Stefano Benini,^{a*} Maria
 Chechik,^b Miguel Ortiz
 Lombardía,^c Sigrun Polier,^d
 Andrew Leech,^e Mikhail B.
 Shevtsov^f and Juan C. Alonso^g

^aLaboratory of Bioorganic Chemistry and
 Crystallography, Faculty of Science and
 Technology, Free University of Bolzano,
 Piazza Università 5, 39100 Bolzano, Italy,

^bYork Structural Biology Laboratory,
 University of York, York YO10 5YW, England,

^cArchitecture et Fonction des Macromolécules
 Biologiques, CNRS, Universités d'Aix-Marseille,
 13288 Marseille CEDEX 9, France, ^dMolecular
 Chaperones Group Genome Damage and
 Stability Centre, University of Sussex, Science
 Park Road, Falmer, Brighton BN1 9RQ,

England, ^eMolecular Interactions Laboratory,
 Department of Biology, University of York,
 York YO10 5YW, England, ^fSchool of Biological
 Sciences, University of Portsmouth,

King Henry Building, King Henry I Street,
 Portsmouth PO1 2DY, England, and
^gBiocnología Microbiana, Centro Nacional de
 Biotecnología, (CNB-CSIC), Darwin 3, Campus
 de Cantoblanco, 28049 Madrid, Spain

Correspondence e-mail: stefano.benini@unibz.it

Received 23 October 2012

Accepted 14 February 2013

PDB Reference: DNA-binding domain of the
 bacteriophage SF6 small terminase subunit,
 2cmp

© 2013 International Union of Crystallography
 All rights reserved

The 1.58 Å resolution structure of the DNA-binding domain of bacteriophage SF6 small terminase provides new hints on DNA binding

DNA packaging in tailed bacteriophages and in evolutionarily related herpesviruses is controlled by a viral-encoded terminase. As in a number of other phages, in the *Bacillus subtilis* bacteriophages SF6 and SPP1 the terminase complex consists of two proteins: G1P and G2P. The crystal structure of the N-terminal DNA-binding domain of the bacteriophage SF6 small terminase subunit G1P is reported. Structural comparison with other DNA-binding proteins allows a general model for the interaction of G1P with the packaging-initiation site to be proposed.

1. Introduction

Tailed bacteriophages and the evolutionarily related herpesviruses package their genomes into preformed empty procapsids using a molecular motor embedded in one of the fivefold-symmetrical vertices of the icosahedral viral shell (Casjens, 2011). There are two general mechanisms of packaging concatemeric viral double-stranded DNA. One mechanism involves the recognition of a specific site on the viral DNA for the initiation and termination of packaging, thus determining the exact length of the viral chromosome (*e.g.* λ phage; Black, 1989; Catalano *et al.*, 1995). In the other mechanism, the initiation event is site-specific while termination arises by a so-called headful cleavage mechanism in which the length of the viral genomic DNA is not precisely determined and the DNA is cut with a redundant end (Black, 1989).

In the *Bacillus subtilis* bacteriophages SF6 and SPP1 the terminal redundancy generated by the imprecise DNA cuts is ~3% of the total length (the genome is 44.0 kb, while the packaged chromosome is 45.9 kb). The headful packaging mechanism in SF6 and SPP1 is sequence-independent and the new terminus generated at the end of the DNA encapsidation serves as a starting point for packaging into the next procapsid. The *pac* site is used only once per approximately every four packaging events, generating a heterogeneous population of terminally redundant and circularly permuted viral chromosomes (Tavares *et al.*, 1992). In the first packaging event the non-encapsidated end of *pacL* is degraded by nucleases, while the rest of the DNA that contains *pacR* is encapsidated (Tavares *et al.*, 1992; Camacho *et al.*, 2003) until the capsid is full.

DNA encapsidation in SF6 and SPP1 is a concerted event involving the formation of a protein complex between G1P (terminase small subunit) and G2P (terminase large subunit) to form the terminase complex binding to the viral DNA concatemer. G1P is involved in DNA binding and recruitment of G2P, which is itself composed of two domains with different activities. The G2P N-terminal domain has ATPase activity, while its C-terminal domain is an endonuclease (Smits *et al.*, 2009). The G1P–G2P–DNA heterocomplex interacts with the portal protein G6P located at the unique vertex of the procapsid.

The SPP1 DNA-packaging motor has been characterized *in vitro* and a model for DNA packaging has been proposed (Camacho *et al.*, 2003). In this model G1P specifically recognizes a region on the phage DNA called the *pac* site; this is followed by cleavage of DNA by G2P and encapsidation. The packaging site contains a *pacL* segment

(100 bp) corresponding to the part of DNA which is not encapsidated in the first insertion cycle, a *pacC* segment (83 bp) containing two **b** boxes (10 bp each) with the cleavage site within one of them and a *pacR* segment (31 bp) located at the end of the DNA which is encapsidated after the DNA is cut (Bravo *et al.*, 1990).

The SPP1 *pac* site was cloned and analysed by Bravo *et al.* (1990). A minimum length of 83 base pairs was found to be sufficient to direct the encapsidation of the viral DNA. Chai and coworkers proposed that SPP1 G1P specifically recognizes the *pac* site by binding to *pacL* and *pacR* but not to *pacC*, which includes the G2P cleavage site (Chai *et al.*, 1995). About 20 subunits of G1P (two decamers) were proposed to bind to the *pac* DNA (Chai *et al.*, 1995). It was also suggested that the *pac* DNA could assume an intrinsically bent conformation owing to the presence of AT-rich segments every 10–11 bp (period) in the DNA (Gual & Alonso, 1998; Crothers *et al.*, 1990). The bending of *pac* DNA upon binding to G1P was demonstrated by DNA-looping experiments (Chai *et al.*, 1995), which supported a model of *pacL* DNA wrapping around G1P. DNA footprinting revealed that one strand of *pacL* contains nine protected sites intercalated by DNAase I-sensitive regions separated by 10 ± 1 bases (one helical turn, assuming 10.5 bp per turn; Chai *et al.*, 1995). These results support a model of a G1P–DNA complex with DNA binding to nine DBDs.

The *pacR* site contains only two protected regions. This may explain the lower affinity of *pacR* towards G1P when compared with *pacL*. Preliminary results of DNA-binding experiments carried out in our group showed that *pacL* binds to the G1P oligomer but fails to bind to the N-terminal DNA-binding domain of G1P (data not shown). This could mean that (i) oligomer formation is necessary for binding to the DNA, (ii) there is cooperativity between the subunits or (iii) binding is enhanced by shape complementarity between the intrinsically bent *pacL* DNA and the G1P oligomer. Replacing parts of the *pac* sequence by nonspecific DNA stretches showed that a single box **a** within the intrinsically bent DNA is sufficient to allow DNA binding (Gual & Alonso, 1998). This result again suggests that the *pac* DNA may specifically interact with the G1P oligomer because of strong shape complementarity rather than the recognition of a specific nucleotide sequence.

The natural sequence-directed arc existing in the *pacL* site was proposed to be the element that triggers the formation of the G1P–DNA complex based on evidence from distamycin-inhibition experiments (Chai & Alonso, 1996). Distamycin, which has been shown to remove DNA curvature (Wu & Crothers, 1984), inhibits the G1P–*pacL* interaction (Chai & Alonso, 1996). The addition of divalent cations reverses distamycin inhibition by increasing the DNA bend, thus favouring interaction of *pacL* with G1P (Chai & Alonso, 1996). Surface plasmon resonance experiments carried out on full-length G1P showed that the strongest interaction between the *pac* DNA and G1P was with a 428 bp fragment of *pac1* which included *pacL*, *pacC* and *pacR* (Büttner *et al.*, 2012).

SF6 and SPP1 G1P share a sequence identity of 84%. One of the highly conserved segments is in the N-terminus: residues 11–49. It contains the helix–turn–helix motif (residues 23–44) that recognizes and binds the *pacR* and *pacL* subsites (Chai *et al.*, 1994). Other conserved segments situated in the C-terminal half of the protein have been proposed to be involved in G1P homo-oligomer formation (Gual & Alonso, 1998; Chai *et al.*, 1994). This was confirmed by recent findings in which the full-length protein (145 residues) formed nine-subunit assemblies, whereas a truncated form of the protein (residues 1–120) lacking the C-terminal part formed mixtures of nine- and ten-subunit assemblies (Büttner *et al.*, 2012).

To date the crystal structures of a few terminases have been determined, but none of them are in a complex with DNA (de Beer

et al., 2002; Zhao *et al.*, 2010; Roy & Cingolani, 2012). The recent determination of the 4 Å resolution crystal structure of full-length SF6 G1P has allowed the proposal of a DNA-binding model characterized by the DNA wrapping around the outside of a nine-subunit G1P oligomer as in the nucleosome (Büttner *et al.*, 2012). Structural information on the interaction between the G1P DNA-binding domain and DNA is not available to date.

Here, we report the 1.58 Å resolution crystal structure of the N-terminal DNA-binding domain of bacteriophage SF6 G1P (G1PNT). Based on structural similarity to the human telomeric protein TRF1 (Court *et al.*, 2005; PDB entry 1w0t; TRF1 in complex with DNA), we propose a model for the interaction of G1P with the packaging-initiation site.

2. Materials and methods

2.1. Gene cloning

The gene sequence corresponding to full-length G1P was cloned from the plasmid pBT411 (provided by J. C. Alonso) by PCR and inserted into a pET28a vector (Novagen) using *NdeI* and *HindIII* restriction sites and forward 5'-AGGAGGGGATCATATGAAA-GAAC-3' and reverse 5'-AAATGCGGCGTAAGCTTT-3' primers. PCR was performed using Vent_R DNA polymerase (New England BioLabs). The ligation product (pYM04) was used to transform competent *Escherichia coli* BL21 (DH5 α) for plasmid production. The N-terminal DNA-binding domain of SF6 G1P from residues 1 to 60 (G1PNT) was subcloned from plasmid pYM04 by a similar procedure with the DNA amplified using forward 5'-GACGC-TCGTTTGAAGTAAATAAACGAGAAGAAAATCCTC-3' and reverse 5'-GGAGGATTTTCTTCTCGTTTATTTACTTCAAACG-AGCGTC-3' primers. The plasmid resulting from ligation was named pYM11. Both clones (pYM04 and pYM11) were sequenced to confirm the absence of mutations.

2.2. Protein expression

G1PNT plasmid pYM11 was transformed into competent *E. coli* BL21 Rosetta pLysS cells for protein expression. 10 ml LB supplemented with 34 $\mu\text{g ml}^{-1}$ chloramphenicol and 30 $\mu\text{g ml}^{-1}$ kanamycin was inoculated and grown overnight at 303 K. The overnight culture was used to inoculate 1 l TB medium containing the same antibiotics. The cells were grown at 310 K until the OD₆₀₀ reached 0.6. The temperature was decreased to 289 K and the culture was induced with 1 mM IPTG. The cells were harvested by centrifugation 16 h after induction. For overexpression of SeMet-containing protein, the plasmid pYM11 was transformed into competent *E. coli* strain B834 (DE3). For protein production, the cells were grown overnight at 303 K in 10 ml LB supplemented with 30 $\mu\text{g ml}^{-1}$ kanamycin, harvested by centrifugation and washed twice with M9 medium supplemented with an amino-acid mixture containing SeMet, vitamins, FeSO₄, MgSO₄ and glucose and resuspended in 1 l of the same medium containing 30 $\mu\text{g ml}^{-1}$ kanamycin. The cells were further grown as for native G1PNT.

2.3. Protein purification

G1PNT was purified by resuspending 5 g bacterial pellet in 50 ml ice-cold buffer consisting of 20 mM sodium phosphate pH 7.5, 0.5 M NaCl, 5% glycerol, 10 mM imidazole (buffer A). Protease inhibitors (0.5 $\mu\text{g ml}^{-1}$ leupeptin, 0.7 $\mu\text{g ml}^{-1}$ pepstatin, 1 mM AEBSF) and 1 mg ml⁻¹ hen egg-white lysozyme (Sigma Chemical Co.) were added to the cell suspension and the mixture was kept on ice for 30 min. The

Table 1

Data-collection statistics.

Values in parentheses are for the highest resolution bin.

	Peak	Remote	Native
Wavelength (Å)	0.9785	0.9185	0.9765
Resolution range (Å)	25.0–2.50 (2.59–2.50)	25.0–2.49 (2.59–2.49)	23.4–1.58 (1.64–1.58)
No. of images	100	100	100
Crystal-to-detector distance (mm)	265	285	155
$\Delta\phi$ per image (°)	1	1	0.5
Raw measurements	25672	26311	81718
Unique reflections	2310	2326	8787
Multiplicity	11.1 (11.0)	11.3 (11.2)	9.3 (5.9)
Completeness (%)	100 (100)	100 (100)	100 (99.9)
$R_{\text{merge}}^{\dagger}$	0.072	0.051	0.036
>3 σ (%)	95.4	95.4	83.2
>2 σ in high-resolution bin (%)	86.7	87.1	57.1
$\langle I/\sigma(I) \rangle$	23.9 (18.5)	27.2 (17.7)	38.7 (3.2)
Space group	$P2_13$	$P2_13$	$P2_13$
Unit-cell parameters (Å)	$a = b = c = 57.2$	$a = b = c = 57.3$	$a = b = c = 57.3$
Matthews coefficient (Å ³ Da ⁻¹)	2.3	2.3	2.3
Solvent content (%)	45	45	45

$\dagger R_{\text{merge}} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is an individual intensity measurement and $\langle I(hkl) \rangle$ is the average intensity for this reflection.

cell lysate was sonicated using a Soniprep 150 MSE (with a 19 mm probe tuned to 23 kHz) at half power to prevent foaming with ten pulses of 60 s each and with 30 s intervals between pulses. The supernatant was loaded onto a 5 ml Ni-charged His-Trap chelating column (Amersham Bioscience) previously equilibrated with buffer A. The protein was eluted with a 10–500 mM imidazole gradient in 100 ml buffer A. G1PNT eluted at 120 mM imidazole. To remove the hexahistidine tag, the fractions were pooled, concentrated by ultrafiltration (Vivascience; MWCO 5 kDa) and buffer-exchanged into 20 mM Tris pH 8.5, 150 mM NaCl, 2.5 mM CaCl₂. Thrombin (Roche) was added to the protein solution (3 mg ml⁻¹) at a ratio of one unit of thrombin per 4 mg protein and incubated for 18 h at room temperature on a rotary shaker. The efficiency of thrombin cleavage was checked by SDS–PAGE. After increasing the NaCl concentration to 250 mM, the supernatant from the previous step was loaded onto a 5 ml His-Trap column equilibrated with buffer B (20 mM Tris–HCl pH 8.5, 250 mM NaCl). The column was washed with buffer B and a 0–500 mM imidazole gradient in 100 ml was applied. Cleaved G1PNT was eluted during the wash. The fractions were pooled, concentrated by ultrafiltration in Vivaspin concentrators (MWCO 5 kDa) and buffer-exchanged into 50 mM HEPES pH 7.5, 50 mM NaCl (buffer C). The protein was then loaded onto a 1 ml HR 5/5 Mono S column (Amersham Bioscience) equilibrated with buffer C and eluted with a gradient from 50 mM to 1 M NaCl in 20 ml. G1PNT eluted in three peaks corresponding to 90, 140 and 190 mM NaCl. The protein from each peak was pooled, buffer-exchanged into 50 mM HEPES pH 7.5, 10 mM NaCl and concentrated separately to 15–20 mg ml⁻¹. The same purification procedure was used for the SeMet derivative of G1PNT.

2.4. Protein crystallization

Crystallization screening of G1PNT was carried out using the protein in the third peak of the Mono S purification at a concentration of 17 mg ml⁻¹ in 50 mM HEPES pH 7.5, 10 mM NaCl. Crystallization screens were set up at 292 K using a Mosquito Robot (TTP LabTech) and 96-well Greiner sitting-drop plates with 150 nl protein solution mixed with an equal amount of precipitant. The screens used were Index (Hampton Research), Clear Strategy Screens I and II (Brzozowski & Walton, 2001; Molecular Dimensions) at pH 6.5 and

Table 2

Final model and refinement statistics.

No. of residues	56
Protein atoms	435
Water molecules	52
Protein–water hydrogen bonds	44
Inter-residue hydrogen bonds	146
Water–water hydrogen bonds	38
Buried residues	8
Temperature factor for protein atoms (Å ²)	45.0
Temperature factor for solvent (Å ²)	55.9
R.m.s.d., bond lengths (Å)	0.014
R.m.s.d., bond angles (°)	1.5
Ramachandran plot	
Residues in most favoured region (%)	93.8
Residues in additionally allowed region (%)	6.2
$R/R_{\text{free}}^{\dagger}$ (%)	22.5/27.5
PDB code	2cmp

$\dagger R$ and $R_{\text{free}} = \sum_{hkl} |F_{\text{obs}}| - |F_{\text{calc}}| / \sum_{hkl} |F_{\text{obs}}|$; a subset of reflections (5%) was randomly chosen as a test set for calculation of R_{free} .

pH 8.5, PACT (Molecular Dimensions) and PEG/Ion (Hampton Research). The best hits were found in the Index screen in conditions containing 2.1 M DL-malic acid pH 7.0 or 2.4 M sodium malonate pH 7.0. Diffraction-quality crystals were grown in 24-well Falcon plates at 292 K by mixing 1 μ l protein solution with 1 μ l precipitant solution containing 2.2–2.6 M sodium malonate pH 7.0. Crystals of the SeMet-derivative protein were obtained using the same conditions.

2.5. Data collection and structure solution

Native and SeMet-derivative crystals of G1PNT with dimensions of ~ 0.25 mm³ were flash-cooled in liquid nitrogen directly from the crystallization drop, exploiting the cryoprotection offered by the high concentration of sodium malonate. X-ray data were collected at 100 K on the BM14 beamline at the ESRF using a MAR MX-225 CCD detector and were processed using *DENZO* and *SCALEPACK* (Otwinowski *et al.*, 2003; Otwinowski & Minor, 1997). Data-collection statistics are reported in Table 1. The structure was solved by MAD using data collected at the peak and remote wavelengths. Initial phasing was performed using *HKL2MAP* (Pape & Schneider, 2004), the graphical user interface of *SHELX* (Sheldrick, 2008). The correct solution had correlation coefficients CC_{all} and CC_{weak} of 59.1 and 46.9, respectively. The solvent content for solvent flattening was set to 50% and after 20 cycles of refinement the correct solution had a pseudo-free CC of 57.0% for the original hand, while the pseudo-free CC of the inverted-hand solution was 17.4%. The contrasts for the original and inverted solutions were 0.62 and 0.14, respectively, with connectivities of 0.91 and 0.80, respectively. The initial phases were extended to 1.58 Å resolution with *DM* (Cowtan, 2010) using the native data set and keeping 5% of the reflections as an R_{free} set. *ARP/wARP* (Morris *et al.*, 2004) was used to automatically build a model that was $\sim 93\%$ complete. The missing residues located at the N-terminus were added manually using *Coot* (Emsley *et al.*, 2010). After initial cycles of refinement with *REFMAC* (Murshudov *et al.*, 2011), water molecules were added using *ARP/wARP* followed by TLS refinement (Winn *et al.*, 2003). The model was refined to a final R of 22.5% and R_{free} of 27.5% (Table 2). The quality of the model was assessed with *PROCHECK* (Laskowski *et al.*, 1993). Data analysis by *SFHECK* (Vaguine *et al.*, 1999) showed that the crystal was partially twinned with a twinning fraction of 10%, which explained the relatively high R factors. The structure of G1PNT (PDB entry 2cmp) was used by Büttner and coworkers as a model to fit the electron density of the low-resolution (4 Å) full-length G1P structure (Büttner *et al.*, 2012).

Figure 1

$2F_{\text{obs}} - F_{\text{calc}}$ electron-density map of the conserved turn between helix 2 and helix 3 contoured at 1σ . This figure was prepared with *PyMOL* (DeLano, 2002).

2.6. Sequence analysis

A *BLASTp* (McGinnis & Madden, 2004) search of full-length G1P (CAK29441; gi|10416207) against the nonredundant protein databases was carried out using the online service at the National Center for Biotechnology Information (NCBI). Sequences sharing up to 47% sequence identity were selected for multiple alignment. The output from the multiple alignment carried out with *COBALT* (Papadopoulos & Agarwala, 2007) was used as input, together with the coordinate file, to create a structure–sequence alignment with *ESPrpt* (Gouet *et al.*, 1999).

3. Results

3.1. G1PNT crystal structure analysis

The four N-terminal residues of G1P could not be modelled in the electron-density map because of disorder, while the C-terminal end

Figure 2

Cartoon representation of G1PNT. This figure was prepared with *PyMOL* (DeLano, 2002).

of the protein is ordered with clear electron density corresponding to the main-chain carboxyl of the last residue. The stretch containing the first five modelled residues (Lys5–Lys9) shows two major alternate conformations deviating at the main chain. Fig. 1 shows the electron-density map of the loop between helices 2 and 3.

The G1PNT model consists of 56 residues (Lys5–Lys60) folded into four α -helices connected by short loops (Fig. 2). In the structure, a loop (Lys5–Ser7) is followed by the first α -helix (Pro8–Asn20), a short turn I' (Asp21–Asn23), a second α -helix (Ala24–Ala31), a loop (Gly32–Ser37) and a third α -helix (Ala38–Gln47) separated from the last α -helix (Pro49–Leu59) by Lys48.

The first three helices of G1P form the typical helix–turn–helix DNA-binding motif, which is followed by a fourth helix. The fourth helix acts as a linker between the DNA-binding domain and the oligomerization domain (Büttner *et al.*, 2012). The same feature is found in the DBD of *Shigella* bacteriophage Sf6 small terminase, in which the fourth helix following the HTH links the DBD to the oligomerization domain (Zhao *et al.*, 2010). Likewise, in bacteriophage λ the fourth helix following the DBD domain (α D) has also been proposed to link the DBD to the oligomerization domain (de Beer *et al.*, 2002; Bain *et al.*, 2001). A comparison between the three DBDs is reported as supporting information in Büttner *et al.* (2012). Sequence alignment between SF6 terminase and SPP1-like terminases sharing a pairwise sequence identity of at least 47% and a minimum sequence coverage of 85% showed a high conservation among them, supporting the hypothesis that the G1PNT fold and DNA-binding mechanism could be conserved in all SPP1-like bacteriophages (Fig. 3).

To find a structural model of DNA binding to G1P, an *SSM* (Krissinel & Henrick, 2004) search against all structures in the PDB using the first 48 residues of G1PNT (a model including only the first three helices of the DBD and excluding the linker helix) was carried out. The best match was with the human telomeric protein TRF1 in complex with telomeric DNA (Court *et al.*, 2005; PDB entry 1w0t), with a *Q* score of 0.56 and an r.m.s.d. of 1.85 Å. Analysis of TRF1 DBD reveals only small differences in secondary structure compared with G1PNT; it therefore represents an ideal model to propose a reliable DNA-binding mode for the SF6 small terminase DBD.

4. Discussion

The structure of G1PNT shows a HTH motif from Ala24 to Gln47, in agreement with previous sequence-based predictions (Gual &

Figure 3 Sequence alignment of selected viral small terminases sharing a pairwise sequence identity of at least 47% as shown in Table 3. This figure was prepared with *ESPrpt* (Gouet *et al.*, 1999); red boxes with white characters indicate strict identity, while red characters indicates similarity in a group.

Table 3 Results of the *BLASTp* (McGinnis & Madden, 2004) search using the full-length G1P protein sequence (CAK29441; gi|10416207) against nonredundant protein databases, showing only selected sequences of viral small terminases sharing a pairwise sequence identity of at least 47% and a minimum sequence coverage of 85%.

Protein ID	Total score	Query cover (%)	Maximum identity (%)
gi 104162071 emb CAK29441.1	291	100	100
gi 104162069 emb CAK29440.1	245	97	84
gi 56964604 ref YP_176335.1	124	91	49
gi 420162678 ref ZP_14669433.1	121	93	47
gi 357050646 ref ZP_09111844.1	120	97	47
gi 148379680 ref YP_001254221.1	119	93	51
gi 194466765 ref ZP_03072752.1	112	85	50
gi 66395762 ref YP_240073.1	109	93	48

Alonso, 1998; Chai & Alonso, 1996). Sequence alignment suggests that this fold is common to a number of viral terminase DBDs sharing sequence similarity with the SF6 G1P (Fig. 3).

4.1. Model of G1PNT–DNA interaction

In the crystal structure of TRF1, which is the closest available structural homologue of G1P, helix 3 is inserted into the major groove of the telomeric DNA. The interactions of helix 3 with the DNA are mediated by contacts with residues located in helix 2. The N-terminal arm inserted into the minor groove further enhances the DNA-binding specificity and affinity (Court *et al.*, 2005). Using this structure as a template, it is possible to predict DNA-binding residues in G1P and propose a model for the complex of G1PNT with the *pac* DNA, as the overall similarity of the DNA-binding domain suggests a similar mode of interaction (Fig. 4). In G1PNT the four missing residues at the N-terminus (MKEP) could easily be modelled onto the first four residues of the TRF1 DBD (chain A of PDB entry 1w0t) N-terminal arm (KRQA) in such a way as to mimic a similar DNA-binding pattern. In TRF1 DBD there are 11 residues involved in direct protein–DNA hydrogen-bond interactions. G1PNT has ten residues in similar positions that could potentially form contacts with

the DNA. It is likely that the TRF1 DBD binds DNA more strongly than G1PNT, as TRF1 functions as a dimer and therefore requires a tighter binding per subunit than G1P, which is an oligomer of nine subunits (Büttner *et al.*, 2012).

Figure 4 Cartoon representation of the *SSM* superposition of G1PNT with hTRF1 DBD in complex with DNA. G1PNT is coloured magenta, TRF1 DBD is coloured green and the DNA surface is coloured by atom type. Helix $\alpha 1$ is parallel to the DNA, while helix $\alpha 3$ is located in the DNA major groove. This figure was prepared with *PyMOL* (DeLano, 2002).

The model of DNA binding to G1PNT proposed here further validates the model of the G1P–DNA complex proposed by Büttner and coworkers based on the low-resolution structure of the full-length protein (Büttner *et al.*, 2012). In that structure only two of the three DBDs in the asymmetric unit are visible in the electron-density map, suggesting freedom of movement of the DBDs thanks to the connecting linker. It is possible that the flexibility between the DBDs and the oligomerization domain facilitates positional adjustment of individual DBDs during their interaction with the DNA. Thus, SF6 and SPP1 *pacL* DNA could coil around G1P like the nucleosome DNA coils around a histone octamer (Luger *et al.*, 1997). This model would possibly allow interactions between G1P and the portal protein G6P in the form of multisubunit rings stacked upon each other. The central hole could be used for the insertion of the DNA processed by G2P into the procapsid *via* interaction between the portal protein G6P and the terminase to generate a powerful molecular motor.

This work was supported by a Wellcome Trust fellowship. We thank David Waterman for critically reading the manuscript. The authors thank Dr Martin Walsh for the precious help provided during MAD data collection at the BM14 ESRF beamline.

References

- Bain, D. L., Berton, N., Ortega, M., Baran, J., Yang, Q. & Catalano, C. E. (2001). *J. Biol. Chem.* **276**, 20175–20181.
- Beer, T. de, Fang, J., Ortega, M., Yang, Q., Maes, L., Duffy, C., Berton, N., Sippy, J., Overduin, M., Feiss, M. & Catalano, C. E. (2002). *Mol. Cell*, **9**, 981–991.
- Black, L. W. (1989). *Annu. Rev. Microbiol.* **43**, 267–292.
- Bravo, A., Alonso, J. C. & Trautner, T. A. (1990). *Nucleic Acids Res.* **18**, 2881–2886.
- Brzozowski, A. M. & Walton, J. (2001). *J. Appl. Cryst.* **34**, 97–101.
- Büttner, C. R., Chechik, M., Ortiz-Lombardía, M., Smits, C., Ebong, I. O., Chechik, V., Jeschke, G., Dykeman, E., Benini, S., Robinson, C. V., Alonso, J. C. & Antson, A. A. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 811–816.
- Camacho, A. G., Gual, A., Lurz, R., Tavares, P. & Alonso, J. C. (2003). *J. Biol. Chem.* **278**, 23251–23259.
- Casjens, S. R. (2011). *Nature Rev. Microbiol.* **9**, 647–657.
- Catalano, C. E., Cue, D. & Feiss, M. (1995). *Mol. Microbiol.* **16**, 1075–1086.
- Chai, S. & Alonso, J. C. (1996). *Nucleic Acids Res.* **24**, 282–288.
- Chai, S., Kruff, V. & Alonso, J. C. (1994). *Virology*, **202**, 930–939.
- Chai, S., Lurz, R. & Alonso, J. C. (1995). *J. Mol. Biol.* **252**, 386–398.
- Court, R., Chapman, L., Fairall, L. & Rhodes, D. (2005). *EMBO Rep.* **6**, 39–45.
- Cowan, K. (2010). *Acta Cryst.* **D66**, 470–478.
- Crothers, D. M., Haran, T. E. & Nadeau, J. G. (1990). *J. Biol. Chem.* **265**, 7093–7096.
- DeLano, W. L. (2002). *PyMOL*. <http://www.pymol.org>.
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* **D66**, 486–501.
- Gouet, P., Courcelle, E., Stuart, D. I. & Métoz, F. (1999). *Bioinformatics*, **15**, 305–308.
- Gual, A. & Alonso, J. C. (1998). *Virology*, **242**, 279–287.
- Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* **D60**, 2256–2268.
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
- Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). *Nature (London)*, **389**, 251–260.
- McGinnis, S. & Madden, T. L. (2004). *Nucleic Acids Res.* **32**, W20–W25.
- Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vornrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* **D67**, 355–367.
- Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Acta Cryst.* **A59**, 228–234.
- Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
- Papadopoulos, J. S. & Agarwala, R. (2007). *Bioinformatics*, **23**, 1073–1079.
- Pape, T. & Schneider, T. R. (2004). *J. Appl. Cryst.* **37**, 843–844.
- Roy, A. & Cingolani, G. (2012). *J. Biol. Chem.* **287**, 28196–28205.
- Sheldrick, G. M. (2008). *Acta Cryst.* **A64**, 112–122.
- Smits, C., Chechik, M., Kovalevskiy, O. V., Shevtsov, M. B., Foster, A. W., Alonso, J. C. & Antson, A. A. (2009). *EMBO Rep.* **10**, 592–598.
- Tavares, P., Santos, M. A., Lurz, R., Morelli, G., de Lencastre, H. & Trautner, T. A. (1992). *J. Mol. Biol.* **225**, 81–92.
- Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* **D55**, 191–205.
- Winn, M. D., Murshudov, G. N. & Papiz, M. Z. (2003). *Methods Enzymol.* **374**, 300–321.
- Wu, H. M. & Crothers, D. M. (1984). *Nature (London)*, **308**, 509–513.
- Zhao, H., Finch, C. J., Sequeira, R. D., Johnson, B. A., Johnson, J. E., Casjens, S. R. & Tang, L. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 1971–1976.