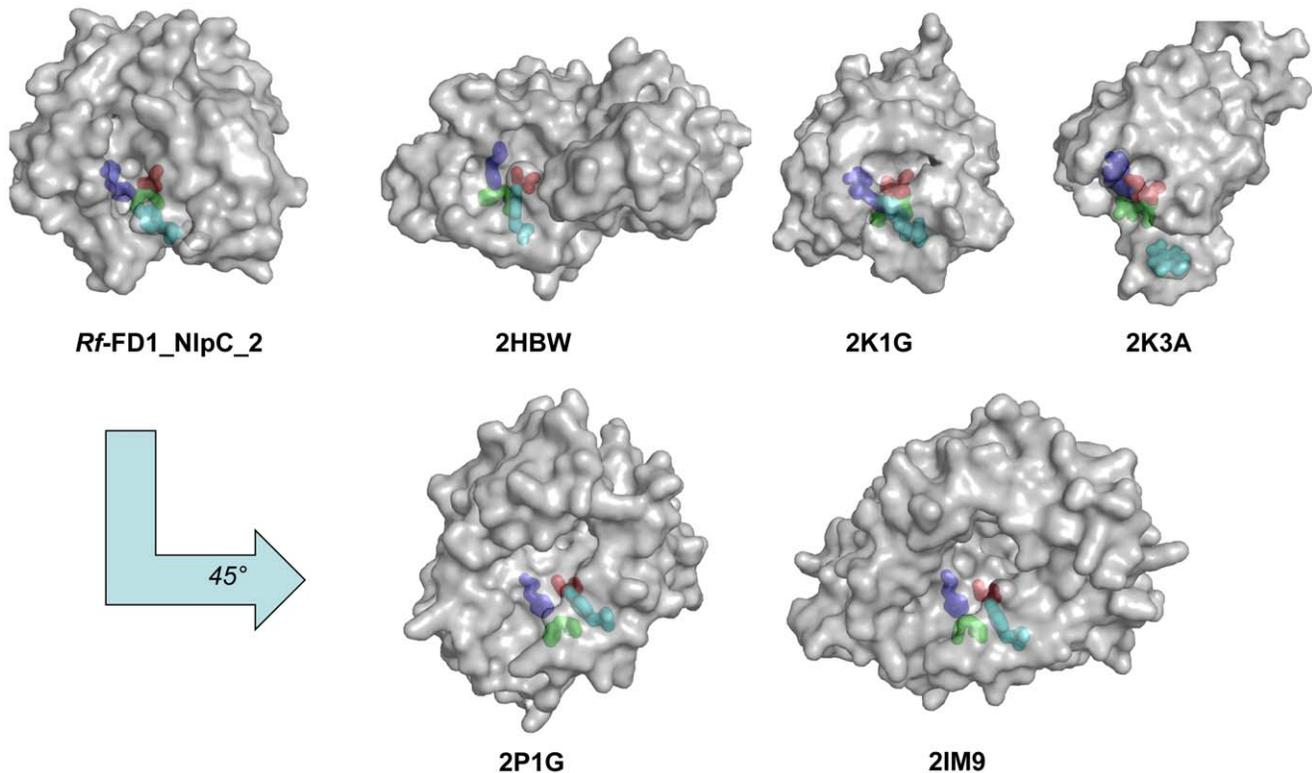


Figure 8. Location of the active-site grooves of various members of the NlpC/P60 superfamily represented on the molecular surface (annotated by PDB entry number). The catalytic triad residues are colored as follows: cysteine in red, histidine in blue, and the third polar residue in green. The active-site grooves of 2HBW, 2K1G, 2K3A are located on the same face as that of NlpC/P60_2. The active-site residues of all structures were first superimposed; in order to view the grooves of 2P1G and 2IM9. The latter structures were then rotated approximately 45° counter-clockwise, relative to NlpC/P60_2.
doi:10.1371/journal.pone.0056138.g008

cation originated during expression and/or purification, perhaps producing a suicide complex during enzymatic reaction. Further experiments will be necessary in order to determine the precise nature of this modification. Various alternatives for the observed electron density can also be considered, including ethanethioamide, acetone or acetamide. In any case, is it likely that the modified cysteine would lead to an inactivated enzyme, which may explain why attempts to demonstrate protease activity were unsuccessful.

Active Site Groove

The potential active site of NlpC/P60_2 is located in the channel groove that runs along the interface between the N- and C-terminal domains (Figure 7). The entrance to the groove is formed by the following amino acid residues: Asp73, Tyr74, Val75, Asn108, Phe138, Gly159, Thr162, Val163, Ser139, Ser140, Tyr141, Arg207, Ile210, Ala211, Asp212, Asn213 and Leu240. The middle section of the groove consists of Tyr71, Met110, Pro137, Trp161, Thr209 and Leu207. Residues lining the bottom of the groove are Cys109, Gly160, Val165, Trp194, Met204, His214, His215, Ile216, Gly232, Pro233, Ala253, Val260 and Leu262.

Compared to several other proteins from this superfamily whose structures have been determined (PDB codes 2HBW [26], 2K1G [32] and 2K3A [33]), the relative position of the active-site groove of NlpC/P60_2 is located approximately on the same face of the molecule (Figure 8). In contrast, in the structures of two other proteins (2P1G and 2IM9, both unpublished), the relative

orientation of the active-site grooves is rotated approximately 45° about the X axis with respect to NlpC/P60_2.

Potential Role of RflaF_05439 in the Cellulosome System of *R. flavefaciens*

Information concerning the various cellulosomal proteins and, in particular, uncommon components that are not of the classic cellulosome mold is necessary in order to extend the cellulosome paradigm of protein complex assembly beyond the realm of glycoside hydrolases and other carbohydrate-active enzymes. This has been enabled in recent years by extensive genome-sequencing efforts and the gradual accumulation of genome sequences of cellulosome-producing bacteria, thus providing a genome-wide view of the cellulosome components and related proteins – a recently emerging field termed cellulomics [1]. For example, the sequenced genome of the prominent anaerobic, thermophilic cellulosome-producing bacterium, *Clostridium thermocellum*, revealed the presence of genes encoding dockerin-containing serine proteinase inhibitors (serpins) [34] and a subtilisin-like serine protease [35]. Nevertheless, the great majority of the >70 dockerin-containing proteins in the *C. thermocellum* cellulosome are indeed carbohydrate-active enzymes.

In contrast to *C. thermocellum*, the cellulosomal system in *R. flavefaciens* includes an abundance of over 220 different dockerin-containing proteins, only about half of which can be classified as carbohydrate-active enzymes. The remainder includes putative peptidases, serpins, structural proteins, and proteins of unidentified function [9].

The crystal structure of the protein module described in this work allows the assignment of a putative function for the RflaF_05439 protein, which is identified to be a member of the NlpC/P60 superfamily, possessing the characteristic conserved catalytic triad residues (Cys109/His215/Gln230). Interestingly, the association of this protein module with this superfamily of peptidases was not evident from the primary sequence, but could only be determined from its structural characteristics. All functionally approved members of the NlpC/P60 family are γ -glutamyl D,L-endopeptidases. They were detected in *Bacillus subtilis* [36–39], *Streptococcus mitis* [40], *Streptococcus aureus* [41] and in bacteriophage [2]. The same structural arrangement of the catalytic core of the NlpC/P60 superfamily suggests that the mechanism of proteolysis is expected to be similar. However, at this point, definitive proof is still wanting, and it can only be speculated that RflaF_05439 functions as a cysteine peptidase.

Interestingly, NlpC/P60 proteins are usually characterized by a single catalytic NlpC/P60 domain. One exception is YwtD from *B. subtilis* that carries 3 copies of this domain while only the second repeat is functional [38]. RflaF_05439 is the only current example of a duplicated NlpC/P60 domain [42].

The dockerin module of RflaF_05439 interacts with ScaE-Coh (unpublished results), and ScaE-Coh is attached covalently to the

cell surface via a sortase-like signal motif [12,43]. Its localization on the cell surface would support involvement in interactions with the environment, and RflaF_05439 may have a novel type of functionality that remains to be determined. In this context, it might play a role in supporting cell-cell interactions or dynamics or might have other unknown regulatory roles, serving, for example, as a protease that protects the bacterium from environmental hazards. Further studies are necessary in order to understand the function of this cellulosome-related protein from *R. flavefaciens* FD-1.

Acknowledgments

We would like to thank the ESRF for synchrotron beam time and the staff scientists of the ID23-1 station for their assistance. E.A.B. is the incumbent of The Maynard I. and Elaine Wishner Chair of Bio-Organic Chemistry.

Author Contributions

Conceived and designed the experiments: IB BW RL EB FF. Performed the experiments: MLA MVG GW LS IB FF IRG SJ. Analyzed the data: MLA MVG IB FF. Contributed reagents/materials/analysis tools: RL EB FF. Wrote the paper: MLA MVG FF EB.

References

- Bayer EA, Lamed R, White BA, Flint HJ (2008) From cellulosomes to cellulosomes. *Chem Rec* 8: 364–377.
- Nelson KE, Zinder SH, Hance I, Burr P, Odongo D, et al. (2003) Phylogenetic analysis of the microbial populations in the wild herbivore gastrointestinal tract: insights into an unexplored niche. *Environ Microbiol* 5: 1212–1220.
- Jullian V, de Vaux A, Millet L, Fonty G (1999) Identification of *Ruminococcus flavefaciens* as the predominant cellulolytic bacterial species of the equine cecum. *Appl Environ Microbiol* 65: 3738–3741.
- Krause DO, Dalrymple BP, Smith WJ, Mackie RI, McSweeney CS (1999) 16S rDNA sequencing of *Ruminococcus albus* and *Ruminococcus flavefaciens*: design of a signature probe and its application in adult sheep. *Microbiology* 145 (Pt 7): 1797–1807.
- Wedekind KJ, Mansfield HR, Montgomery L (1988) Enumeration and isolation of cellulolytic and hemicellulolytic bacteria from human feces. *Appl Environ Microbiol* 54: 1530–1535.
- Flint HJ (1997) The rumen microbial ecosystem—some recent developments. *Trends Microbiol* 5: 483–488.
- Berg Miller ME, Antonopoulos DA, Rincon MT, Band M, Bari A, et al. (2009) Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of *Ruminococcus flavefaciens* FD-1. *PLoS One* 4: e6650.
- Rincon MT, Ding SY, McCrae SI, Martin JC, Aurilia V, et al. (2003) Novel organization and divergent dockerin specificities in the cellulosome system of *Ruminococcus flavefaciens*. *J Bacteriol* 185: 703–713.
- Rincon MT, Dassa B, Flint HJ, Travis AJ, Jindou S, et al. (2010) Abundance and diversity of dockerin-containing proteins in the fiber-degrading rumen bacterium, *Ruminococcus flavefaciens* FD-1. *Plos One* 5, e12476.
- Jindou S, Brule JM, Levy-Assaraf M, Rincon MT, Flint HJ, et al. (2008) Cellulosome gene cluster analysis for gauging the diversity of the ruminal cellulolytic bacterium *Ruminococcus flavefaciens*. *FEMS Microbiol Lett* 285: 188–194.
- Rincon MT, Cepeljnik T, Martin JC, Lamed R, Barak Y, et al. (2005) Unconventional mode of attachment of the *Ruminococcus flavefaciens* cellulosome to the cell surface. *J Bacteriol* 187: 7569–7578.
- Rincon MT, Cepeljnik T, Martin JC, Barak Y, Lamed R, et al. (2007) A novel cell surface-anchored cellulose-binding protein encoded by the sca gene cluster of *Ruminococcus flavefaciens*. *J Bacteriol* 189: 4774–4783.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
- Van Duyn GD, Standaert RF, Karplus PA, Schreiber SL, Clardy J (1993) Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J Mol Biol* 229: 105–124.
- Chayen NE, Stewart PDS, Maeder DL, Blow DM (1990) An automated-system for microbatch protein crystallization and screening. *J Appl Crystallogr* 23: 297–302.
- Bergfors T (2003) Seeds to crystals. *J Struct Biol* 142: 66–76.
- Cosier J, Glazer AM (1986) A Nitrogen-gas-stream cryostat for general X-ray-diffraction studies. *J Appl Crystallogr* 19: 105–107.
- Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276: 307–326.
- Sheldrick GM (2010) Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr D, Biol Crystallogr* 66: 479–485.
- Pape T, Schneider TR (2004) HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. *J Appl Crystallogr* 37: 843–844.
- Cowtan K, Main P (1998) Miscellaneous algorithms for density modification. *Acta Crystallogr D Biol Crystallogr* 54: 487–493.
- Perrakis A, Morris R, Lamzin VS (1999) Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* 6: 458–463.
- Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66: 213–221.
- Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D, Biol Crystallogr* 66: 486–501.
- Holm L, Kaariainen S, Rosenstrom P, Schenkel A (2008) Searching protein structure databases with DALI-Lite v.3. *Bioinformatics* 24: 2780–2781.
- Xu QP, Sudek S, McMullan D, Miller MD, Geierstanger B, et al. (2009) Structural basis of murein peptide specificity of a γ -D-glutamyl-L-diamino acid endopeptidase. *Structure* 17: 303–313.
- Anantharaman V, Aravind L (2003) Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome Biol* 4: R11.
- Jindou S, Borovok I, Rincon MT, Flint HJ, Antonopoulos DA, et al. (2006) Conservation and divergence in cellulosome architecture between two strains of *Ruminococcus flavefaciens*. *J Bacteriol* 188: 7971–7976.
- Chayen NE (2003) Protein crystallization for genomics: throughput versus output. *J Struct Funct Genomics* 4: 115–120.
- McGrath ME (1999) The lysosomal cysteine proteases. *Annu Rev Biophys Biomol Struct* 28: 181–204.
- Gallo M, Pennestri M, Bottomley MJ, Barbato G, Eliseo T, et al. (2009) Binding of a noncovalent inhibitor exploiting the S' region stabilizes the hepatitis C virus NS3 protease conformation in the absence of cofactor. *J Mol Biol* 385: 1142–1155.
- Aramini JM, Rossi P, Huang YJ, Zhao L, Jiang M, et al. (2008) Solution NMR structure of the NlpC/P60 domain of lipoprotein Spr from *Escherichia coli*: structural evidence for a novel cysteine peptidase catalytic triad. *Biochemistry* 47: 9715–9717.
- Rossi P, Aramini JM, Xiao R, Chen CX, Nwosu C, et al. (2009) Structural elucidation of the Cys-His-Glu-Asn proteolytic relay in the secreted CHAP domain enzyme from the human pathogen *Staphylococcus saprophyticus*. *Proteins* 74: 515–519.
- Kang S, Barak Y, Lamed R, Bayer EA, Morrison M (2006) The functional repertoire of prokaryote cellulosomes includes the serpin superfamily of serine proteinase inhibitors. *Mol Microbiol* 60: 1344–1354.
- Zverlov VV, Kellermann J, Schwarz WH (2005) Functional subgenomics of *Clostridium thermocellum* cellulosomal genes: identification of the major catalytic components in the extracellular complex and detection of three new enzymes. *Proteomics* 5: 3646–3653.

36. Margot P, Pagni M, Karamata D (1999) *Bacillus subtilis* 168 gene *lytF* encodes a γ -D-glutamate-*meso*-diaminopimelate muropeptidase expressed by the alternative vegetative sigma factor, σ^D . *Microbiology* 145: 57–65.
37. Ohnishi R, Ishikawa S, Sekiguchi J (1999) Peptidoglycan hydrolase LytF plays a role in cell separation with CwlF during vegetative growth of *Bacillus subtilis*. *J Bacteriol* 181: 3178–3184.
38. Suzuki T, Tahara Y (2003) Characterization of the *Bacillus subtilis* *ywtD* gene, whose product is involved in gamma-polyglutamic acid degradation. *J Bacteriol* 185: 2379–2382.
39. Yamaguchi H, Furuhashi K, Fukushima T, Yamamoto H, Sekiguchi J (2004) Characterization of a new *Bacillus subtilis* peptidoglycan hydrolase gene, *yvcE* (named *cwlO*), and the enzymatic properties of its encoded protein. *J Biosci Bioeng* 98: 174–181.
40. Llull D, Lopez R, Garcia E (2006) Skl, a novel choline-binding N-acetylmuramoyl-L-alanine amidase of *Streptococcus mitis* SK137 containing a CHAP domain. *Febs Letters* 580: 1959–1964.
41. Kajimura J, Fujiwara T, Yamada S, Suzawa Y, Nishida T, et al. (2005) Identification and molecular characterization of an N-acetylmuramyl-L-alanine amidase Sle1 involved in cell separation of *Staphylococcus aureus*. *Molecular Microbiology* 58: 1087–1101.
42. Copley SD (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr Opin Chem Biol* 7: 265–272.
43. Alber O, Noach I, Rincon MT, Flint HJ, Shimon LJ, et al. (2009) Cohesin diversity revealed by the crystal structure of the anchoring cohesin from *Ruminococcus flavefaciens*. *Proteins* 77: 699–709.
44. Stout GH, Jensen LH (1968) X-ray Structure Determination. A Practical Guide. London: Macmillan.