Acta Crystallographica Section F
Structural Biology
and Crystallization
Communications

ISSN 1744-3091

Expression, purification, crystallization and preliminary crystallographic analysis of a putative *Clostridium difficile* surface protein Cwp19

Jonathan M. Kirby,[a,b] Nethaji Thiyagarajan,[a] April K. Roberts,[b] Clifford C. Shone[b] and K. Ravi Acharya[a]*

[a]Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, England, and [b]Research Department, Health Protection Agency, Porton Down, Salisbury SP4 0JG, England

Correspondence e-mail: bsskra@bath.ac.uk

Received 16 March 2011
Accepted 3 May 2011

Cwp19 is a putatively surface-located protein from *Clostridium difficile*. A recombinant N-terminal protein (residues 27–401) lacking the signal peptide and the C-terminal cell-wall-binding repeats (PFam04122) was crystallized using the sitting-drop vapour-diffusion method and diffracted to 2 Å resolution. The crystal appeared to belong to the primitive monoclinic space group $P2_1$, with unit-cell parameters $a = 109.1$, $b = 61.2$, $c = 109.2$ Å, $\beta = 111.85°$, and is estimated to contain two molecules of Cwp19 per asymmetric unit.

1. Introduction

Clostridium difficile is a Gram-positive spore-forming primarily nosocomial pathogen that is the aetiological agent in antibiotic-associated diarrhoea and pseudomembranous colitis (Bartlett, 2010). Changes in epidemiology and disease severity, particularly in strains that have emerged over the last ten years, *e.g.* the 027 ribotype, highlight the need to understand more about this worldwide pathogen (Freeman *et al.*, 2010).

The elucidation of structural information for *C. difficile* proteins has understandably been directed towards the main virulence factors, the toxins (Albesa-Jové *et al.*, 2010; Ho *et al.*, 2005; Pruitt *et al.*, 2009, 2010; Sundriyal *et al.*, 2009). Despite adherence and subsequent colonization by *C. difficile* representing key milestones in infection, there are considerable gaps in the understanding of how the surface proteins of *C. difficile* interact with both themselves and the environment to mediate these key steps. To date, there is only one report of high-resolution structural information for a *C. difficile* surface protein: the low-molecular-weight subunit of the S-layer (PDB entry 3cvz; Fagan *et al.*, 2009).

The *C. difficile* S-layer is derived from post-translational cleavage of SlpA into low-molecular-weight and high-molecular-weight subunits (LMW SLP and HMW SLP, respectively). HMW SLP contains three PFam04122 repeats which putatively mediate attachment to the bacterial cell surface (cell-wall-binding domains; CWBDs). A total of 28 other proteins in the *C. difficile* 630 genome have been found to contain these CWBDs at the N-terminus or the C-terminus, with a 'functional domain' at the other terminus (Sebaihia *et al.*, 2006). Recently, Dang *et al.* (2010) identified one such CWBD-containing protein, Cwp19 (CD2767; *C. difficile* 630 genome numbering; Fagan *et al.*, 2011; Sebaihia *et al.*, 2006), during a pull-down assay of ABP-labelled Cwp84. Cwp19 has an N-terminal DUF187 domain (together with three C-terminal CWBDs) which belongs to a glycosyl hydrolase clan of enzymes that possess a TIM barrel (a conserved protein fold consisting of eight α-helices and eight parallel β-strands that alternate along the peptide backbone, as originally identified in the conserved glycolytic enzyme triosephosphate isomerase). Other members include α-amylases and cellulases.

To understand the molecular structure of this protein, the N-terminal domain of Cwp19, lacking the CWBDs, has been expressed, purified and crystallized for structural studies.

Table 1
Statistics for the processing of X-ray data from the rCwp19$_{27–401}$ crystal in various possible space groups using *iMOSFLM*.

Values in parentheses are for the highest resolution shell

Space group	$C222/C222_1$	$C2$	$P2/P2_1$	$P1$
Unit-cell parameters (Å, °)	a = 122.4, b = 181.18, c = 61.3, $\alpha = \beta = \gamma = 90.0$	a = 122.3, b = 180.8, c = 61.2, $\alpha = \gamma = 90.0$, $\beta = 89.98$	a = 109.1, b = 61.2, c = 109.2, $\alpha = \gamma = 90.0$, $\beta = 111.9$	a = 61.23, b = 109.2, c = 109.3, $\alpha = 111.8$, $\beta = 90.1$, $\gamma = 89.9$
Resolution range (Å)	50–2.00 (2.11–2.00)	50–2.00 (2.11–2.00)	50–2.00 (2.11–2.00)	50–2.00 (2.11–2.00)
$R_{\rm merge}$	0.502 (0.656)	0.489 (0.666)	0.135 (0.538)	0.100 (0.439)
$R_{\rm p.i.m.}$	0.173 (0.258)	0.234 (0.354)	0.074 (0.306)	0.082 (0.351)
$\langle I/\sigma(I)\rangle$	3.6 (2.2)	2.8 (1.3)	6.3 (2.6)	5.4 (2.0)
Completeness (%)	98.3 (97.8)	97.9 (95.9)	91.4 (82.9)	84.3 (75.6)
Total No. of reflections	348929 (43885)	372261 (47129)	363675 (45705)	371347 (46962)
No. of unique reflections	45719 (6575)	87734 (12542)	83202 (10948)	150187 (19737)
Multiplicity	7.6 (6.7)	4.2 (3.8)	4.4 (4.2)	2.5 (2.4)
Wilson B factor (Å^2)	23.1	22.1	21.0	21.2
Average mosaicity (°)	1.2	1.1	1.1	1.1

2. Materials and methods

2.1. Cloning

A synthesized gene (GENEART, Germany) corresponding to the N-terminus lacking the predicted signal peptide and CWBDs (residues 27–401) of *cwp19* from *C. difficile* QCD32g-58 was cloned into pET28a using *Nde*I and *Eco*RI. The resulting rCwp19$_{27–401}$ protein had a 21-amino-acid leader sequence including a His$_6$ tag (MGSSHHHHHHSSGLVPRGSHM).

2.2. Expression and purification

The *cwp19* construct was transformed into *Escherichia coli* BL21 (DE3) Star (Invitrogen). A single colony was used to inoculate 50 ml Terrific Broth (TB) medium (Sigma) with 50 µg ml^{-1} kanamycin supplemented with 0.5% glucose and grown overnight at 303 K. The starter culture was then inoculated into 950 ml of the aforementioned supplemented TB medium and grown until the OD reached ~0.6. Cultures were then cooled to 289 K, induced with 1 mM IPTG and grown for a further 16 h before harvesting by centrifugation. Cell pellets were either used directly or frozen at 253 K.

The cell pellet was thawed on ice, resuspended in immobilized metal-affinity chromatography (IMAC) binding/wash buffer (50 mM Tris, 0.5 M NaCl, 20 mM imidazole pH 8.0), sonicated and centrifuged to remove cell debris. IMAC was performed on an ÄKTA design FPLC (GE Healthcare) using a HisTrap HP (GE Healthcare) column equilibrated with binding/wash buffer. Elution was performed using an imidazole gradient (elution buffer: 50 mM Tris, 0.5 M NaCl, 0.5 M imidazole pH 8.0). Early elution peak fractions were dialysed into 50 mM Tris, 150 mM NaCl pH 8.0, 0.2 µm filtered and then concentrated in a Vivaspin-20 10k MWCO spin concentrator to approximately 167 mg ml^{-1} (as measured by the Bradford assay using 1 mg ml^{-1} BSA as the standard). Purity was assessed by SDS–PAGE and anti-His$_6$ Western blot.

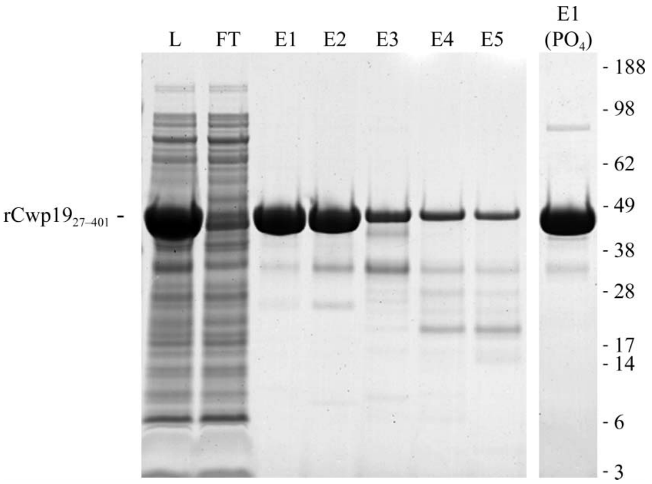

Figure 1
Purification of rCwp19$_{27–401}$. The first seven lanes contain material obtained using Tris-based IMAC buffers. Lane L, *E. coli* lysate. Lane FT, unbound material. Lanes E1–5, eluted fractions from early (E1) and late (E5) in the eluted peak. Lane E1 (PO$_4$), early-eluted fraction from sodium phosphate (monobasic) based IMAC buffers.

2.3. Crystallization

Using a nanodispensing robot (Art Robbins Instruments), sitting-drop vapour-diffusion crystallization trials were set up in 96-well Intelli-Plates (Art Robbins Instruments) and incubated at 289 K. Appropriate amounts of protein solution and reservoir solution were dispensed to give 2:1, 1:1 and 1:2 ratios (using 150 or 300 nl volumes). The following screens were assessed: PACT *premier*, JCSG-*plus*, Structure Screen 1 and 2 HT-96, MemGold and Morpheus (Molecular Dimensions). A large crystal appeared after ~4 months in well D10 of Structure Screen 1 and 2 HT-96 [0.05 M potassium dihydrogen phosphate, 20%(*w*/*v*) PEG 8000] using a 1:1 protein:reservoir ratio.

2.4. X-ray data collection and processing

A total of 250 images were recorded from a single crystal of rCwp19$_{27–401}$ using a Quantum-4 CCD detector (ADSC Systems, California, USA) with an oscillation angle of 1.0° per image, a crystal-to-detector distance of 300 mm and an exposure time of 3 s per image at 100 K (no cryoprotectant was used) on the PX beamline I04 at the Diamond Light Source (Didcot, Oxon, England). The diffraction data were processed using the *iMOSFLM* X-ray data-processing package (Battye *et al.*, 2011) and were scaled using *SCALA* (part of the *CCP*4 program suite; Winn *et al.*, 2011). Data-collection and processing statistics are listed in Table 1. Molecular-replacement trials were attempted using the *PHENIX* suite of crystallography programs (Adams *et al.*, 2010).

3. Results and discussion

3.1. Protein expression and crystallization

Despite the identification of 28 SlpA paralogues containing a Pfam 04122 (cell-wall-binding domain, CWBD), only 11 have been either identified on the cell surface or have had their transcription demonstrated (Calabi *et al.*, 2001; Karjalainen *et al.*, 2001; Wright *et al.*, 2005). The role of CWBD-containing surface proteins in the physiology and pathogenesis of *C. difficile* has therefore only started to be understood and requires further work.

To obtain pure rCwp19 it was necessary to express only the N-terminal functional domain, residues 27–401 (minus the predicted signal peptide, residues 1–26), containing the predicted glycosidase catalytic core. The full-length protein (including the CWBDs but also lacking the signal peptide) exhibited extensive truncation/degradation and purification issues. IMAC purification yielded a pure (>90%) 47 kDa species in one step, particularly early in the elution peak (Fig. 1). rCwp19$_{27–401}$ had a tendency to dimerize when purified or dialysed in phosphate buffers. However, we could concentrate the protein to a final concentration of 167 mg ml^{-1}.

Using an automated high-throughput sitting-drop vapour-diffusion technique, crystals were obtained in condition D10 of Structure Screen 1 and 2 HT-96 [0.05 *M* potassium dihydrogen phosphate, 20%(*w*/*v*) PEG 8000]. The crystal (Fig. 2) grew after approximately four months and diffracted to 2.0 Å resolution (Fig. 3).

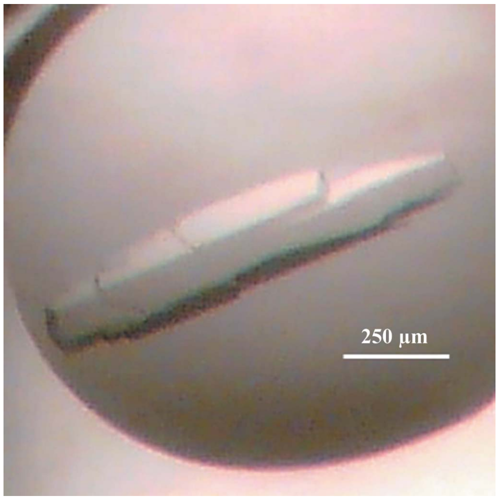

Figure 2
Crystal of rCwp19$_{27–401}$.

3.2. Space-group ambiguity

The X-ray diffraction data for the crystal of rCwp19$_{27–401}$ were analyzed by processing the data in all suggested space groups using the *iMOSFLM* software suite (Battye *et al.*, 2011). The data were processed in centred orthorhombic, centred and primitive monoclinic and primitive triclinic space groups. The final data-processing statistics for all of these possible space groups are given in Table 1. *POINTLESS* (Winn *et al.*, 2011) suggested the primitive monoclinic system as a possible space group for the rCwp19$_{27–401}$ crystal; however, we also analysed the data for the presence of pseudotranslational symmetry (Adams *et al.*, 2010; Winn *et al.*, 2011; Vagin & Teplyakov, 1997; Vaguine *et al.*, 1999) and complete/partial merohedral twinning (Padilla & Yeates, 2003; French & Wilson, 1978; Adams *et al.*, 2010; Winn *et al.*, 2011). These analyses were performed for data processed in centred orthorhombic, primitive monoclinic and primitive triclinic space groups using *TRUNCATE* (Winn *et al.*, 2011; French & Wilson, 1978), *phenix.xtriage* (Adams *et al.*, 2010), the *L*-test (Adams *et al.*, 2010; Padilla & Yeates, 2003) and the *H*-test (Lebedev *et al.*, 2006). Patterson maps were calculated using *MOLREP* (Vagin & Teplyakov, 2010) and *POLARRFN* from the *CCP*4 package (Winn *et al.*, 2011).

3.2.1. Twinning analysis. *TRUNCATE* analysis showed normalized structure amplitudes $\langle E \rangle$ of 0.928 and 0.889 for the centred orthorhombic and primitive monoclinic space groups, respectively. The expected value for an untwinned data set is 0.886 and that for a perfectly twinned data set is 0.94. Thus, *TRUNCATE* indicated the presence of partial twinning in the centred orthorhombic space group with a twin fraction of 0.218. Twinning was not detected by *TRUNCATE* in the primitive monoclinic space group.

The *L*-test analysis (Adams *et al.*, 2010; Padilla & Yeates, 2003) gave multivariate *Z* scores of 20.34 and 4.59 for the centred orthorhombic and primitive monoclinic space groups, respectively (Figs. 4*a* and 4*b*), indicating the presence of perfect twinning in the centred

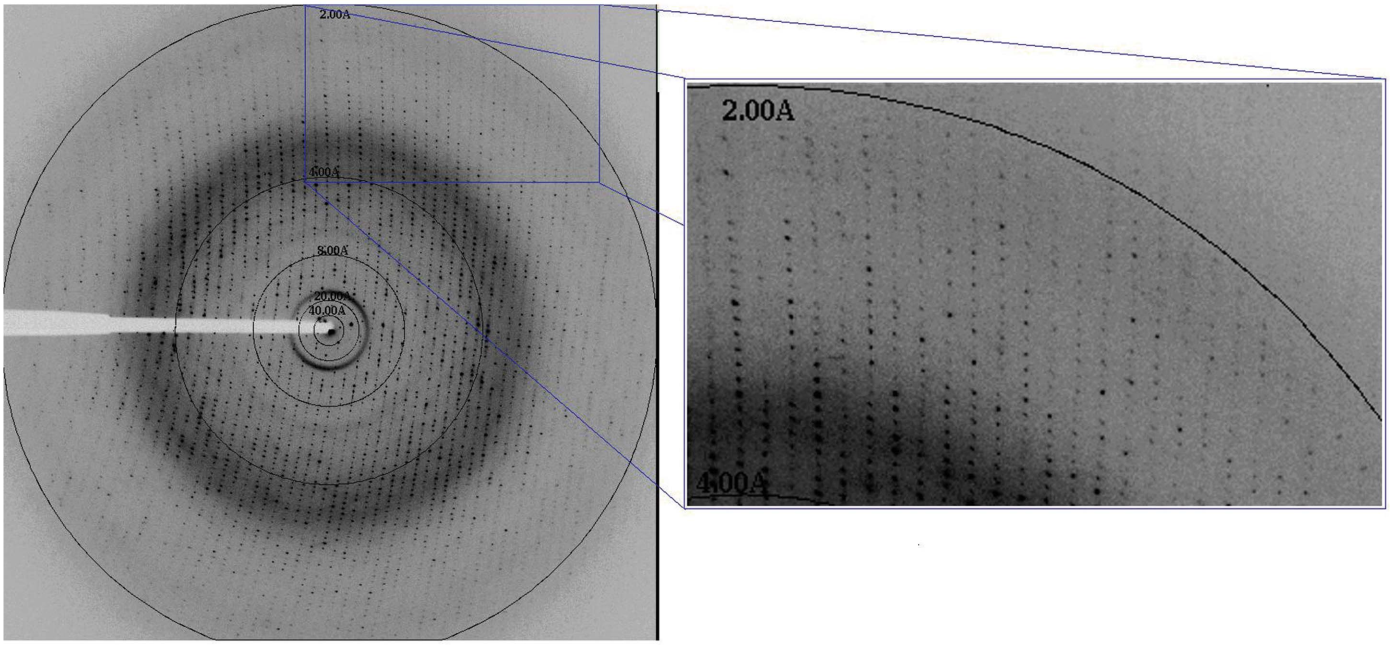

Figure 3
X-ray diffraction image collected from the crystal of rCwp19$_{27–401}$ at Diamond Light Source (Oxon, England).

orthorhombic system. For untwinned data and where pseudosymmetry may be absent, the Z score is expected to be <3.5; this is not the case for the primitive monoclinic space group. The mean $|L|$ values were 0.334 and 0.432 for the centred orthorhombic and primitive monoclinic systems, respectively. For a perfectly twinned case this value should be 0.375 and for an untwinned data set the value should be 0.500. In the present case, the value for the primitive monoclinic space group is closer to that for untwinned data. A similar L-test analysis for the primitive triclinic system resulted in a mean $|L|$ value of 0.442 and a multivariate Z score of 3.593.

The H-test (Lebedev *et al.*, 2006) analysis gave a twin fraction of 0.022 for both the primitive monoclinic and primitive triclinic space groups. In the case of untwinned data the expected mean $|H|$ value should be 0.50; values of 0.482 and 0.499 were found for the primitive monoclinic and primitive triclinic space groups, respectively. The H-test was not performed for the centred orthorhombic system as there are no twin laws available for this space group.

The various twinning tests may appear to have erratic or high twin-fraction results because the data do not scale well in centred space groups (*C*2 or *C*222; Table 1). However, twinning may be absent in the primitive monoclinic space group.

3.2.2. Pseudotranslational symmetry analysis. The presence of noncrystallographic symmetry (NCS) was tested for using *MOLREP* (Vagin & Teplyakov, 2010) and *phenix.xtriage* (Adams *et al.*, 2010). Both indicated the presence of pseudotranslational NCS in the centred orthorhombic and primitive monoclinic space groups. A strong off-origin peak was found in all these space groups. In the primitive monoclinic and primitive triclinic systems the strength of

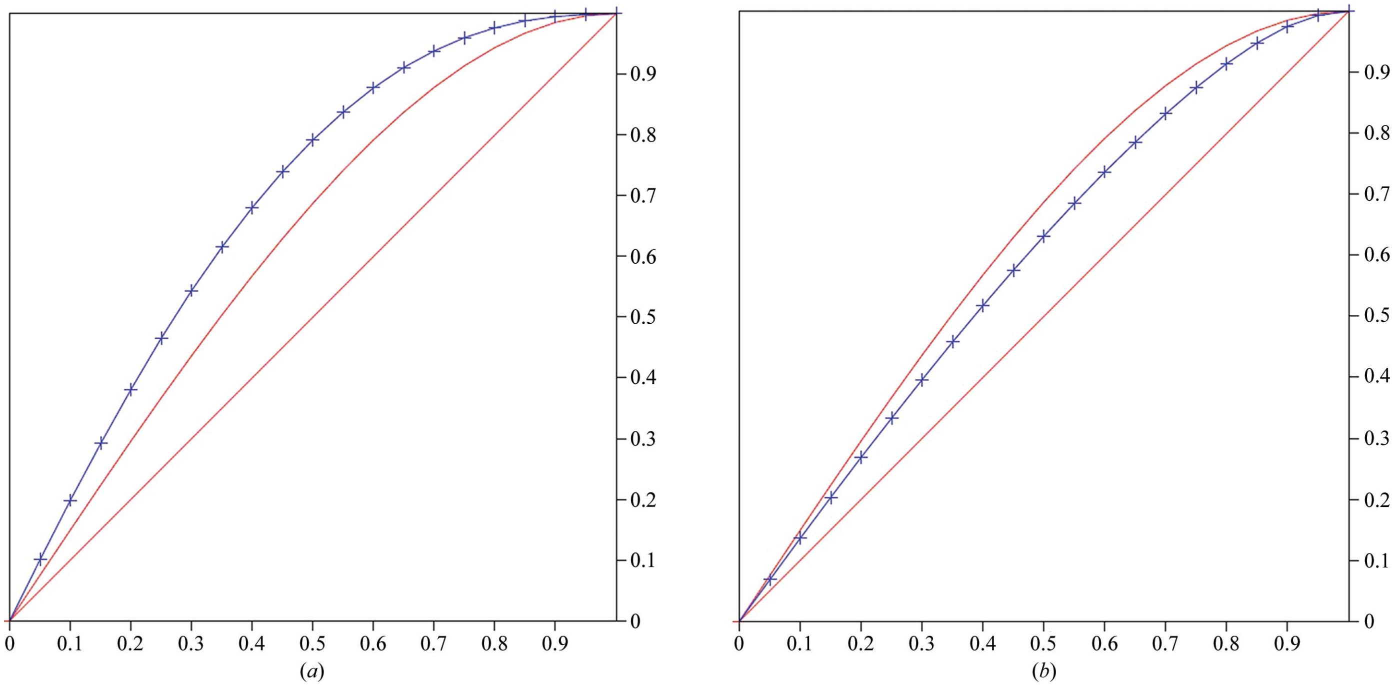

Figure 4
L-test analysis for space groups *C*222/*C*222$_1$ (*a*) and *P*2/*P*2$_1$ (*b*). Curved line, perfect twin; straight line, untwinned; blue line with marks, observed data.

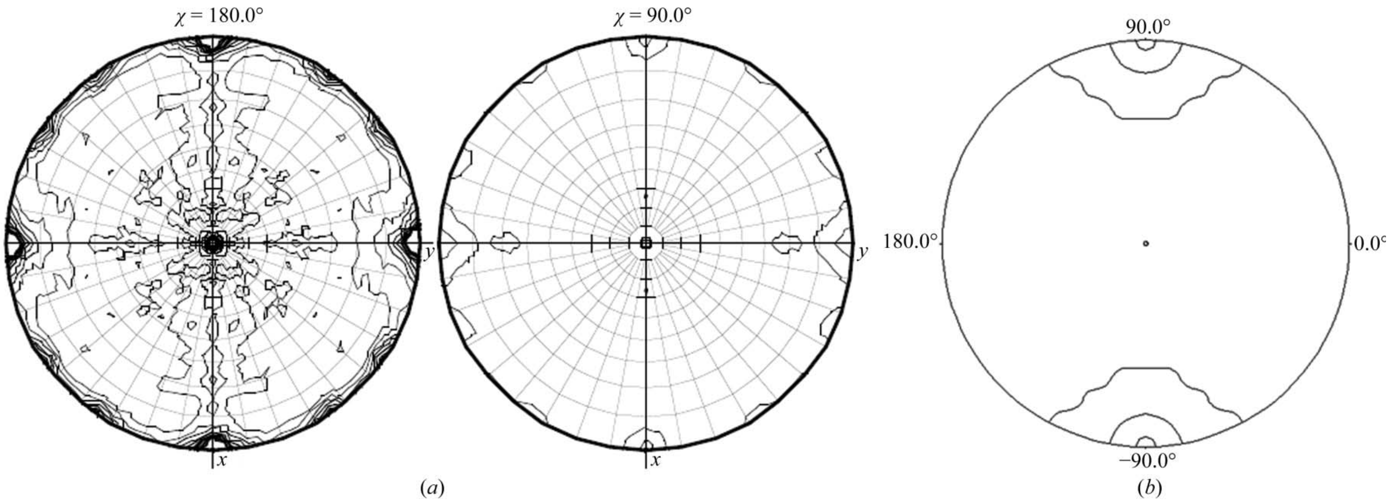

Figure 5
Self-rotation Patterson maps for space group *C*222 as calculated by (*a*) *MOLREP* and (*b*) *POLARRFN* (κ = 90°).

the off-origin peak was 50% of the origin peak, whereas in the centred orthorhombic space group it was only 23%. The corresponding p-values (calculated using *phenix.xtriage*) are 0.00520, 6.8×10^{-5} and 7.2×10^{-5} for the centred orthorhombic, primitive monoclinic and primitive triclinic systems, respectively (a p-value of <0.05 indicates the presence of pseudotranslational NCS). A self-rotation function was also calculated in the centred orthorhombic (Figs. 5*a* and 5*b*), primitive monoclinic (Figs. 6*a* and 6*b*) and primitive triclinic (Fig. 7) space groups using *MOLREP* and *POLARRFN* (Winn *et al.*, 2011).

3.2.3. Data-processing statistics and point-group analysis. The X-ray data-processing statistics indicated that the centred orthorhombic space group had an overall $\langle I/\sigma(I)\rangle$ of 3.6 and an overall merging R of 0.50, compared with the primitive monoclinic space group which had an overall $\langle I/\sigma(I)\rangle$ of 6.3 and an overall merging R of 0.135. The corresponding values for the centred monoclinic space group were 2.8 and 0.489 for the overall $\langle I/\sigma(I)\rangle$ and overall merging R, respectively. For the primitive triclinic system these values were 5.4 and 0.100 for the overall $\langle I/\sigma(I)\rangle$ and overall merging R, respectively. Similarly, the overall $R_{\rm p.i.m.}$ (Evans, 2006; Leslie, 1992) values were also high for the centred orthorhombic and centred monoclinic space groups compared with the primitive monoclinic and primitive triclinic systems (Table 1).

Analysis of systematic absences (Adams *et al.*, 2010) confirmed the presence of a twofold 2_1 screw axis in both the centred orthorhombic and primitive monoclinic space groups. There were three and two violations with $\langle I/\sigma(I)\rangle > 3.0$ for the centred orthorhombic space groups $C222$ and $C222_1$, respectively, whereas for the primitive monoclinic space groups $P2$ and $P2_1$ there were zero and four violations with $\langle I/\sigma(I)\rangle > 3.0$, respectively. However, the likelihoods for the centred orthorhombic and primitive monoclinic space groups are 7 and 1.7, respectively (as calculated using *phenix.xtriage*; Adams *et al.*, 2010).

A point-group test performed by *phenix.xtriage* (Adams *et al.*, 2010) suggested the reprocessing of data that were processed previously in the centred orthorhombic space group, which could have resulted as a consequence of over-merging of pseudo-symmetry and/or twinned data, *i.e.* this is possibly not the correct space group. A similar point-group test was carried out for data processed in the primitive monoclinic space group, which suggested this could be the correct space group, with unit-cell parameters a = 109.1, b = 61.2, c = 109.2 Å, β = 111.9°. A point-group test in the primitive triclinic

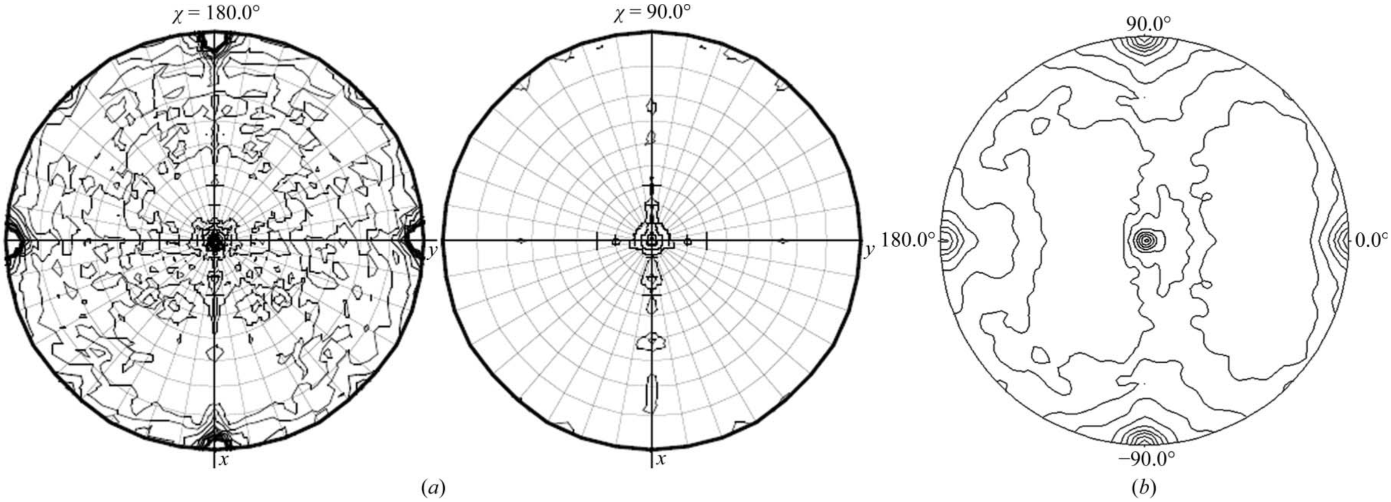

Figure 6
Self-rotation Patterson maps for space group $P2$ as calculated by (*a*) *MOLREP* and (*b*) *POLARRFN* (κ = 180°).

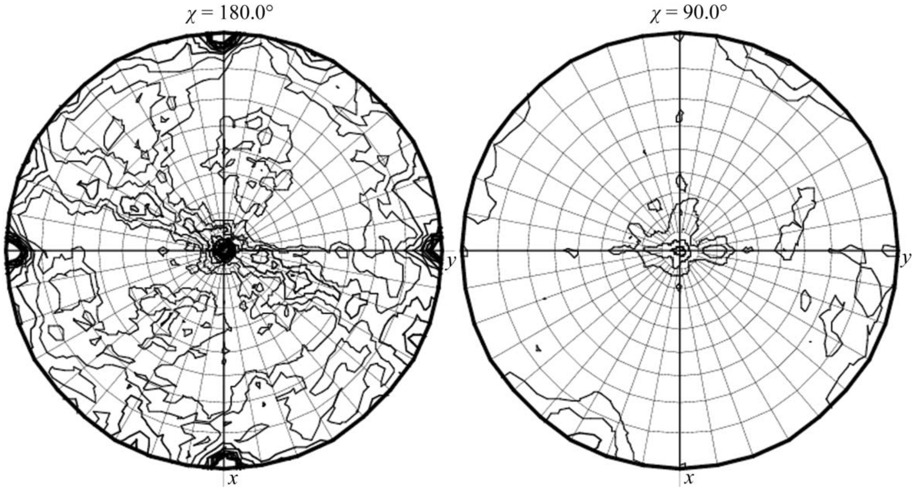

Figure 7
Self-rotation Patterson maps for space group $P1$ as calculated by *MOLREP*.

system also suggested a primitive monoclinic space group with identical unit-cell parameters and a likelihood score of 3.0.

Based on the various analyses performed, the data-processing statistics and suggestions from *POINTLESS* (Winn *et al.*, 2011) and *phenix.xtriage* (Adams *et al.*, 2010), we conclude that the crystal of rCwp19$_{27–401}$ could belong to a primitive monoclinic space group. In addition, *phenix.xtriage* analysis of data processed in the primitive monoclinic space group detected the presence of pseudo-translational noncrystallographic symmetry (which could be the reason for the elevated intensity ratios observed) and twinning could be present. Hence, twin laws are applicable to this crystal symmetry and this could be the reason for the departure of the intensity statistics from normality.

3.3. Low sequence homology

BLASTP (http://blast.ncbi.nlm.nih.gov) analysis revealed that Cwp19$_{27–401}$ has low sequence homology to known protein structures in the PDB; the closest available structure (PDB entries 1eh9 and 1eha; Feese *et al.*, 2000) shares 24% identity (44% similarity) but only over 35% of Cwp19$_{27–401}$. Given the proposed space group, molecular-replacement trials were attempted in space group $P2_1$ using homology models generated by *SWISS-MODEL* (Arnold *et al.*, 2006) [based on PDB entries 2gsj (Cavada *et al.*, 2006) and 3bxw (Meng *et al.*, 2010)] or *Phyre* (Kelley & Sternberg, 2009) [based on PDB entry 1m7x (Abad *et al.*, 2002)], but were unsuccessful presumably owing to low sequence identity (12.7% for 2gsj, 7.8% for 3bxw and 17% for 1m7x). Molecular modelling using the aforementioned servers together with *HHPred* (Söding *et al.*, 2005) and *I-TASSER* (Roy *et al.*, 2010) suggests that Cwp19$_{27–401}$ has homology to proteins with a TIM-barrel structure. We are currently attempting to solve the structure of rCwp19$_{27–401}$ using experimental phasing methods.

This work was funded by the National Institute for Health Research Centre for Health Protection Research at the Health Protection Agency (a postgraduate studentship to JMK). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. We would also like to thank the Wellcome Trust (UK) for the equipment grant (No. 088464), the Royal Society for an Industry Fellowship to KRA and the scientists on PX beamline I04 at the Diamond Light Source (Oxon, England).

References

Abad, M. C., Binderup, K., Rios-Steiner, J., Arni, R. K., Preiss, J. & Geiger, J. H. (2002). *J. Biol. Chem.* **277**, 42164–42170.

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Albesa-Jové, D., Bertrand, T., Carpenter, E. P., Swain, G. V., Lim, J., Zhang, J., Haire, L. F., Vasisht, N., Braun, V., Lange, A., von Eichel-Streiber, C., Svergun, D. I., Fairweather, N. F. & Brown, K. A. (2010). *J. Mol. Biol.* **396**, 1260–1270.

Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. (2006). *Bioinformatics*, **22**, 195–201.

Bartlett, J. G. (2010). *Ann. N. Y. Acad. Sci.* **1213**, 62–69.

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. (2011). *Acta Cryst.* D**67**, 271–281.

Calabi, E., Ward, S., Wren, B., Paxton, T., Panico, M., Morris, H., Dell, A., Dougan, G. & Fairweather, N. (2001). *Mol. Microbiol.* **40**, 1187–1199.

Cavada, B. S. *et al.* (2006). *FEBS J*, **273**, 3962–3974.

Dang, T. H., de la Riva, L., Fagan, R. P., Storck, E. M., Heal, W. P., Janoir, C., Fairweather, N. F. & Tate, E. W. (2010). *ACS Chem. Biol.* **5**, 279–285.

Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.

Fagan, R. P., Albesa-Jové, D., Qazi, O., Svergun, D. I., Brown, K. A. & Fairweather, N. F. (2009). *Mol. Microbiol.* **71**, 1308–1322.

Fagan, R. P., Janoir, C., Collignon, A., Mastrantonio, P., Poxton, I. R. & Fairweather, N. F. (2011). *J. Med. Microbiol.*, doi:10.1099/jmm.0.028472-0.

Feese, M. D., Kato, Y., Tamada, T., Kato, M., Komeda, T., Miura, Y., Hirose, M., Hondo, K., Kobayashi, K. & Kuroki, R. (2000). *J. Mol. Biol.* **301**, 451–464.

Freeman, J., Bauer, M. P., Baines, S. D., Corver, J., Fawley, W. N., Goorhuis, B., Kuijper, E. J. & Wilcox, M. H. (2010). *Clin. Microbiol. Rev.* **23**, 529–549.

French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.

Ho, J. G., Greco, A., Rupnik, M. & Ng, K. K.-S. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 18373–18378.

Karjalainen, T., Waligora-Dupriet, A. J., Cerquetti, M., Spigaglia, P., Maggioni, A., Mauri, P. & Mastrantonio, P. (2001). *Infect. Immun.* **69**, 3442–3446.

Kelley, L. A. & Sternberg, M. J. (2009). *Nature Protoc.* **4**, 363–371.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Acta Cryst.* D**62**, 83–95.

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.

Meng, G., Zhao, Y., Bai, X., Liu, Y., Green, T. J., Luo, M. & Zheng, X. (2010). *J. Biol. Chem.* **285**, 39898–39904.

Padilla, J. E. & Yeates, T. O. (2003). *Acta Cryst.* D**59**, 1124–1130.

Pruitt, R. N., Chagot, B., Cover, M., Chazin, W. J., Spiller, B. & Lacy, D. B. (2009). *J. Biol. Chem.* **284**, 21934–21940.

Pruitt, R. N., Chambers, M. G., Ng, K. K.-S., Ohi, M. D. & Lacy, D. B. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 13467–13472.

Roy, A., Kucukural, A. & Zhang, Y. (2010). *Nature Protoc.* **5**, 725–738.

Sebaihia, M. *et al.* (2006). *Nature Genet.* **38**, 779–786.

Söding, J., Biegert, A. & Lupas, A. N. (2005). *Nucleic Acids Res.* **33**, W244–W248.

Sundriyal, A., Roberts, A. K., Shone, C. C. & Acharya, K. R. (2009). *J. Biol. Chem.* **284**, 28713–28719.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Wright, A., Wait, R., Begum, S., Crossett, B., Nagy, J., Brown, K. & Fairweather, N. (2005). *Proteomics*, **5**, 2443–2452.