from the original NColE7 sequence (starting with K446 according to the original numbering of ColE7 protein). It is known that the native ColE7 gene is toxic for the cells, owing to the unwanted minor expression level of the protein during the cloning process [55]. Thus, the success of the cloning and expression of a ColE7 mutant indicates that the given protein is not toxic for the cells. According to the PCR followed by agarose gel electrophoresis, the genes were successfully inserted into the vector and cloned in either DH10B or Mach1 cells. The transformed BL21 (DE3) cells based on the change in OD_{600} after the induction of the protein expression were grown in a manner similar to that for the cells expressing GST itself, unlike those of the toxic variant of NColE7, where the cells started to die shortly after induction (Fig. 1b). The SDS-PAGE of the expressed proteins showed intense bands around the expected molecular mass (Fig. 1c). However, the DNA sequencing and the mass spectra of the proteins showed that although the GST-ΔN25-NColE7 sequence was correct, instead of GST-ΔN4-NColE7 we had expressed a new mutant, GST-ΔN4-NColE7-C*, with a modified C-terminus—but containing all the amino acids necessary for binding of the Zn^{2+} ion. This strongly suggests that GST-ΔN4-NColE7 was cytotoxic. The sequences of the proteins after the GST cleavage are depicted in Fig. 1a.

To check the effect of the C-terminal modification in GST-ΔN4-NColE7-C*, we have also shown that the ΔN4-NColE7 mutant expressed from the pET21a plasmid without an N-terminal tag but having the correct sequence at the C-terminus is non-toxic for the cells [43]. These results together proved that the N-terminal basic amino acids are necessary for the cell-killing activity of the enzyme if it is overexpressed in bacterial cells. Looking at the available crystal structures of NColE7 (Table 1), we see that these amino acid residues with special emphasis on residue R447 were observed in only a few of them. In those few including this amino acid, however, the R447 side chain is situated close to the Zn^{2+} ion in the active centre [5, 6, 10]. The two positively charged residues are bridged by a phosphate ion (Fig. 1d), which is most probably replaced by the scissile phosphodiester group of the DNA in the catalytically active complex [16]. In the NColE7–DNA crystals, R447 is mostly missing from the solved structure [16, 29, 32], but it is close to the phosphate backbone in the structure of a metal ion deficient mutant [57].

In view of the possible allosteric control, it is important to know what the function of the N-terminal residues with positively charged side chains (one arginine and two lysines) is and the role of the whole N-terminal chain—a component without autonomous secondary structure—regarding the activity of the NColE7 protein. We attempted to get closer to the solution of this problem by means of the investigation of the Zn^{2+}- and DNA-binding abilities of the expressed mutant proteins.

Table 1 Crystal structures of the nuclease domain of colicin E7 (NColE7) containing amino acid sequences of different lengths

PDB ID	Mutation	Complex	Sequence in PDB file[a]	Reason for inactivity
1M08 [6]	K446M	Protein–Zn–PO_4[b]	446 MRNK-HRGK 576	–
1MZ8 [5]	–	Protein–Zn–PO_4–Im7	447 RNKP-IDIH 573	–
1PT3 [16]	–	Protein–8 bp DNA	449 KPGK-HRGK 576	No metal ion
1ZNS [32]	K443M/H545E	Protein–Zn–12 bp DNA	450 PGKA-DIHR 574	Mutation
1ZNV [32]	K443M/H545E	Protein–Ni–PO_4–Im7	450 PGKA-HRGK 576	–
7CEI [10]	–	Protein–Zn–Im7	447 RNKP-IDIH 573	–
2IVH [29]	H545Q	Protein–Zn–18 bp DNA	449 KPGK-IDIH 573	Mutation
2JAZ [28]	N560D	Protein–Zn–PO_4–Im7	450 PGKA-HRGK 576	–
2JB0 [28]	H573A	Protein–Zn–Im7	449 KPGK-HIDI 572	–
2JBG [28]	N560A	Protein–Zn–SO_4–Im7	448 NKPG-HRGK 576	–
3GJN [56]	H545A	Protein–Zn–Im[c]	450 PGKA-HRGK 576	–
3GKL [56]	H545A	Protein–Zn–Im[c]	450 PGKA-HRGK 576	–
3FBD [57]	D493Q	Protein–18 bp DNA	445 SKRN-HRGK 576	No metal ion

PDB Protein Data Bank

[a] All the proteins were expressed in the presence of the immunity protein. A general sequence of NColE7 was MLDKES+446–576, with the exception of the one with PDB ID 7CEI, where an N-terminal hexahistidine tag in a form of MRGSHHHHHHGSES was attached to the 446–576 sequence

[b] Charges are omitted for simplicity

[c] Mutant immunity proteins were applied in these experiments. The expression of NColE7 was not described in detail in the original article [56]

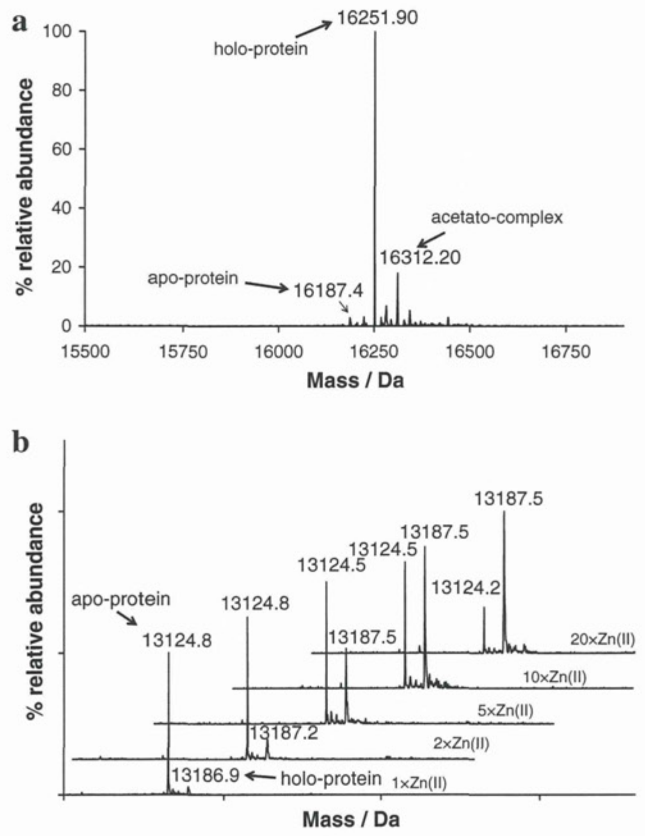

Fig. 2 **a** Mass spectrum of the ΔN4-NColE7-C* mutant. The main peak corresponds to the mass of the holoprotein. The theoretical average mass of the apoprotein is calculated to be 16,188.1 Da, whereas the mass of the Zn^{2+} complex is 16,253.5 Da. **b** Mass spectra of the ΔN25-NColE7 mutant in the presence of onefold to 20-fold molar excess of Zn^{2+} ions. The theoretical average mass of the apoprotein is 13,123.7 Da, whereas the mass of the Zn^{2+} complex is calculated to be 13,189.1 Da

Protein–Zn^{2+} interaction

Mass spectrometry investigations

Intact protein mass spectrometry was used to identify the truncated proteins and to further investigate their Zn^{2+} ion binding abilities. Figure 2a shows the mass spectrum of the purified ΔN4-NColE7-C* mutant recorded in the volatile ammonium acetate buffer without addition of Zn^{2+} ions. The apparent mass of the main peak, i.e.,16,251.9 Da, corresponded to the mass of the holoprotein. This clearly demonstrates that the ΔN4-NColE7-C* mutant was purified in its Zn^{2+}-bound form. The multiply charged spectrum showed at almost all charge states the presence of a significant amount of acetato complex as a result of a noncovalent interaction. Its amount increased with the decrease in the protein's charge state (see Fig. S2 for the *m*/*z* spectrum). Since the metal binding site consists of three histidine imidazole nitrogen ligands from the HNH motif, the presence of the acetate ligand, completing the tetrahedral coordination around the Zn^{2+} ion, is expected here instead of that of the phosphate ion, which usually occurs in the crystal structures.

In contrast to ΔN4-NColE7-C*, Fig. 2b shows that the purified ΔN25-NColE7 mutant did not contain Zn^{2+} ions, and that it was not able to complete the metallation of the apo form even in the presence of 20-fold molar excess of Zn^{2+} ions at pH 6.7 (pH of the ammonium acetate buffer used for mass spectrometry measurements). The apparent mass of the apoprotein (13,124.8 Da) fits very well with the calculated theoretical mass. As a result of increasing amounts of Zn^{2+} ions, the molar ratio of the holoprotein increased. In the presence of tenfold metal ion excess, the molar ratio of the apoprotein and the holoprotein is approximately 1:1, whereas 20-fold excess of Zn^{2+} ions is required to achieve approximately 85 % metallation. The K_D value calculated from the ratio of the ion signal intensities of the apoprotein and the holoprotein in the *m*/*z* spectra of the 11 times charged ion (see the series of spectra in Fig. S3) was 74 ± 18 μM, assuming that no dissociation occurs during the transmission through the mass spectrometer and the metal ion binding to the protein does not alter the ionization efficiency of the non-covalent complex [59]. It should, however, be noted that the estimated constant mentioned above would largely depend on the protonation state of the protein molecule. This means that the stability of the metal ion complex is lower than that reported for the Zn^{2+} binding of the nuclease domain of ColE9 (nanomolar K_D) and is similar to that for Ni^{2+} binding of the same protein [33]. These data unambiguously indicate that the 21 N-terminal amino acids of the ΔN4-NColE7-C* mutant play an important role in the metal binding in the HNH motif at the C-terminus of the protein. The amino acids of the N-terminal loop may, e.g., affect the dynamics of the protein folding and promote the formation of the proper structure of the protein.

Fluorimetry

Fluorimetry can also be applied to monitor the Zn^{2+} ion binding of proteins by probes that are fluorescent in their zinc(II) complexes [45], such as the TFLZn probe used by us. The maximal fluorescence intensity at 490 nm was monitored (Fig. S4). In the solutions containing the TFLZn probe and the truncated NColE7 proteins in 2:1 molar ratio, different behaviour was observed for ΔN25-NColE7 and ΔN4-NColE7-C*. There was no significant change in the fluorescence intensity of TFLZn in the presence of ΔN25-NColE7, whereas the addition of 1 equiv [$c(Zn^{2+}) = c$(protein)] of Zn^{2+} ions caused a large increase in the intensity. At the same time, an increase of the fluorescence was observed upon addition of ΔN4-NColE7-C* to the TFLZn solution. The resulting fluorescence intensity was in both cases significantly higher in the presence of the proteins (and metal ion) than in the Zn^{2+}–TFLZn binary system. This suggests that in agreement with the mass spectrometry result, ΔN25-NColE7 does not contain Zn^{2+} ions, and that the proteins cannot completely replace the TFLZn probe in the coordination sphere of the Zn^{2+} ion. The latter can be explained supposing that the proteins do not fill all the coordination sites around the metal ion (coordination occurs through the three histidine side chains). Therefore, according to the thermodynamics of the system, the formation of Zn^{2+}(protein)(TFLZn) ternary complexes is also possible, in which an enhancement of the fluorescence can be observed. The addition of DNA to protein-containing solutions slightly decreased the fluorescence, probably owing to the replacement of the dye in the ternary complex. The slightly larger extent of the change for the ΔN4-NColE7-C* protein points to its stronger DNA binding (see later).

SRCD spectroscopy results

In a chiral environment there is a difference between the absorption of the left and right circularly polarized light, and a plot of the difference in their absorption coefficients ($\Delta\varepsilon = \varepsilon_{\mathrm{left}} - \varepsilon_{\mathrm{right}}$) versus wavelength yields a characteristic circular dichroism spectrum of the sample. The relative position of chiral amide chromophores in proteins, i.e. the secondary structure, and its changes are responsible for this effect in the wavelength region of UV light (180–250 nm). Since SRCD spectroscopy provides an optimal and even flux of UV light in a highly controlled manner, it can be applied for accurate study of the solution structure and interactions of proteins [60].

The effect of metal ion binding on the structure of the mutant NColE7 proteins was investigated by monitoring the changes in their SRCD spectra on addition of Zn^{2+} ions and/or EDTA to their solutions, as described in "Materials and methods". Figure 3a shows the spectra obtained for ΔN4-NColE7-C* protein. As can be seen, the addition of Zn^{2+} ions did not affect the SRCD spectrum (not even at fivefold Zn^{2+} excess—data not shown). This again suggests that the ΔN4-NColE7-C* protein already includes a bonded metal ion. At the same time, an excess of EDTA caused a slight decrease in the intensity. This suggests that the removal of the Zn^{2+} ions from the protein by EDTA caused only a negligible change in the secondary structure composition of the protein, suggesting that the structure is also stable without the metal ion.

In similar experiments with ΔN25-NColE7, both the intensity and the shape of the SRCD spectra changed continuously (the spectral pattern becoming more similar to that of the ΔN4-NColE7-C* spectrum) upon gradual addition of up to 10 equiv of Zn^{2+} ions, and the extent of this change became negligible at higher metal ion excess. The addition of an excess of EDTA resulted in a spectrum similar to that recorded in the absence of metal ions (Fig. 3b). These results further show that the shorter protein binds Zn^{2+} ions more weakly than ΔN4-NColE7-C*, which could result from the more extensive distortion of the metal ion binding site upon the deletion of the further 21 amino acids.

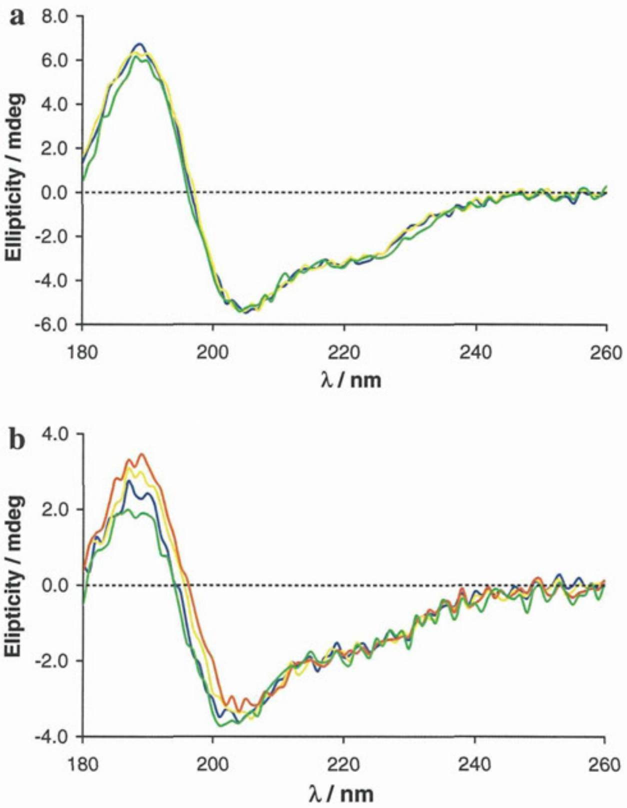

Fig. 3 Comparison of the synchrotron-radiation circular dichroism (SRCD) spectra recorded for **a** ΔN4-NColE7-C* ($c = 36$ μM) and **b** ΔN25-NColE7 ($c = 18$ μM) under various conditions. The spectra of the aqueous solutions of the proteins are in *blue*. The *yellow curves* belong to the systems where 1 equiv of Zn^{2+} ions has been added to the protein solutions. For ΔN25-NColE7, the SRCD spectrum recorded in the presence of 10 equiv of Zn^{2+} ions has also been plotted (*orange curve*), since the change here is more expressed than in the case of ΔN4-NColE7-C*. Finally, an excess of EDTA was added to the previous solutions and the spectra were recorded (*green*). In all cases the average of three measurements was plotted

Protein–DNA interactions

SRCD spectroscopy results

SRCD spectroscopy was also applied to study the dsDNA binding of the mutant proteins. The spectra recorded in the presence of dsDNA are presented in Fig. S5. Although the gel mobility shift experiments (see later) proved there was DNA binding, the recorded spectra showed that the addition of DNA did not change the structure of ΔN4-NColE7-C* in solution. This is in line with the crystal structures of NColE7 and the nuclease domain of ColE9 bound to dsDNA [16, 29, 31, 32, 57]. There was, however, a clear difference between the calculated and experimental spectra upon addition of DNA to ΔN25-NColE7, suggesting that the dsDNA binding induces a slight conformational change in the more flexible mutant, probably by the stabilization of the wild-type structure.

Gel mobility shift assays

To compare the DNA-binding ability of the truncated mutants, a gel mobility shift experiment was conducted, in which increasing amounts of proteins were added to a 0.874 μM solution of an approximately 400 bp dsDNA sample (Fig. 4). For ΔN25-NColE7, approximately ten times greater protein concentration than for ΔN4-NColE7-C* was applied to achieve a substantial gel mobility shift, in agreement with its weaker DNA binding.

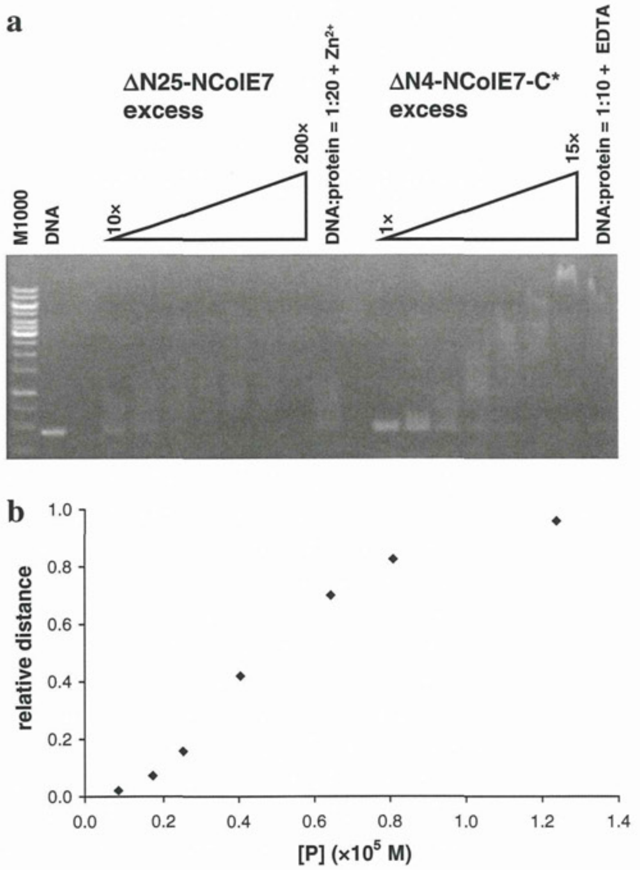

Fig. 4 **a** Gel mobility shift assay for studying the DNA-binding ability of ΔN25-NColE7 (*left*) and ΔN4-NColE7-C* (*right*). The first lane from the left contains a 1,000 bp marker DNA; the second lane contains a 0.874 μM double-stranded DNA sample. In the following lanes equal amounts of the same DNA sample incubated for 1 h with increasing amounts of mutant proteins in a constant volume of 10 μl were loaded. The excess of the proteins was tenfold, 20-fold, 30-fold, 50-fold, 80-fold, 100-fold and 200-fold for ΔN25-NColE7 and onefold, twofold, threefold, fivefold, eightfold, tenfold and 20-fold for ΔN4-NColE7-C*. **b** The relative gel mobility shift of the ΔN4-NColE7-C* mutant (i.e. the normalized distance of the shifted band and the band of the unbound double-stranded DNA; the saturation distance was taken as 1.0) versus the equilibrium protein concentration, [P]. At the inflection point [P] = K_D

Addition of Zn^{2+} ions to the ΔN25-NColE7 solution containing 20-fold excess of protein (see Fig. 4a) did not result in a change in the position of the band. Two reasons may account for this: (1) the protein had already bound metal ions, however, this would be in contrast with our previous results, or (2) the binding of Zn^{2+} ions is not necessary for DNA binding—similarly as for NColE7. In agreement with this latter observation, an excess of EDTA added to the ΔN4-NColE7-C*–DNA system did not cause any change in the position of the DNA band (Fig. 4a).

For ΔN4-NColE7-C* an apparent stability constant was estimated on the basis of the gel mobility shift assay. Simplifying conditions were introduced assuming 1:1 DNA binding site (10 bp DNA)—protein complex (P-DNA) formation and 100 % complex formation at the saturation of the curve. In Fig. 4b the relative gel mobility shift versus equilibrium protein concentration ([P]) is plotted. The latter was estimated as [P] = c_P − [P-DNA], where c_P is the total concentration of the protein, and [P-DNA] is the equilibrium concentration of the protein–DNA complex, which is proportional with the relative distance of the shifted band from the unbound DNA on the gel. K_D = ([P] × [DNA])/[P-DNA] at the inflection point, where 50 % of the DNA binding sites are occupied by the protein, i.e. [DNA] = [P-DNA], and thus [P] = K_D, where K_D is the apparent dissociation constant related to the formation of protein–DNA complexes at each binding site. By the above considerations, K_D was estimated to be approximately 5.0 μM (pK_D ~ 5.3) for a ΔN4-NColE7-C*–DNA binding site complex.

Molecular dynamics calculations

Structural changes of NColE7 and the mutant proteins (it should be noted that in the calculations the native sequences were applied without any tags) in explicit SPC/E water were tracked by 22 ns molecular dynamics calculations. Figure 5a describes the change in the root mean square deviation (RMSD) of backbone atoms in the molecule with respect to the reference structure at the 500th picosecond during the simulation. According to the RMSD diagram for the NColE7 (446–576) simulation, a relatively stable structure is formed after 2.5 ns of solvation, causing a 0.2 nm difference, as compared with the start of the simulation. The structure slightly changes until 10 ns, and then it fluctuates around 0.2 nm.

The structure of the ΔN4-NColE7 (450–576) mutant behaves similarly to the wild-type NColE7, but the ΔN25-

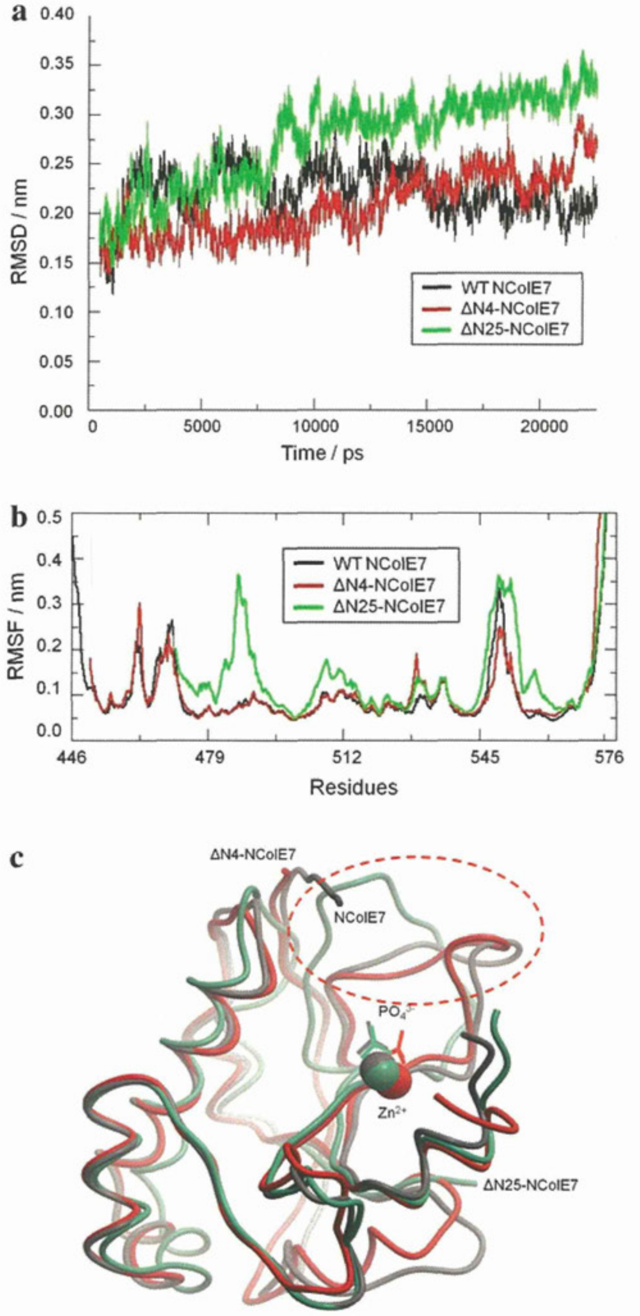

Fig. 5 **a** The root mean square deviation (*RMSD*) versus time as a result of 22 ns molecular dynamics calculations in explicit SPC/E water. **b** The average motion of backbone atoms during the molecular dynamics simulation in proteins. **c** The average structure of the proteins in the 20th to 25th nanosecond range of the simulations. The loop between the two β-strands of the HNH motif is highlighted by a *circle*. *RMSF* root mean square fluctuation, *WT* wild type

NColE7 (471–576) mutant goes through more serious changes, reaching an RMSD of 0.35 nm by the end of simulation. The RMSD for both mutants increased with time during the whole simulation, in contrast to that of wild-type NColE7. This means that shortening the protein caused remarkable changes in protein dynamics even if only four residues were cut at the N-terminus.

Figure 5b shows the average motion of each backbone atoms (root mean square fluctuation) during the whole simulation. The HNH motif is at the C-terminal part of the protein, and as such is at the right side of the diagram. The intense peak at about the 550th residue corresponds to the loop that joins the two β-sheets of the HNH motif. It is a functionally important part of the protein: the conserved residue N560 is located here, and is responsible for orienting the general base H545.

A significant difference can be observed between NColE7 and the shortened mutants. Two regions of ΔN25-NColE7, i.e. amino acid residues 485–487 and 511–515, show increased motion. These parts of NColE7 are loops leaning approximately parallel to the original N-terminal part that is missing from ΔN25-NColE7. The α-helices in the neighbourhood remained unchanged. The region including residues 530–535 also forms a loop at the N-terminal end of ΔN25-NColE7, and it interacts with the helix of the HNH motif. Interestingly, these residues also show an increased root mean square fluctuation in the case of ΔN4-NColE7, which suggests that the deletion of the last four residues at the N-terminus has an influence on the dynamics of the middle part of the protein. Residues 547–560 form the loop between the β-strands of the HNH motif. Changes in the dynamics of this loop can strongly influence the function of the protein as mentioned above.

Figure 5c shows the average structure of each mutants taken from the 20th to 25th nanosecond region of the simulation. The most obvious effect of shortening the N-terminus is the change in the orientation of the loop between the two β-strands of the HNH motif (highlighted by a circle in Fig. 5c). In case of ΔN25-NColE7, the missing N-terminal loop caused the two neighbouring loops to approach one another. Therefore, the loop in the HNH motif lost its original orientation. A smaller but not negligible movement can also be seen in ΔN4-NColE7: the HNH loop is also shifted in this case. As mentioned above, there are catalytically important residues here. The change of the average orientation and flexibility of the loop between β-strands of the HNH motif could be a reason for the decreased nuclease activity of the shortened proteins. The N-terminal loop can be considered as a structural spacer between the HNH loop and the DNA binding loop of the protein. It is also worth mentioning, that the N-terminal loop remained in an unchanged position in ΔN4-NColE7 even without the positively charged amino acids being deleted in this mutant. This has also been observed for the ΔN4-NColE7-C* mutant [43].

In a previous study of NColE7 [32], it was supposed that NColE7 can bind the DNA substrate in two different manners: coordinating also a water molecule or without it.

That is, the Zn^{2+} ion may have a temporary fifth coordination site that can provide a general acid (assisting in the protonation of the leaving group) in the form of an induced coordinated water molecule. During the 25 ns simulations, no such structure was found: there was no water molecule near the metal ion. However, a change was detected in the solvent distribution around the Zn^{2+} ion in the mutants. The shorter the protein, the looser the structure, which allows more water molecules to get near to or in the active centre (Fig. S6).

Semiempirical quantum chemical calculations

PM6/MOZYME/conductor-like screening model semiempirical quantum chemical calculations were performed to further investigate the fine changes in the active centre of the protein. Figure 6 shows the active centre in the optimized structures of NColE7, ΔN4-NColE7 and ΔN25-NColE7. The proteins were superimposed with PyMOL [58] using the whole length of the corresponding sequences. The RMSD of the full-length backbone relative to the initial structure of the calculations (PDB ID 1M08) was 1.010 nm for NColE7 (127 atoms fitted), 0.604 nm for ΔN4-NColE7 (109 atoms fitted) and 0.893 for ΔN25-NColE7 (100 atoms fitted). Aligning with the optimized structure of the NColE7 resulted in an RMSD of 0.527 nm (110 atoms) for ΔN4-NColE7 and 0.824 (89 atoms) for ΔN25-NColE7. The active centre of the ΔN4-NColE7 mutant is similar to that of the wild-type enzyme, as the orientation of the histidine side chains is almost identical (Fig. 6b). However, the small differences in the structures lead to different orientations of the phosphate and Zn^{2+} ions. This is even more obvious in the case of ΔN25-NColE7. The changed geometry around Zn^{2+} and phosphate ions could be a reason for the decreased metal- and DNA-binding ability of the ΔN25-NColE7 mutant. This again points to the fact that the removal of the N-terminal part has a significant effect on the structure of the C-terminal active centre.

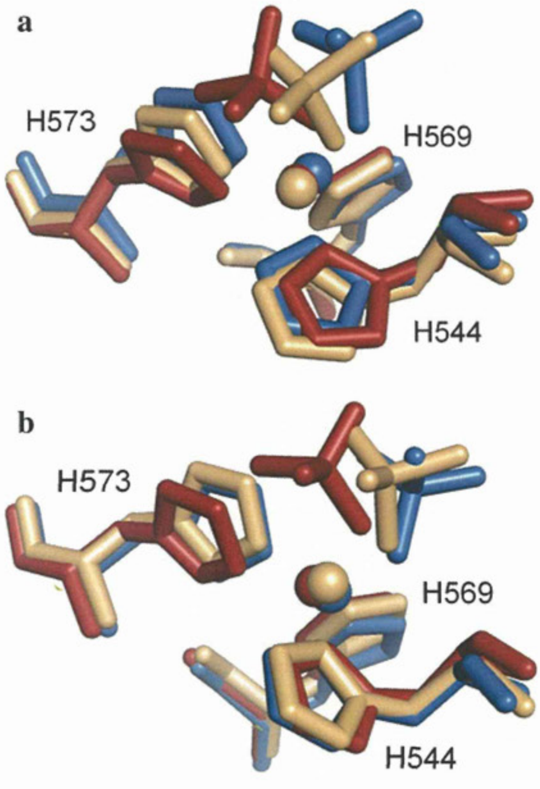

Fig. 6 Superposition of the metal ion binding residues in the optimized structures of NColE7 (*blue*), ΔN4-NColE7 (*yellow*) and ΔN25-NColE7 (*red*) with phosphate and zinc ions. The backbone atoms of the proteins were aligned with **a** the 1M08 structure [6] and **b** the optimized structure of NColE7

Conclusions

The necessity of the arginine residue at the N-terminus for hydrolytic activity of NColE7 poses the possibility of positive allosteric control in this protein. Mass spectrometry, SRCD and fluorescence spectroscopy and agarose gel mobility shift assays provided information on the effect of the removal of N-terminal sequences on the Zn^{2+} ion and DNA binding in ΔN4-NColE7-C* and ΔN25-NColE7 mutants. The longer protein bound both Zn^{2+} ion and DNA more strongly than its shorter counterpart owing to the structural stabilization effect of the N-terminal amino acids. The C-terminal mutation in ΔN4-NColE7-C* might affect these properties, but our results here and in [43] strongly suggest that the C-terminal flanking sequence does not participate in the metal ion or DNA binding. Molecular dynamics and semiempirical quantum chemical calculations performed in parallel showed that the absence of the N-terminal sequences resulted in significantly increased movement of the backbone atoms in regions of possible interactions with the N-terminal loop: residues 485–487, 511–515 and 570–571 for ΔN25-NColE7, and residues 467–468, 530–535 and 570–571 for ΔN4-NColE7. The distortion of the active centre predicted by semiempirical quantum chemical calculations could also be the reason for weak Zn^{2+} binding of ΔN25-NColE7. These results lead to the conclusion that the N-terminal loop plays an important role in the positioning of the arginine residue for the control of the DNAse activity. The question arose whether this could be a common feature among the HNH family of endonucleases. Since the amino acid sequences of the bacterial colicins and pyocins display high similarity, the arginine is frequently found in a position similar to that in NColE7. Also in the available crystal structures of different members of the HNH family, e.g. Sm endonuclease, Vvn

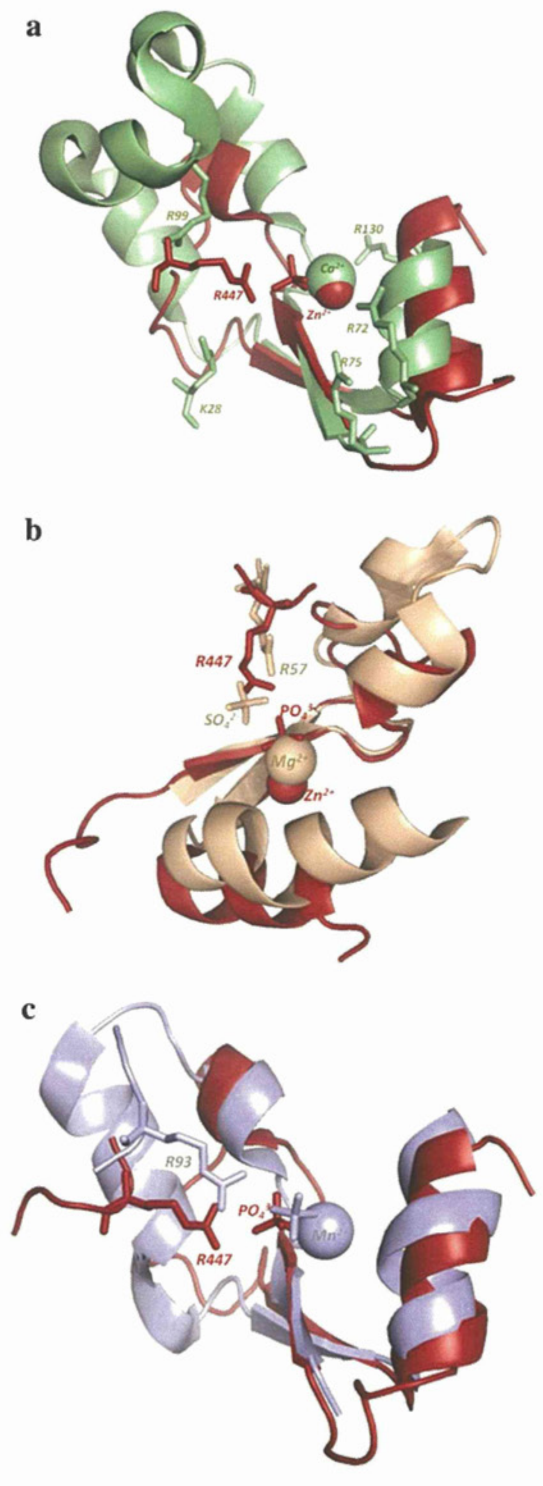

Fig. 7 The alignment of the HNH motifs of NColE7 (*red*, PDB ID 1MZ8) and selected proteins belonging to the HNH superfamily. **a** Vvn endonuclease (PDB ID 1OUP). **b** Sm endonuclease (PDB ID 1G8T). **c** Nuclease *A* (PDB ID 1ZM8)

endonuclease or nuclease A, we found arginine side chains in the spatial vicinity of the catalytic centre (Fig. 7). The answer thus seems to be positive. However, since the arginines are not always situated at the N-termini of the proteins, it is difficult to identify them without knowing the 3D structure. Therefore, a detailed bioinformatic study is required.

Acknowledgments This work has received support through the Hungarian Science Foundation (OTKA-NKTH CK80850), TÁMOP-4.2.1/B-09/1/KONV-2010-0005 and TÁMOP-4.2.2/B-10/1-2010-0012. The computational resources at High Performance Computing of the University of Szeged and financial support from the European Union Research Infrastructure Action FP7 (Integrated Activity on Synchrotron and Free Electron Laser Science, contract no. FP7/2007-2013; no. 226716) are also gratefully acknowledged. B.G. thanks the Japan Society for the Promotion of Science, and I.N.J-S. and A.C. thank the Hungarian Scholarship Board for the fellowships provided.

References

1. Chak K-F, Kuo W-S, Lu F-M, James R (1991) J Gen Microbiol 137:91–100
2. Lin Y-H, Liao C-C, Liang P-H, Yuan HS, Chak K-F (2004) Biochem Biophys Res Commun 318:81–87
3. Liao C-C, Hsia K-C, Liu Y-W, Leng P-H, Yuan HS, Chak K-F (2001) Biochem Biophys Res Commun 284:556–562
4. Cheng Y-S, Shi Z, Doudeva LG, Yang W-Z, Chak K-F, Yuan HS (2006) J Mol Biol 356:22–31
5. Sui M-J, Tsai L-C, Hsia K-C, Doudeva LG, Ku W-Y, Han GW, Yuan HS (2002) Protein Sci 11:2947–2957
6. Cheng Y-S, Hsia K-C, Doudeva LG, Chak K-F, Yuan HS (2002) J Mol Biol 324:227–236
7. Chak K-F, Safo MK, Ku W-Y, Hsieh S-Y, Yuan HS (1996) Proc Natl Acad Sci USA 93:6437–6442
8. Hsieh S-Y, Ko T-P, Tseng M-Y, Ku W-Y, Chak K-F, Yuan HS (1997) EMBO J 16:1444–1454
9. Dennis CA, Videler H, Paupit RA, Wallis R, James R, Moore GR, Kleanthous C (1998) Biochem J 333:183–191
10. Ko T-P, Liao C-C, Ku W-Y, Chak K-F, Yuan HS (1999) Structure 7:91–102
11. Kleanthous C, Walker D (2001) Trends Biochem Sci 26:624–631
12. Kolade OO, Carr SB, Kühlmann UC, Pommer A, Kleanthous C, Bouchcinsky CA, Hemmings AM (2002) Biochimie 84:439–446
13. Orlowski J, Bujnicki JM (2008) Nucleic Acids Res 36:3552–3569
14. Eastberg JH, Eklund J, Monnat R, Stoddard BL (2007) Biochemistry 46:7215–7225
15. Mehta P, Katta K, Krishnaswamy S (2004) Protein Sci 13:295–300
16. Hsia K-C, Chak K-F, Liang P-H, Cheng Y-S, Ku W-Y, Yuan HS (2004) Structure 12:205–214
17. Michel-Briand Y, Baysse C (2002) Biochimie 84:499–510
18. Shen BW, Landthaler M, Shub DA, Stoddard BL (2004) J Mol Biol 342:43–56
19. Ghosh M, Meiss G, Pingoud A, London RE, Pedersen LC (2005) J Biol Chem 280:27990–27997
20. Kriukiene E, Lubiene J, Lagunavicius A, Lubys A (2005) Biochim Biophys Acta 1751:194–204
21. Saravanan M, Bujnicki JM, Cymerman IA, Rao DN, Nagaraja V (2004) Nucleic Acids Res 32:6129–6135
22. Saravanan M, Vasu K, Ghosh S, Nagaraja V (2007) J Biol Chem 282:32320–32326
23. Cymerman IA, Obarska A, Skowronek KJ, Lubys A, Bujnicki MJM (2006) Proteins 65:867–876
24. Jakubauskas A, Giedriene J, Bujnicki JM, Janulaitis A (2007) J Mol Biol 370:157–169

25. Sokolowska M, Czapinska H, Bochtler M (2009) Nucleic Acids Res 37:3799–3810
26. Veluchamy A, Mary S, Acharya V, Mehta P, Deva T, Krishnaswamy S (2009) Bioinformation 6:80–83
27. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A (2008) Nucleic Acids Res 36:D281–D288
28. Huang H, Yuan HS (2007) J Mol Biol 368:812–821
29. Wang Y-T, Yang W-J, Li C-L, Doudeva LG, Yuan HS (2007) Nucleic Acids Res 35:584–594
30. Papadakos G, Wojdyla JA, Kleanthous C (2012) Q Rev Biophys 45:57–103
31. Mate MJ, Kleanthous C (2004) J Biol Chem 279:34763–34769
32. Doudeva LG, Huang H, Hsia K-C, Shi Z, Li C-L, Shen Y, Cheng C-L, Yuan HS (2006) Protein Sci 15:269–280
33. Pommer AJ, Kuhlmann UC, Cooper A, Hemmings AM, Moore GR, James R, Kleanthous C (1999) J Biol Chem 274:27153–27160
34. Hannan JP, Whittaker SBM, Hemmings AM, James R, Kleanthous C, Moore GR (2000) J Inorg Biochem 79:365–370
35. Keeble AH, Hemmings AM, James R, Moore GR, Kleanthous C (2002) Biochemistry 41:10234–10244
36. van den Bremer ETJ, Jiskoot W, James R, Moore GR, Kleanthous C, Heck AJR, Maier CS (2002) Protein Sci 11:1738–1752
37. Hannan JP, Whittaker SB, Davy SL, Kuhlmann UC, Pommer AJ, Hemmings AM, James R, Kleanthous C, Moore GR (1999) Protein Sci 8:1711–1713
38. van den Bremer ETJ, Keeble AH, Visser AJWG, van Hoek A, Kleanthous C, Heck AJR, Jiskoot W (2004) Biochemistry 43:4347–4355
39. Ku W-Y, Liu Y-W, Hsu Y-C, Liao C-C, Liang P-H, Yuan HS, Chak K-F (2002) Nucleic Acids Res 30:1670–1678
40. Shi Z, Chak K-F, Yuan HS (2005) J Biol Chem 280:24663–24668
41. Li C-L, Hor L-I, Chang Z-F, Tsai L-C, Yang W-Z, Yuan HS (2003) EMBO J 22:4014–4025
42. Gyurcsik B, Czene A (2011) Future Med Chem 3:1935–1966
43. Tóth E, Czene A, Gyurcsik B, Otten H, Poulsen J-CN, Larsen S, Christensen HEM, Nagata K (2013) Acta Crystallogr Sect D
44. Limao-Vieira P, Giuliani A, Delwiche J, Parafita R, Mota R, Duflot D, Flament JP, Drage E, Cahillane P, Mason NJ, Hoffmann SV, Hubin-Franskin MJ (2006) Chem Phys 324:339–349
45. Fahrni CJ, O'Halloran TV (1999) J Am Chem Soc 121:11448–11458
46. Berendsen HJC, van der Spoel D, van Drunen R (1995) Comput Phys Commun 91:43–56
47. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) J Chem Theory Comput 4:435–447
48. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF (2004) J Comput Chem 25(13):1656–1676
49. Olsson MHM, Søndergaard CR, Rotkowski M, Jensen JH (2011) J Chem Theory Comput 7:525–537
50. Stewart JJP (2008) MOPAC2009. Stewart Computational Chemistry, Colorado Springs. http://openmopac.net
51. Stewart JJP (2007) J Mol Model 13:1173–1213
52. Stewart JJP (2009) J Mol Model 15:765–805
53. Stewart JJP (1996) Int J Quantum Chem 58:133–146
54. Klamt A, Schüümann G (1993) J Chem Soc Perkin Trans 2 799–805
55. Anthony LC, Suzuki H, Filutowicz M (2004) J Microbiol Methods 58:243–250
56. Levin KB, Dym O, Albeck S, Magdassi S, Keeble AH, Kleanthous C, Tawfik DS (2009) Nat Struct Mol Biol 16:1049–1055
57. Wang Y-T (2009) Wright JD, Doudeva LG, Jhang H-C, Lim C, Yuan HS. J Am Chem Soc 131:17345–17353
58. DeLano WL (2006) PyMOL version 0.99rc6. DeLano Scientific, San Carlos
59. Jecklin MC, Schauer S, Dumelin CE, Zenobi R (2009) J Mol Recognit 22:319–329
60. Miles AJ, Wallace BA (2006) Chem Soc Rev 35:39–51

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The structure of the deacetylase domain of *Escherichia coli* PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family

Takashi Nishiyama, Hiroki Noguchi, Hisashi Yoshida, Sam-Yong Park and Jeremy R. H. Tame*

Protein Design Laboratory, Graduate School of Nanobioscience, Yokohama City University, Suehiro 1-7-29, Yokohama, Kanagawa 230-0045, Japan

Correspondence e-mail:
jtame@tsurumi.yokohama-cu.ac.jp

Received 5 July 2012
Accepted 8 October 2012

PDB Reference: N-terminal domain of PgaB, 3vus

Bacterial biofilm formation is an extremely widespread phenomenon involving the secretion of a protective exopolysaccharide matrix which helps the bacteria to attach to surfaces and to overcome a variety of stresses in different environments. This matrix may also include proteins, lipids, DNA and metal ions. Its composition depends on the bacterial species and growth conditions, but one of the most widely found components is polymeric β-1,6-*N*-acetyl-D-glucosamine (PGA). Several studies have suggested that PGA is an essential component of biofilm and it is produced by numerous bacteria, including *Escherichia coli*, *Staphylococcus epidermis*, *Yersinia pestis*, *Bordetella* spp. and *Actinobacillus* spp. In *E. coli*, PGA production and export are dependent on four genes that form a single operon, *pgaABCD*, which appears to have been transferred between various species. Biofilms themselves are recognized as environments in which such horizontal gene transfer may occur. The *pga* operon of *E. coli*, which is even found in innocuous laboratory strains, is highly homologous to that from the plague bacterium *Yersinia pestis*, and biofilm is believed to play an important role in the transmission of *Yersinia*. The crystal structure of the N-terminal domain of PgaB, which has deacetylase activity, is described and compared with models of other deacetylases.

1. Introduction

Bacteria may adopt either a free-living ‘planktonic’ or a sessile lifestyle. Sessile bacterial growth of a structured community of cells attached to a surface is known as a biofilm (Hall-Stoodley *et al.*, 2004). Biofilm formation is a general feature of microorganisms, and has been of intense interest for decades owing to its medical relevance; metabolically quiescent bacteria within biofilms are hidden from the immune system and are difficult to treat with antibiotics, and biofilms frequently form on catheters or other surgical implants (Costerton *et al.*, 1999). Polymeric β-1,6-*N*-acetyl-D-glucosamine (PGA) is a principal component of the exopolysaccharide matrix of many bacteria. Without the *pgaABCD* operon, *Escherichia coli* is unable to form biofilms (Wang *et al.*, 2004), and cleavage of PGA breaks up biofilms (Itoh *et al.*, 2005). The enzymes encoded by the operon produce a linear homopolymer of β-1,6-linked *N*-acetylglucosamine. This polymer was first described in studies of *Staphylococcus epidermis* and was referred to as polysaccharide intracellular adhesin (PIA; Mack *et al.*, 1996). As well as being involved in bacterial attachment to biofilm (Agladze *et al.*, 2005), PGA can play important roles in

doi:10.1107/S0907444912042059

host–microbe interactions. It has been implicated in the colonization and virulence of both Gram-positive and Gram-negative bacteria (Cerca *et al.*, 2007; Vuong, Kocianova *et al.*, 2004; Vuong, Voyich *et al.*, 2004). Biofilm formation by *E. coli*, in particular the role of the *pgaABCD* operon, has been studied in detail by the group of Romeo (Itoh *et al.*, 2008; Wang *et al.*, 2004). Transcription of the *pgaABCD* operon is repressed by CsrA (carbon storage regulatory A), an RNA-binding protein that binds to the untranslated leader of target mRNA (Wang *et al.*, 2005). *pgaABCD* transcription requires NhaR (a LysR-family DNA-binding protein), which switches on PGA production in response to high pH and also to high concentrations of salt (Goller *et al.*, 2006).

PgaC and PgaD are inner membrane proteins, with the latter having no sequence similarity to any known structure; PgaC is a member of the GT-2 glycosyltransferase family. They consume UDP-GlcNAc from the cytoplasm and release PGA into the periplasm. PgaA and PgaB are not directly involved in PGA synthesis, but are required for its export (Itoh *et al.*, 2008). PgaA is believed to form a β-barrel in the outer membrane through which the polymer passes to exit the cell; a homologous structure from *Pseudomonas aeruginosa* has recently been crystallized and shown to form an 18-stranded β-barrel (Whitney *et al.*, 2011). PgaB is anchored to the outer membrane by attachment of lipid; it has a classical signal sequence followed by a cysteine (Cys21), which is the lipid-attachment site. Sequence analysis of PgaB suggests that the protein has an N-terminal deacetylase domain, a member of carbohydrate esterase family 4 (CE4) in the CAZy classification scheme (http://www.cazy.org; Cantarel *et al.*, 2009). Other members of this family act on carbohydrate polymers such as xylan or chitin, but only PgaB has been reported to be involved in PGA synthesis or export. The C-terminal domain of PgaB has an unknown function, but may bind to PgaA or to PGA itself. The deacetylase activity of PgaB appears to be necessary for PGA export, presumably by the exposure of amine groups, which acquire a positive charge (Itoh *et al.*, 2008). NMR analysis of PGA obtained by overexpression of the *pgaABCD* operon showed that the polymer has a high molecular weight (about 400 kDa) but contains fewer than 3% deacetylated residues (Wang *et al.*, 2004). PGA from other species may show a much higher proportion; that from *S. epidermis*, for example, shows 15–20% deacetylation, although the proportion may vary according to the growth conditions (Vuong, Kocianova *et al.*, 2004). Published studies therefore suggest that the deacetylase activity of PgaB is essential for biofilm formation but that it is only weakly active. The growing polymer chain presumably passes close to PgaB in order to be acted upon by it, yet fewer than one residue in 20 undergoes modification. The proportion of residues which must be deacetylated in order to promote secretion appears to be low, but the possibility exists that inhibition of this weak enzyme activity could be an effective route to prevent biofilm formation by *E. coli.* Biofilm formation by the plague bacterium *Yersinia pestis* involves an operon (*hmsHFRS*) that is highly homologous to *E. coli pgaABCD* and appears to play a significant role in bacterial transmission (Bobrov *et al.*, 2008; Darby, 2008). We have therefore solved the crystal structure of the catalytic domain of PgaB (PGABN), the first CE4 model from this organism, and compared it with known structures.

2. Materials and methods

2.1. Cloning

pgaB was amplified by PCR from *E. coli* JM109 chromosomal DNA and cloned into a modified pET28 expression vector using *Bam*HI and *Xho*I restriction sites. The primer sequences were CGGGATCCGCCCAGTCAAGAACATC-ATTTATACCG and CCGAGCCTCGAGTTACTGGAG-GTTTTCGTCATAAAC. The PCR product was digested with *Bam*HI and *Xho*I at 310 K for 2 h before purification using a QIAquick PCR Purification Kit (Qiagen). The purified PCR product was ligated into the cut vector using T4 DNA ligase (Wako) at room temperature for 1 h. The ligation mixture was used to transform *E. coli* DH5α, and pET28b-pgaBN was prepared from cultures using QIAprep (Qiagen). This construct directs expression of residues Gln24–Gln330 of PgaB with a hexahistidine tag at the N-terminus that is cleavable with TEV protease.

2.2. Expression and purification

pET28b-pgaBN was transformed into *E. coli* BL21 (DE3) and cells were grown at 310 K with shaking in 3 l LB medium containing kanamycin (50 μg ml^{-1}). When the OD$_{600}$ of the culture reached 0.5–0.6, PGABN expression was induced by adding IPTG to a final concentration of 0.5 m*M* and growth was continued overnight at 288 K. The cells were collected by centrifugation at 3000*g* at 277 K for 30 min. The pellet was suspended in 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl and was then lysed by sonication on ice. The lysate was centrifuged at 38 000*g* and 277 K for 30 min. The supernatant solution was loaded onto a 10 ml nickel Sepharose column (GE Healthcare) equilibrated with 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl, 10 m*M* imidazole; after washing, it was eluted with 50 m*M* Tris–HCl pH 8.0, 250 m*M* imidazole. The major protein fractions were collected and digested with TEV protease overnight at 277 K during dialysis into 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl. The protease:PgaB ratio was 1:50. The protein was reloaded onto the washed nickel Sepharose column and eluted with 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl. The pooled fractions containing PGABN were dialyzed into 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl before concentration to 30 mg ml^{-1} using Amicon centrifugal filter units (Millipore).

2.3. Crystallization and structure determination

Crystallization experiments were performed at 293 K using the hanging-drop vapour-diffusion method. Crystals grew in 16%(*w*/*v*) PEG 3350, 0.18 *M* sodium acetate, 0.1 *M* bis-Tris–HCl pH 6.5. Data were collected on beamline 17A of the Photon Factory, Tsukuba. The highest resolution data were collected from a crystal which had been soaked briefly in 1 m*M* mercury chloride but which appeared to be native in phasing trials. The data were used in the final refinement and

Table 1
Data-collection and refinement statistics.

Values in parentheses are for the outer shell.

Data set	Native	Hg, remote	Hg, inflection	Hg, peak	Pt
Space group	$P2_1$	$P2_1$			$P2_1$
Wavelength (Å)	1.00000	0.99321	1.00938	0.99957	1.00000
Unit-cell parameters (Å, °)	a = 39.6, b = 53.1, c = 144.2, β = 95.2	a = 39.0, b = 51.9, c = 144.7, β = 95.5			a = 39.3, b = 53.0, c = 144.4, β = 95.5
Resolution range (Å)	50.0–1.65 (1.68–1.65)	50.0–2.50 (2.54–2.50)			50.0–2.30 (2.34–2.30)
Reflections (measured/unique)	337723/63364	87449/19886	88848/19847	86805/19838	106033/25579
Completeness (%)	96.7/83.7	98.0/96.6	98.1/97.1	98.0/97.3	96.0/86.2
$R_{\rm merge}$† (%)	4.9/37.5	4.9/19.1	4.4/13.3	5.4/12.4	7.5/39.8
Multiplicity	4.9	4.4	4.5	4.4	4.2
$\langle I/\sigma(I)\rangle$	49.0	44.3	47.9	48.2	34.9
Refinement statistics					
Resolution range (Å)	25.0–1.65				
R factor/free R factor	0.205/0.257				
R.m.s.d. bond lengths (Å)	0.027				
R.m.s.d. bond angles (°)	2.29				
No. of water molecules	175				
Average B factor (protein/water) (Å^2)	33.4/35.2				
Ramachandran plot, residues in (%)					
Most favourable regions	87.6				
Additional allowed regions	11.1				
Generously allowed regions	0.9				
Disallowed regions	0.4				

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of an observation, $\langle I(hkl)\rangle$ in the mean value for that reflection and the summations are over all equivalents.

revealed a bound Hg atom with low occupancy on refinement. A total of 250 images of 1° oscillation were collected for each data set. Data processing and scaling were carried out with *HKL*-2000 and *SCALEPACK* (Otwinowski & Minor, 1997). The space group was found to be $P2_1$, with two molecules in the asymmetric unit. Data statistics are given in Table 1. Multiple-wavelength data were collected to 2.5 Å resolution on the same beamline using a crystal soaked in 2 m*M* mercury chloride for 14 h. A single data set to 2.3 Å resolution was also collected using a crystal soaked in 2 m*M* K_2PtCl_4 for 24 h. The native data set and the Pt-soak data set were collected using incident radiation of 1.000 Å wavelength. Phases were calculated using *PHENIX* (Adams *et al.*, 2010). Model building was carried out with *ARP*/*wARP* (Langer *et al.*, 2008; Morris *et al.*, 2003) and *Coot* (Emsley *et al.*, 2010; Emsley & Cowtan, 2004). Refinement was carried out with *REFMAC* (Murshudov *et al.*, 2011) and the *CCP*4 suite (Winn *et al.*, 2011). Noncrystallographic symmetry restraints and TLS group refinement were not applied. The Ramachandran plot showed several residues in unusual positions, but the agreement between the two copies of the molecule was very good. Slight main-chain disorder at residues 66 and 100 led to several pairs of equivalent residues in chains *A* and *B* lying at different positions in the Ramachandran plot. Two Ramachandran outliers were the active-site residues His55 and Asp115. Isotropic temperature factors were refined with default restraints. Figures were prepared with *PyMOL* (DeLano, 2002). Water molecules were checked manually. A single Hg atom was modelled into the structure with an occupancy of 20%. Density around the Zn atom of chain *B* showed less than full occupancy of the acetate, and a partially ordered water molecule was also modelled at this site. The final model and structure factors have been deposited in the Protein Data Bank as entry 3vus.

3. Results

3.1. Overall structure

Initial sequence analysis of the *pgaB* gene from *E. coli* suggested that the deacetylase activity resided in an N-terminal domain (PGABN) of roughly 300 residues in length. A threading search of known structures using the *Wurst* server (Torda *et al.*, 2004) suggested a strong match to aldolase from *Trypanosoma brucei* (PDB entry 1f2j), which has 14% sequence identity to the PgaB N-terminus (Chudzik *et al.*, 2000). This TIM β-barrel structure has no enzymatic activity in common with PgaB, but has eight $\alpha\beta$ structure repeats, whereas the CE4 enzymes have seven. PGABN shows sequence similarity to PdaA, an *N*-deacetylase from *Bacillus subtilis*, which is a CE4 member of known structure (Blair & van Aalten, 2004). On the basis of these results, a construct was made to express PgaB from Gln24 to Gln330 by PCR from *E. coli* genomic DNA, omitting the signal peptide and associated cysteine residue. This construct yielded 25 mg purified protein (with the N-terminal histidine tag removed) per litre of culture. Analytical ultracentrifugation showed the protein to exist as a monomer in solution (data not shown).

Crystals were grown that diffracted to almost 1.6 Å resolution and the structure (shown in Fig. 1*a*) was solved using two heavy-atom derivatives and multiple-wavelength anomalous measurements. Two copies of the molecule were found in the asymmetric unit. The ordered residues visible in the electron-density map began at Pro43 and ended at Val308 or

Gln309, so that roughly 20 disordered residues at each end of the polypeptide were not modelled. EXAFS analysis of a native crystal clearly showed the presence of zinc, a metal not purposely added to the protein at any stage of preparation or purification. In each copy of the molecule a single metal ion was readily located and modelled (Fig. 1*b*). A search for related models in the PDB using *DALI* (Holm & Rosenström, 2010) yielded the best match as SlCE4 (PDB entry 2cc0), an acetylxylan esterase from *Streptomyces lividans* involved in plant cell-wall degradation which has a structure similar to PdaA (Taylor *et al.*, 2006). Both enzymes are members of Pfam family PF01522. Overlaying the structures with *SSM* (Krissinel & Henrick, 2004) matched 128 residues of PdaA with PGABN with 18% sequence identity, giving an C^α r.m.s.d. of 2.3 Å. Similar results were obtained matching PGABN to SlCE4, giving an alignment of 138 residues (19.6% identity) and 2.4 Å r.m.s.d. Straightforward sequence searches against the PDB gave much poorer alignments, with the best fits being about 30% identical sequences of roughly 50 residues in length. Although most of these hits were to CE4 enzymes, they also included a subunit of the yeast ribosome and clearly were not all biologically relevant. Examination of the models quickly showed that PGABN is circularly permuted with respect to known CE4 members and that the last β-strand of their TIM-like barrel corresponds to the first strand of PGABN. Simple sequence alignments of PGABN with other members of the CE4 family therefore suggested quite different matches to those found from comparisons of the actual models and did not match conserved active-site residues. A sequence alignment between SlCE4 and PGABN is shown in Fig. 2(*a*), demonstrating the shift of one strand of the central barrel from the N-terminus of PGABN to the C-terminus of other CE4 enzymes. The result of fitting PGABN and SlCE4 by *SSM* is shown in Fig. 2(*b*).

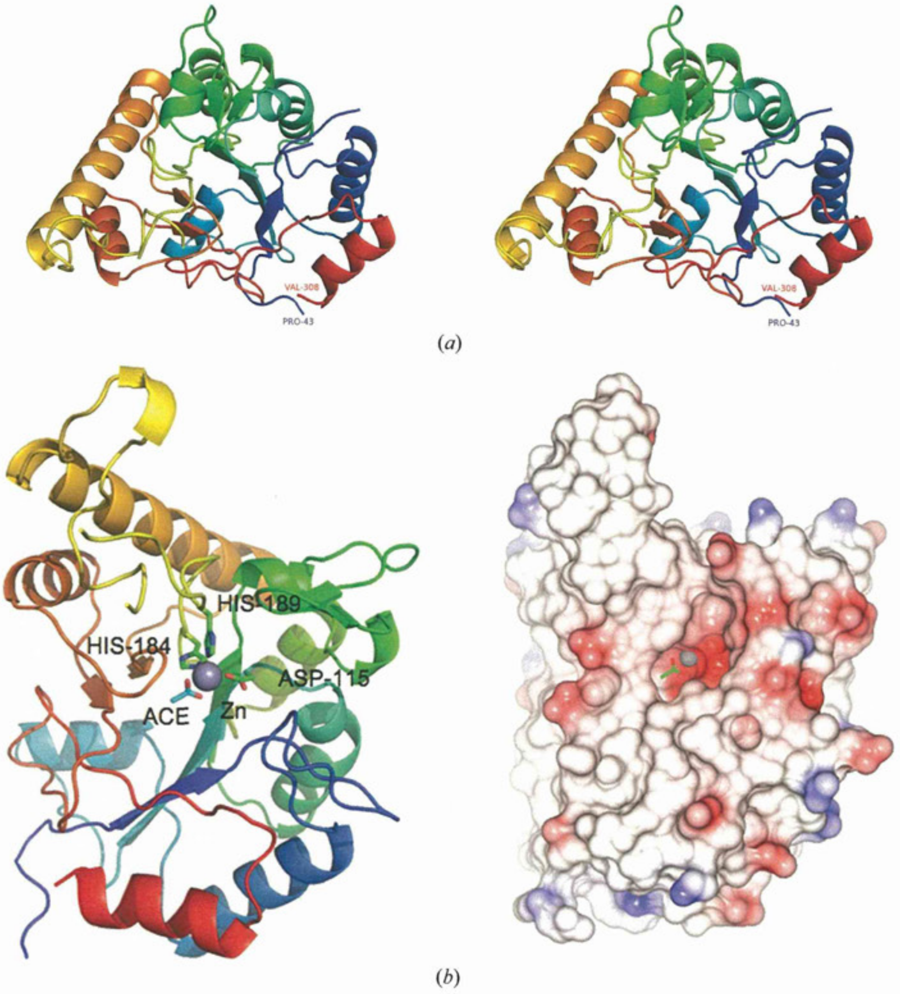

Figure 1
(*a*) A stereoview of the C^α trace of the PgaB N-terminal domain (PGABN), looking into the central barrel. The trace is coloured from blue (Pro43) to red (Val308), with α-helices shown as coils and β-strands as arrows. The figure was drawn using *PyMOL* (DeLano, 2002). Secondary structure was determined automatically. (*b*) The C^α trace of the PGABN monomer, coloured as in (*a*), is shown on the left, with the active-site zinc ion shown as a grey sphere. Residues coordinating the metal are shown as sticks, with O atoms coloured red and N atoms blue. The right-hand panel shows a surface representation of the monomer in the same orientation coloured by electrostatic potential. The zinc ion and acetate ligand were omitted from the calculation of the potential and are shown as ball-and-stick models over the protein surface. The strong negative charge of the active site is apparent. The remainder of the protein surface shows no distinct pattern of charge.

3.2. Active site

The active site of the CE4 family is found at the centre of the barrel-like β-sheet and involves a metal ion in some, but not all, cases. PdaA notably has no metal in the active site on purification and does not bind zinc ions even when these are added; cadmium ions do bind, but only at very high concentrations (Blair & van Aalten, 2004). Other members of the CE4 family have been reported to require cobalt for activity or to prefer cobalt to zinc (Taylor *et al.*, 2006). Nevertheless, two active-site residues (histidine and aspartic acid) were identified from a comparison of the PdaA and PgaB sequences; mutation of either residue in PgaB to alanine (D115A and H184A) blocked activity (Itoh *et al.*, 2008). PGABN shares the common coordination pattern of a zinc ion bound by two histidines and an aspartic acid residue (Fig. 3). The initial crystallization screen and subsequent optimization showed that the inclusion of acetate in the buffer greatly improved the crystal quality, and an acetate ion can be found coordinated to the metal ion of each monomer in the asymmetric unit. However, in one of these sites the occupancy seems to be less than 1.0, and a water molecule with partial occupancy was also modelled close to the zinc ion. Well diffracting

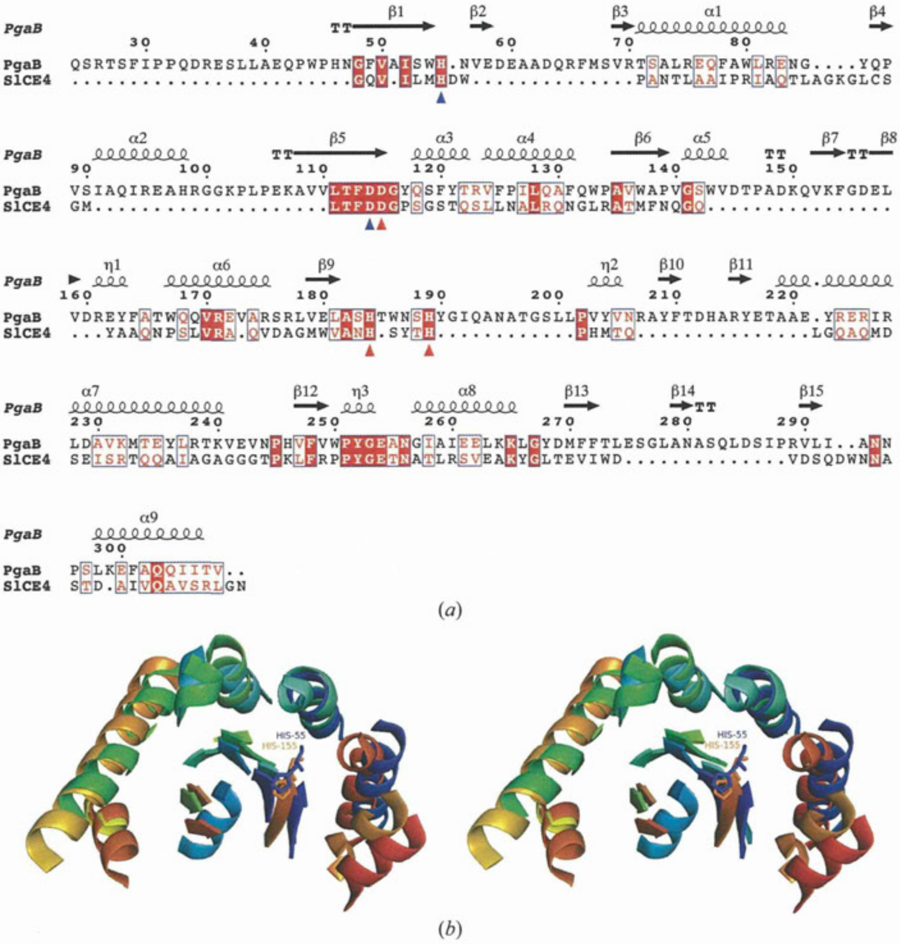
PgaB QSRTSFIPPQDRESLLAEQPWPHNGFVAISWH.NVEDEAADQRFMSVRTSALREQFAWLRENG....YQP
SlCE4 ........................GQV.ILMHDW.............PANTLAAIPRIAQTLAGKGLCS
PgaB VSIAQIREAHRGGKPLPEKAVVLTFDDGYQSFYTRVFPILQAFQWPAVWAPVGSWVDTPADKQVKFGDEL
SlCE4 GM...................LTFDDGPSGSTQSLLNALRQNGLRATMFNQGQ................
PgaB VDREYFATWQQVREVARSRLVELASHTWNSHYGIQANATGSLLPVYVNRAYFTDHARYETAAE.YRERIR
SlCE4 ...YAAQNPSLVRA.QVDAGMWVANH.SYTH............PHMTQ...............LGQAQMD
PgaB LDAVKMTEYLRTKVEVNPHVFVWPYGEANGIAIEELKKLGYDMFFTLESGLANASQLDSIPRVLI..ANN
SlCE4 SEISRTQQAIAGAGGGTPKLFRPPYGETNATLRSVEAKYGLTEVIWD..............VDSQDWNNA
PgaB PSLKEFAQQIITV..
SlCE4 STD.AIVQAVSRLGN
(a)
HIS-55
HIS-155
(b)

Figure 2
(*a*) A sequence alignment based on the crystal structures of PGABN and SlCE4 (PDB entry 2cc0). Conserved residues are shown in white on a red background. Red triangles show residues that coordinate the active-site zinc ion and blue triangles show nearby residues involved in catalysis. The residue numbering refers to PGABN. The first β-strand of PGABN (residues 48–55) matches the last strand of SlCE4, the sequence of which has been shifted from the C-terminus to the N-terminus in this alignment. His55 of PGABN matches His155 of SlCE4. The figure was produced by *ESPript* (Gouet *et al.*, 1999). (*b*) Overlay of the C$^\alpha$ traces of the PgaB N-domain and SlCE4 (PDB entry 2cc0), showing the helices and strands of both structures, but not coil. Alignment of the models was carried out with *SSM* (Krissinel & Henrick, 2004). Each structure is coloured from blue to red (N-terminus to C-terminus) as in (*a*), but circular permutation of the sequences leads to a marked difference in colouring, despite the close structural similarity. His55 of PGABN lies close to the active site, overlapping His155 of SlCE4, which is a conserved histidine in the CE4 family.

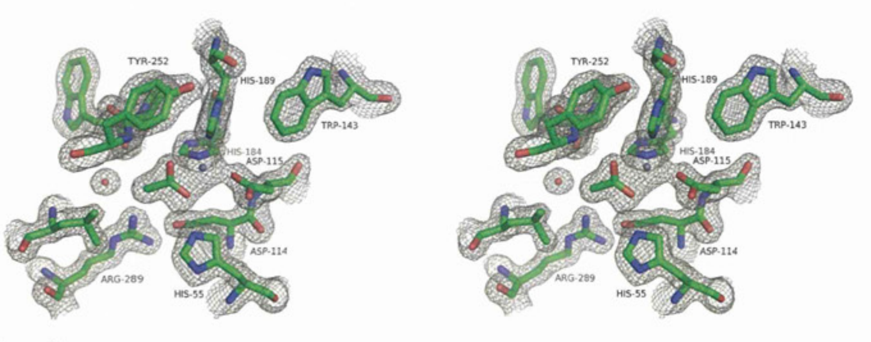

Figure 3
A stereo figure of the $2mF_o - DF_c$ electron-density map for the refined model covering the active site. Electron density is shown at a level of 1σ. The zinc ion is shown as a grey sphere and water molecules are shown as red spheres. The acetate ion coordinating the zinc ion is clearly visible in the density map and is found to hydrogen-bond to the conserved His55. The conserved LT*X*DDG motif of the NodB domain includes Asp114 and Asp115.

crystals could not be grown in the absence of acetate, suggesting that this ligand stabilized the protein. Attempts to remove acetate by soaking the crystals in acetate-free buffer limited the diffraction resolution.

Comparing PGABN with PdaA, it can be seen that many features of the active site are preserved. 128 C$^\alpha$ atoms from the core of the structures show roughly 17% sequence identity, and the zinc-binding residues His184 and His189 of PGABN closely overlap with His124 and His128 of PdaA. Loss of metal binding by PdaA is explained by the fact that it has no third coordinating residue; Asp115 of PGABN is replaced by an asparagine residue (Asn74) that points away from the active site (Fig. 4*a*). Other conserved residues of the active site of PdaA such as Phe98, Trp187 and Leu220 are not found in PGABN, the fold of which is quite different from PdaA in the region of the latter two residues. The active-site residue Asp73 of PdaA is preserved as Asp114 in PGABN, and in both enzymes this aspartic acid forms a salt bridge with an arginine side chain. This arginine residue (Arg163 in PdaA and Arg289 in PGABN) is found on a different β-strand of the barrel; this does not reflect the altered order of the secondary-structure elements but is a consequence of a quite different spatial geometry. Another conserved histidine residue (His55) lies close to the active site at the end of the first β-strand in PGABN, but its equivalent in PdaA (His222) lies at the end of the last β-strand of the barrel. In PdaA this histidine forms a salt bridge (His222–Asp193), but the aspartate has no equivalent in PGABN and His55 is only stabilized by a water molecule. However, His55 does hydrogen-bond to the acetate placed in density near the zinc ion (Fig. 3) and must contact the substrate in the active site. The ion pairs in this region of PdaA have been proposed to play a role in stabilizing charges during catalysis (Blair & van Aalten, 2004), but the details of any such mechanism presumably differ in members of the CE4 family with or without active-site metal ions.

Overlaying SlCE4 with PGABN gives a similar pattern to that with PdaA (Fig. 4*b*). Although SlCE4 is metal-dependent and PdaA has no metal centre, these two enzymes are not circularly permuted relative to one another, and the Asp73–Arg163 pair of PdaA is exactly mirrored by Asp12–Arg100 in SlCE4. His55 of PGABN is preserved as His155 in SlCE4. Thus, although the three enzymes share a common overall three-dimensional structure, each has an active site with characteristics not found in the others. PGABN and SlCE4 presumably share a common mechanism and the altered residues around the active site reflect the substrate preference.

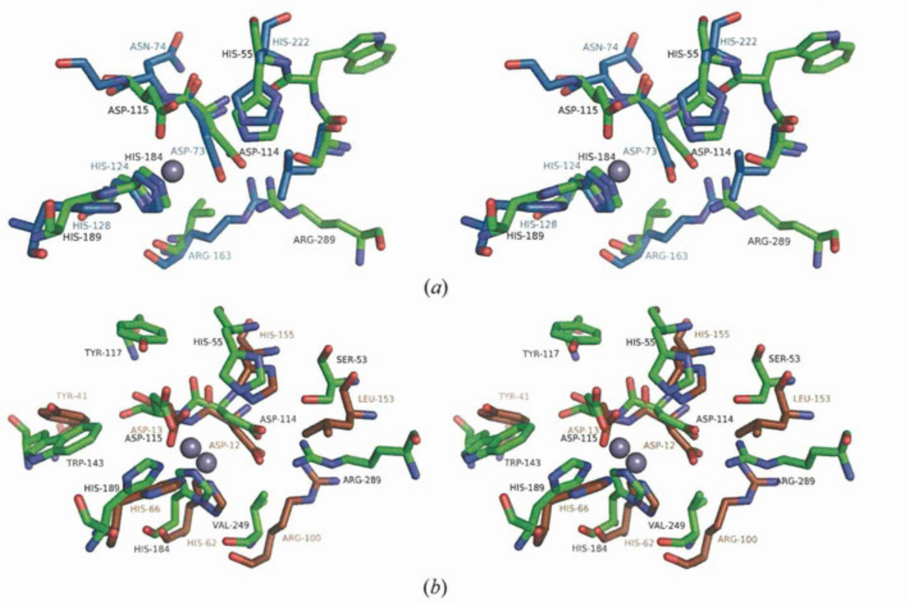

Figure 4
(*a*) A stereo overlay of PGABN (green) and PdaA (light blue) showing the active site. Labels in black indicate the residues in PGABN. O atoms are coloured red and N atoms blue. The models were fitted by overlapping 128 C$^{\alpha}$ atoms of the conserved core residues. The zinc ion of PGABN is shown as a grey sphere. PdaA (PDB entry 1w17) does not bind metal ions owing to the loss of the coordinating aspartate residue. Instead, Asn74 points away from the active site. Arg163 forms a salt bridge to Asp73, but the equivalent arginine in PGABN is Arg289, which sits on a different β-strand to Arg163 of PdaA. (*b*) Stereo overlay of the active sites of PGABN (green) and SlCE4 (brown). Both structures have a bound zinc ion with very similar coordination geometry. Other residues around the active site show low sequence conservation between the two structures, and Tyr117 of PGABN has no counterpart in SlCE4.

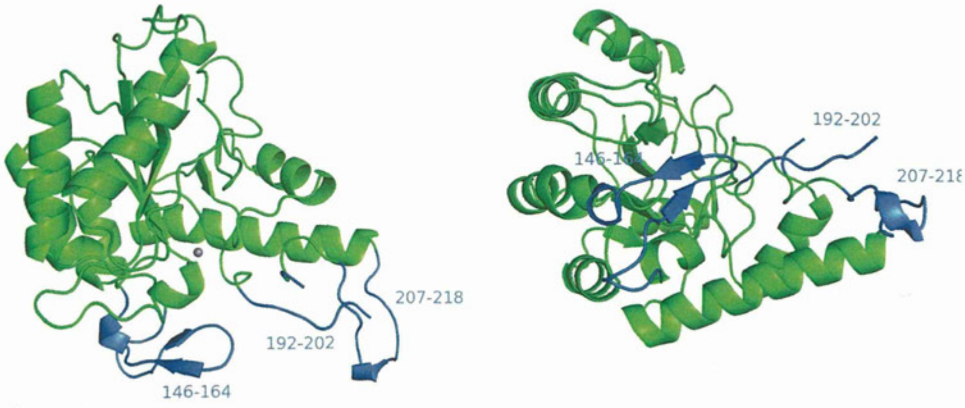

Figure 5
Two views of the C$^{\alpha}$ trace of PGABN rotated 90° relative to one another about a horizontal axis. The zinc ion is shown as a grey sphere. Surface loops near the active site are shown in blue. These regions are presumably flexible in solution and have no counterpart in SlCE4 or PdaA. They may bind to PgaA in the periplasm. Residues 195–200 are not visible in the electron-density map for the copy of the molecule shown.

3.3. Substrate binding

CE4 enzymes have a highly distorted barrel that lacks one of the $\alpha\beta$ repeats of regular TIM barrels, which creates a groove into which the extended polymer substrates of these enzymes can fit. PdaA has several positively charged residues lining this pocket (such as Lys34, Arg35 and Arg166) that are proposed to be involved in binding to the negatively charged peptidoglycan substrate. PGABN has no equivalent residues and acts upon an uncharged substrate; its enzyme activity is also presumably strongly affected by the insertion of two loops blocking one end of the substrate groove, apparently a feature unique to PGABN among the CE4 enzymes (Fig. 5). Residues Thr146–Phe164 face residues Tyr190–Pro202 on the opposite side of the groove. Neither loop appears to be stably fixed in position and residues 195–200 are not modelled in one copy of the molecule, but the presence of these loops may help to control access to the active site. A further loop (Arg207–Thr218) is more distant from the active site but may impede a long polymeric substrate from easy access to it. Overall, the structured domain of PGABN is rather longer than SlCE4 owing to these and other loops, which decorate its surface. Not all of these loops appear to be mobile; the surface loop from Ser276 to Pro288 appears to be fixed in place by hydrogen bonds formed between the side chain of Gln283 and the main-chain atoms of neighbouring residues.

Repeated attempts to grow well diffracting crystals of PGABN without acetate failed. The reason why acetate, the product of the reaction, strongly promotes crystallization is unclear from the structure, but the enzyme appears to be product-inhibited. As with previous structures from this family, we were therefore obliged to attempt to fit substrate analogues into the binding site by docking (Blair *et al.*, 2005). No clearly preferred binding sites were found which can be confidently predicted to bind single glucosamine residues, but the polymeric substrate may bind by weak interactions at several adjacent sites, just as a protease may require a minimum length of peptide in order to cut a single peptide bond. *In vivo*, it is also possible that the substrate is held in proximity to PgaB by interaction with other proteins in the synthetic machinery.

4. Discussion

A number of structures from the CE4 family have been solved by X-ray crystallography: the first of these was solved by a structural genomics consortium (PDB entry 1nyl; Northeast Structural Genomics Consortium, unpublished work), but no accompanying report has been published to date. The first enzyme from this family to be studied in detail was NodB, which removes the acetyl group from a GlcNAc residue in one step of the synthesis of Nod factors, which are molecular signals involved in the symbiosis of legumes and nitrifying bacteria (John *et al.*, 1993). A NodB homology domain was identified in this sequence (Kafetzopoulos *et al.*, 1993) and has been shown to be present in a variety of esterases with different substrates (Caufrier *et al.*, 2003). This conserved region begins roughly 20 residues upstream of the highly conserved LT*X*DDG motif, where *X* is F or Y (Fig. 2*a*). However, accurate sequence alignment of PGABN without the X-ray model was made difficult by the unique inserts that form loops near the substrate-binding site, and sequence analysis alone was unable to show that PGABN is a circularly permuted member of the CE4 family. The NodB domain forms part of a conserved tertiary structure but with associated secondary-structure elements attached to the N-terminus instead of the C-terminus, an arrangement not previously described. PGABN maintains commonly found features at the active site, including coordination to a catalytic zinc ion and nearby aspartic acid and histidine residues. The observed weak activity of the enzyme presumably arises from obstruction of the polymeric substrate by the unique surface loops of PGABN and/or a low intrinsic affinity of the catalytic domain for PGA. The crystallization of a fragment of PgaB has recently been reported. Little and coworkers attempted to crystallize full-length PgaB and identified a crystallizable fragment by proteolysis (Little, Whitney *et al.*, 2012). Our approach was to express both the N-terminal domain and the full-length protein directly, but so far only the N-terminal domain has yielded crystals.

Circular permutation of proteins has been known for some time and has become a route to artificial proteins of enhanced stability (Yu & Lutz, 2011). Hydrolases form one of the largest groups of naturally occurring circular permutants, although many types of protein are found to have such variants (Lo *et al.*, 2009). PGABN is clearly an offshoot of the CE4 family, as shown by the close similarity of the CE4 structures, and the fact that the N- and C-termini of PGABN are close in space, roughly 10 Å apart, a distance which can easily be bridged by a few amino-acid residues. It seems unlikely that a permuted variant of PGABN with the connectivity of the parent family would have markedly different properties from the wild-type protein. Circular permutation has also been noted to play a role in the evolution of calcium-dependent carbohydrate-binding modules involved in xylan recognition (Montanier *et al.*, 2010).

The very high level of conservation shown between PGABN and the equivalent domain of HmsF from *Yersinia* is unequivocal evidence of gene transfer between species. Codon-usage analysis implies that the gene has been introduced into *E. coli.* Whereas *Yersinia* apparently requires biofilm for its preferred method of infection, blocking the gut of fleas to cause them to expel the bacteria into the bloodstream of an animal, exopolysaccharide clearly plays a very different role in a bacterium found in the soil and the gut of higher animals. The gene used in the work described in this paper involved an attenuated laboratory strain of *E. coli.* The *hms* operon of *Yersinia* is named for haemin storage and is found within the *pgm* (pigmentation) locus, a 102 kb long region associated with colouration and iron uptake. The *hmsHFRS* operon was originally shown to be required for the haemin-storage phenotype, but subsequently it was found that an extra gene *hmsT* was also required which lies far from the *hmsHFRS* operon (outside the *pgm* locus) on the genome of *Y. pestis* (Jones *et al.*, 1999). This extra gene is not required for biofilm formation and is not found in *Y. enterolitica*. *E. coli* strain MG1655 has homologues for all five genes (*ycdSRQPT*), but does not show a *hms*$^+$ phenotype. Only *ycdQ* and *ycdP* from *E. coli* MG1655 (equivalent to *pgaC* and *pgaD* in *E. coli* K-12) complement mutations in *hmsR* and *hmsS* in *Y. pestis* (Jones *et al.*, 1999).

Biofilm research is an extremely active and growing field, driven by both medical and biotechnological goals, and the mucoid phenotype of *P. aeruginosa* is a much-studied model (Franklin *et al.*, 2011). This organism causes chronic lung infections in cystic fibrosis patients and produces an exopolysaccharide layer of alginate, a linear polymer of 1,4-linked β-D-mannouronic acid and its C5 epimer α-L-glucuronic acid. The structures of the alginate-export proteins clearly resemble those involved in PGA export in *E. coli* (Keiski *et al.*, 2010; Whitney *et al.*, 2011), raising the possibility that a similar strategy may be employed to tackle biofilm production by both microorganisms. After this paper was reviewed, the structure of PgaB was reported independently by the Howell group (Little, Poloczek *et al.*, 2012).

The authors are extremely grateful to the staff of the Photon Factory for the provision of beamtime in the aftermath of the tsunami disaster of 2011.

References

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

Agladze, K., Wang, X. & Romeo, T. (2005). *J. Bacteriol.* **187**, 8237–8246.

Blair, D. E., Schüttelkopf, A. W., MacRae, J. I. & van Aalten, D. M. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 15429–15434.

Blair, D. E. & van Aalten, D. M. (2004). *FEBS Lett.* **570**, 13–19.

Bobrov, A. G., Kirillina, O., Forman, S., Mack, D. & Perry, R. D. (2008). *Environ. Microbiol.* **10**, 1419–1432.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). *Nucleic Acids Res.* **37**, D233–D238.

Caufrier, F., Martinou, A., Dupont, C. & Bouriotis, V. (2003). *Carbohydr. Res.* **338**, 687–692.

Cerca, N., Maira-Litrán, T., Jefferson, K. K., Grout, M., Goldmann, D. A. & Pier, G. B. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 7528–7533.

Chudzik, D. M., Michels, P. A., de Walque, S. & Hol, W. G. J. (2000). *J. Mol. Biol.* **300**, 697–707.

Costerton, J. W., Stewart, P. S. & Greenberg, E. P. (1999). *Science*, **284**, 1318–1322.

Darby, C. (2008). *Trends Microbiol.* **16**, 158–164.

DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Franklin, M. J., Nivens, D. E., Weadge, J. T. & Howell, P. L. (2011). *Front. Microbiol.* **2**, 167.

Goller, C., Wang, X., Itoh, Y. & Romeo, T. (2006). *J. Bacteriol.* **188**, 8022–8032.

Gouet, P., Courcelle, E., Stuart, D. I. & Métoz, F. (1999). *Bioinformatics*, **15**, 305–308.

Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. (2004). *Nature Rev. Microbiol.* **2**, 95–108.

Holm, L. & Rosenström, P. (2010). *Nucleic Acids Res.* **38**, W545–W549.

Itoh, Y., Rice, J. D., Goller, C., Pannuri, A., Taylor, J., Meisner, J., Beveridge, T. J., Preston, J. F. III & Romeo, T. (2008). *J. Bacteriol.* **190**, 3670–3680.

Itoh, Y., Wang, X., Hinnebusch, B. J., Preston, J. F. III & Romeo, T. (2005). *J. Bacteriol.* **187**, 382–387.

John, M., Röhrig, H., Schmidt, J., Wieneke, U. & Schell, J. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 625–629.

Jones, H. A., Lillard, J. W. & Perry, R. D. (1999). *Microbiology*, **145**, 2117–2128.

Kafetzopoulos, D., Thireos, G., Vournakis, J. N. & Bouriotis, V. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 8005–8008.

Keiski, C. L., Harwich, M., Jain, S., Neculai, A. M., Yip, P., Robinson, H., Whitney, J. C., Riley, L., Burrows, L. L., Ohman, D. E. & Howell, P. L. (2010). *Structure*, **18**, 265–273.

Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nature Protoc.* **3**, 1171–1179.

Little, D. J., Poloczek, J., Whitney, J. C., Robinson, H., Nitz, M. & Howell, P. L. (2012). *J. Biol. Chem.* **287**, 31126–31137.

Little, D. J., Whitney, J. C., Robinson, H., Yip, P., Nitz, M. & Howell, P. L. (2012). *Acta Cryst.* F**68**, 842–845.

Lo, W.-C., Lee, C.-C., Lee, C.-Y. & Lyu, P.-C. (2009). *Nucleic Acids Res.* **37**, D328–D332.

Mack, D., Fischer, W., Krokotsch, A., Leopold, K., Hartmann, R., Egge, H. & Laufs, R. (1996). *J. Bacteriol.* **178**, 175–183.

Montanier, C., Flint, J. E., Bolam, D. N., Xie, H., Liu, Z., Rogowski, A., Weiner, D. P., Ratnaparkhe, S., Nurizzo, D., Roberts, S. M., Turkenburg, J. P., Davies, G. J. & Gilbert, H. J. (2010). *J. Biol. Chem.* **285**, 31742–31754.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2003). *Methods Enzymol.* **374**, 229–244.

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.* D**67**, 355–367.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Taylor, E. J., Gloster, T. M., Turkenburg, J. P., Vincent, F., Brzozowski, A. M., Dupont, C., Shareck, F., Centeno, M. S., Prates, J. A., Puchart, V., Ferreira, L. M., Fontes, C. M., Biely, P. & Davies, G. J. (2006). *J. Biol. Chem.* **281**, 10968–10975.

Torda, A. E., Procter, J. B. & Huber, T. (2004). *Nucleic Acids Res.* **32**, W532–W535.

Vuong, C., Kocianova, S., Voyich, J. M., Yao, Y., Fischer, E. R., DeLeo, F. R. & Otto, M. (2004). *J. Biol. Chem.* **279**, 54881–54886.

Vuong, C., Voyich, J. M., Fischer, E. R., Braughton, K. R., Whitney, A. R., DeLeo, F. R. & Otto, M. (2004). *Cell. Microbiol.* **6**, 269–275.

Wang, X., Dubey, A. K., Suzuki, K., Baker, C. S., Babitzke, P. & Romeo, T. (2005). *Mol. Microbiol.* **56**, 1648–1663.

Wang, X., Preston, J. F. III & Romeo, T. (2004). *J. Bacteriol.* **186**, 2724–2734.

Whitney, J. C., Hay, I. D., Li, C., Eckford, P. D., Robinson, H., Amaya, M. F., Wood, L. F., Ohman, D. E., Bear, C. E., Rehm, B. H. & Howell, P. L. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 13083–13088.

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

Yu, Y. & Lutz, S. (2011). *Trends Biotechnol.* **29**, 18–25.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structures of haemoglobin from woolly mammoth in liganded and unliganded states

Hiroki Noguchi,[a] Kevin L. Campbell,[b] Chien Ho,[c] Satoru Unzai,[a] Sam-Yong Park[a]* and Jeremy R. H. Tame[a]*

[a]Protein Design Laboratory, Yokohama City University, Suehiro 1-7-29, Tsurumi-ku, Yokohama 230-0045, Japan, [b]Department of Biological Sciences, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada, and [c]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Correspondence e-mail:
park@tsurumi.yokohama-cu.ac.jp,
jtame@tsurumi.yokohama-cu.ac.jp

Received 19 April 2012
Accepted 28 June 2012

PDB References: woolly mammoth haemoglobin, deoxy form, 3vre; carbonmonoxy form, 3vrf; met form, 3vrg

The haemoglobin (Hb) of the extinct woolly mammoth has been recreated using recombinant genes expressed in *Escherichia coli*. The globin gene sequences were previously determined using DNA recovered from frozen cadavers. Although highly similar to the Hb of existing elephants, the woolly mammoth protein shows rather different responses to chloride ions and temperature. In particular, the heat of oxygenation is found to be much lower in mammoth Hb, which appears to be an adaptation to the harsh high-latitude climates of the Pleistocene Ice Ages and has been linked to heightened sensitivity of the mammoth protein to protons, chloride ions and organic phosphates relative to that of Asian elephants. To elucidate the structural basis for the altered homotropic and heterotropic effects, the crystal structures of mammoth Hb have been determined in the deoxy, carbonmonoxy and aquomet forms. These models, which are the first structures of Hb from an extinct species, show many features reminiscent of human Hb, but underline how the delicate control of oxygen affinity relies on much more than simple overall quaternary-structure changes.

1. Introduction

The great majority of mammals produce a single major haemoglobin (Hb) component which represents 90% or more of the Hb in the red cells. This protein must therefore adapt to the habitat of the animal in order to both extract sufficient oxygen from the air and deliver it to the tissues efficiently. For instance, vertebrates that live at high altitudes, including high-flying birds, have Hbs with relatively high oxygen affinity (Hiebl *et al.*, 1989; Jessen *et al.*, 1991; Weber, 2007), and evolution has found several elegant solutions to the problem of adjusting this property of the protein (Perutz, 1983). Most vertebrate Hbs are heterotetramers composed of two α-type and two β-type subunits, each of which carries a single haem group for the reversible binding of oxygen molecules. Oxygen transport is effected by the Hb molecule binding oxygen cooperatively, so that it saturates in the lungs and releases oxygen readily within the body. Purified human Hb is found to have a very high oxygen affinity, but this is reduced inside red cells by a roughly equimolar concentration of 2,3-diphosphoglycerate (DPG, also known as BPG). This polyanion binds to the protein near the N-terminus of the β subunits and lowers oxygen affinity. Near actively respiring tissues, oxygen release is also assisted by the presence of carbon dioxide (Bohr *et al.*, 1904), which can bind directly to Hb and, *via* its conversion to carbonic acid, also lowers the pH of the blood, leading to the well known Bohr effect. Finally, Hb affinity can be further modulated by chloride ions, although the precise mechanisms are under debate and remain to be elucidated. The molecular structure of Hb from humans and other animals is known through the pioneering work of Perutz, whose group discovered that the protein adopts two basic quaternary conformations (Perutz, 1970, 1972). These were identified with the T and R states of the allosteric model proposed by Monod, Wyman and Changeux (Monod *et al.*, 1965), so that cooperatively was suggested to arise as the low-affinity T state switched to the high-affinity R state with increasing oxygenation. While this essential picture has been thoroughly tested, it clearly neglects other very important effects. The oxygen affinity of the haems is not determined solely by the quaternary state of the protein, but, as noted above, by heterotropic ligands which bind at various positions on the protein surface (Imai, 1982). It has been demonstrated, for example, that Hb in the R state may have an affinity just as low as the T-state protein: the artificial effector L35 binds to both the T state and R state and reduces the oxygen affinity of both to the same level (Shibayama *et al.*, 2002; Yokoyama *et al.*, 2006). The textbook description of Hb lays strong emphasis on the allosteric switch of the protein, but it is now clear that dynamic effects are also at work, controlling the oxygen affinity of the protein within each quaternary state (Schay *et al.*, 2006; Ho *et al.*, 2011; Yonetani *et al.*, 2002).

The regulation of oxygen affinity by heterotropic effector molecules can also vary among mammalian Hbs. Hb from ungulates such as cows, for example, is found to respond weakly to organophosphates such as DPG; these proteins have an intrinsically low oxygen affinity which makes such effectors unnecessary (Perutz & Imai, 1980). The low intrinsic oxygen affinity of bovine Hb is believed to arise largely from changes in the β-type subunit of the protein, although it has been demonstrated that mutating human HbA to give it a similar N-terminus to the β subunit does not transfer this effect (Fronticelli *et al.*, 1995). Although the pH dependence (Bohr effect) of bovine Hb is similar to that of human Hb, the enthalpy of oxygenation (ΔH) is only about one-third as much at pH 9, where enthalpy effects of protonation can be ignored (roughly -29 kJ mol^{-1} instead of -84 kJ mol^{-1}; Razynska *et al.*, 1990). A reduced enthalpy is highly desirable for animals living in cold environments (Giardina *et al.*, 1990; Weber & Campbell, 2011) since enthalpy is also a measure of the change of free energy of a reaction with temperature, a relationship quantified by the van't Hoff equation. The Hbs of several Arctic species, notably reindeer, have been studied in some detail (Giardina *et al.*, 1990). Unlike human HbA, which has a similar ΔH in the T and R states, for reindeer Hb the enthalpy of oxygenation becomes much smaller with each oxygen ligand. Thus, for human Hb the oxygenation curve is shifted with temperature but maintains the same shape, whereas for reindeer Hb it changes shape with temperature. It is clear that there are numerous possible routes to reducing the ΔH of oxygenation while leaving the Gibbs free energy (ΔG) unchanged, exploiting the well known phenomenon of entropy–enthalpy compensation. The observed heat change on binding of any protein to a ligand in fact reflects a number of different factors including solvent rearrangement (Chervenak & Toone, 1995) and homotropic effects not associated with effector binding.

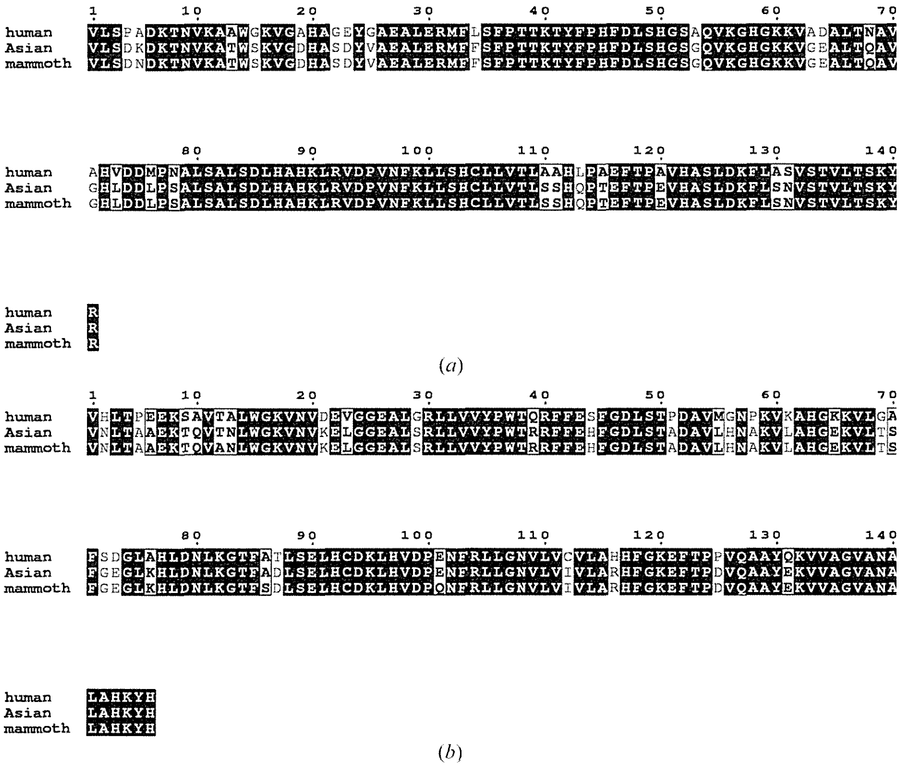
1 10 20 30 40 50 60 70
human VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAV
Asian VLSDKDKTNVKATWSKVGDHASDYVAEALERMFFSFPTTKTYFPHFDLSHGSGQVKGHGKKVGEALTQAV
mammoth VLSDNDKTNVKATWSKVGDHASDYVAEALERMFFSFPTTKTYFPHFDLSHGSGQVKGHGKKVGEALTQAV
80 90 100 110 120 130 140
human AHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY
Asian GHLDDLPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLSSHQPTEFTPEVHASLDKFLSNVSTVLTSKY
mammoth GHLDDLPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLSSHQPTEFTPEVHASLDKFLSNVSTVLTSKY
human R
Asian R
mammoth R
(a)
1 10 20 30 40 50 60 70
human VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGA
Asian VNLTAAEKTQVTNLWGKVNVKELGGEALSRLLVVYPWTRRFFEHFGDLSTADAVLHNAKVLAHGEKVLTS
mammoth VNLTAAEKTQVANLWGKVNVKELGGEALSRLLVVYPWTRRFFEHFGDLSTADAVLHNAKVLAHGEKVLTS
80 90 100 110 120 130 140
human FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANA
Asian FGEGLKHLDNLKGTFADLSELHCDKLHVDPENFRLLGNVLVIVLARHFGKEFTPDVQAAYEKVVAGVANA
mammoth FGEGLKHLDNLKGTFSDLSELHCDKLHVDPQNFRLLGNVLVIVLARHFGKEFTPDVQAAYEKVVAGVANA
human LAHKYH
Asian LAHKYH
mammoth LAHKYH
(b)

Figure 1
(*a*) An alignment of the α-globin sequences of human HbA, Asian (Indian) elephant (*Elephas maximus*) and woolly mammoth (*Mammothus primigenius*). The sequences from *E. maximus* and mammoth differ at one residue (Asn5→Lys), and similarly the modern African elephant (*Loxodonta africana*) shows one amino-acid difference, Ser49→Gly. Residues that are identical in all three sequences are shown in white on black. The figure was produced with *ESPript* (Gouet *et al.*, 1999). (*b*) An equivalent alignment of the β or β/δ subunits. The mammoth β/δ subunit has acquired three mutations relative to the Asian (Indian) elephant, Thrl2→Ala, Ala86→Ser and Glu101→Gln. The single point mutation between modern African and Asian elephants in this subunit, Asp52→Glu, corroborates the evidence from the α subunits that the mammoth and the Asian elephant descended from a common ancestor after the African elephant had diverged.

The ancestors of woolly mammoths (*Mammuthus primigenius*) evolved in subtropical Africa and were phylogenetically closer to modern Asian (also called Indian) elephants than African elephants are. Woolly mammoths only colonized high-latitude environments in the early Pleistocene Ice Ages some 1.2–2.0 million years ago. Presumably under strong selective pressures associated with this new environment, woolly mammoth Hb acquired three

Table 1
Data-collection and refinement statistics.

Values in parentheses are for the outer shell.

	Met	CO	Deoxy
Data collection			
Resolution range (Å)	50.0–1.50 (1.53–1.50)	50.0–1.55 (1.58–1.55)	46.51–2.20 (2.32–2.20)
Space group	*C*2	*C*2	*P*1
Unit-cell parameters			
a (Å)	109.4	109.5	53.6
b (Å)	61.5	61.8	59.1
c (Å)	53.5	53.6	62.1
α (°)			96.7
β (°)	110.1	110.4	114.4
γ (°)			113.9
No. of measured reflections	224380	237928	75756
No. of unique reflections	53178	47308	28619
Completeness (%)	96.7 (96.2)	97.4 (83.2)	94.4 (93.9)
Mean $I/\sigma(I)$	50.09 (14.14)	51.4 (7.99)	6.6 (2.7)
Multiplicity	4.3	2.6	2.6
R_{merge}[†] (%)	4.7 (12.6)	5.0 (18.7)	9.8 (32.7)
Refinement statistics			
Resolution range (Å)	20.0–1.50	20.0–1.55	20.0–2.20
R factor[‡] (%)	18.8	17.2	20.6
R_{free}[‡] (%)	22.0	19.9	26.9
Solvent molecules	242	244	209
R.m.s. deviations from ideal			
Bond lengths (Å)	0.023	0.024	0.012
Bond angles (°)	2.059	2.104	1.359
Chiral volumes (Å^3)	0.129	0.154	0.093
Ramachandran plot, residues in (%)			
Most favourable region	98.2	98.2	97.5
Additional allowed region	1.8	1.8	2.5

[†] $R_{\text{merge}} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of an observation, $\langle I(hkl) \rangle$ is the mean value for that reflection and the summations are over all reflections. [‡] R factor $= \sum_{hkl} \big||F_{\text{obs}}| - |F_{\text{calc}}|\big| / \sum_{hkl} |F_{\text{obs}}|$, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes, respectively. The free R factor was calculated with 5% of the data excluded from the refinement.

amino-acid substitutions, all of which are located on the chimeric β/δ globin chain. Notable among these mutations is the Gluβ/δ101(G3)$\rightarrow$Gln replacement situated within the central cavity of the Hb tetramer near subunit interfaces of the protein that markedly alter its physiochemical properties. In the absence of chloride or the polyanionic effector DPG, mammoth Hb has substantially higher oxygen affinity than elephant Hb, but a very similar heat of oxygenation (Campbell *et al.*, 2010; Yuan *et al.*, 2011). Chloride ions and DPG exert a stronger effect on it, however, so that with these effectors present the oxygen affinity of mammoth Hb is close to that of elephant Hb under the same conditions at 310 K but with a significantly smaller heat release on oxygen binding. A sequence alignment of human and modern Asian elephant globins is shown in Fig. 1.

Five human variant Hbs are known in which the highly conserved β101 position is altered, and all of them show strong increases in intrinsic oxygen affinity. Hb Rush, which carries the same Gluβ101$\rightarrow$Gln mutation as mammoth Hb, is slightly unstable and leads to mild haemolytic anaemia (Adams *et al.*, 1974). Alone among the five variants, Hb Rush also possesses a proton-linked chloride-binding site (Shih *et al.*, 1985). Despite considerable study, the stereochemical basis underlying this difference remains unclear. Here, we have attempted to match the physiochemical properties of woolly mammoth Hb to the structures of the protein in the T and R states. We find no evidence of surface chloride-binding sites and that the unique Gluβ101(G3)$\rightarrow$Gln mutation imposes no significant structural change on the protein.

2. Materials and methods

2.1. Protein crystallization

Mammoth Hb was produced as described by Campbell *et al.* (2010) and stored in the carbonmonoxy form as pellets frozen in liquid nitrogen. To prepare crystals, the protein was dialyzed into 10 m*M* Tris–HCl pH 7.5 and concentrated to 60 mg ml^{-1}. CO was removed as necessary using a table lamp to illuminate a sample rotated on an ice-water bath under a stream of air. Deoxy and CO crystals were grown in rubber-stoppered tubes holding 50 µl samples as described previously (Park *et al.*, 2006). Deoxy crystals were grown under nitrogen using 2.8 *M* ammonium sulfate pH 6.5 as a precipitant. Crystals of the aquo-met (oxidized) form of the protein were grown while exposed to air using 2.4 *M* sodium/potassium phosphate pH 6.7 and the CO form was subsequently also crystallized under the same conditions using 0.1 *M* Tris–HCl pH 6.5, 1.9 *M* ammonium sulfate, and the CO form was subsequently crystallized using 2.2 *M* sodium/potassium phosphate pH 6.7. Crystals were harvested using standard nylon cryoloops (Hampton Research) in mother liquor containing 20% glycerol (25% in the case of the met form) before cooling to 100 K.

2.2. Data collection and structure refinement

Data were collected on beamline BL17A of the Photon Factory, Tsukuba, Japan using an ADSC CCD detector and were processed with *HKL*-2000 (Otwinowski & Minor, 1997). Structure determination was carried out using the *CCP*4 suite of programs (Winn *et al.*, 2011). *MOLREP* (Vagin & Teplyakov, 2010) was used to build initial models by molecular replacement using PDB entries 2dn1 and 2dn2 (Park *et al.*, 2006) as models of human oxy and deoxy Hb. The models were manipulated with *Coot* (Emsley *et al.*, 2010), which was also used for model validation. Refinement was carried out with *REFMAC* (Murshudov *et al.*, 2011). Default restraints (geometric and thermal parameters) were used. In all cases 5% of reflections were used to calculate a free *R* factor. No σ-factor cutoff or resolution cutoff was applied at any stage. Solvent water molecules were modelled into the map where geometrically reasonable with at least 1σ $2mF_{\rm o} - DF_{\rm c}$ electron density to support their inclusion. Structures and structure factors have been deposited in the PDB with codes 3vre (deoxy form), 3vrf (carbonmonoxy form) and 3vrg (met form). Data-collection and refinement statistics are given in Table 1.

3. Results

3.1. Overall structure

The DNA sequences of component globin chains from woolly mammoth have been determined previously (Campbell

Table 2
Indicator hydrogen-bond lengths (Å) in different liganded states of mammoth Hb.

Values in parentheses indicate symmetry-equivalent interactions. A cutoff distance of 3.5 Å was used. Distances are rounded to one decimal place.

	CO	Met	Deoxy
Lys$\alpha_1$40···His $\beta/\delta_2$146	–	–	2.6 (2.8)
Tyr$\alpha_1$42···Asp$\beta/\delta_2$99	–	–	2.6 (2.6)
Asp$\alpha_1$94···Trp$\beta/\delta_2$37	3.7	3.7	2.9 (3.2)
Asp$\alpha_1$94···Asn$\beta/\delta_2$102	2.7	2.8	–
Arg$\alpha_1$141···Asp$\beta/\delta_2$126	–	–	2.7 (2.4)
Asp$\beta/\delta_1$94···His$\beta/\delta_1$146	–	–	2.9 (2.7)
Trp$\beta/\delta_1$37···Asn$\beta/\delta_1$102	3.0	2.9	–

Table 3
Comparison of mammoth Hb models with human and bovine Hb.

Root-mean-square deviations (Å) of the tetramer using the 'Tame frame' (148 C$^\alpha$ atoms from residues α23–48, α57–63, α101–111, α118–125, β51–57, β110–116 and β119–132). The PDB models used were 2dn2 (human deoxy; Park *et al.*, 2006), 2dn1 (human oxy; Park *et al.*, 2006), 1hda (bovine deoxy; Perutz *et al.*, 1993), 1fsx (bovine carbonmonoxy; Safo & Abraham, 2001) and 1bbb (human R2 state; Silva *et al.*, 1992). Figures are quoted to two decimal places. Values in parentheses show the overlaps between alternative $\alpha\beta$ dimers where these are found in the asymmetric unit.

	Tetramer	Dimer
Deoxy mammoth–deoxy human	0.58	0.55 (0.46)
Deoxy mammoth–deoxy bovine	0.58	0.55 (0.41)
Deoxy human–deoxy bovine	0.30	0.22 (0.25)
CO mammoth–CO human	0.95	0.61
CO mammoth–CO bovine	0.88	0.58
CO human–CO bovine	0.82	0.33
CO mammoth–human R2	1.26	0.56
Oxy human–human R2	1.14	0.29
CO bovine–human R2	0.75	0.26

et al., 2010). The sequences are entries D3U1H8 and D3U1H9 in the UniProt database. Rather than a typical β subunit, the β-type chain of mammoth Hb is the product of a chimeric fusion gene (*HBB*/*D*) that arose *via* an unequal crossover event between the parental *HBB* (β) and *HBD* (δ) loci that predates the diversification of paenungulate (elephants, sea cows and hyraxes) mammals (Opazo *et al.*, 2009). As in living elephants, mammoth Hb contains 141 residues in the α subunit and 146 residues in the β/δ subunit. Notably, the mammoth β/δ subunit shows three amino-acid changes compared with the Asian elephant protein, Thr12(A9) to Ala, Ala86(F2) to Ser and Glu101(G3) to Gln, while the α subunit of the Asian species evolved a single replacement, Asn5(A3) to Lys, following its divergence from the mammoth lineage (Campbell *et al.*, 2010).

Despite over 60 million years of independent evolution, the mammoth Hb protein is 81% identical to human HbA, and the crystallographic models are consequently strikingly similar to the human protein in overall structure. One of the most notable features of sequence comparison between the human and mammoth proteins is that HbA has 14 proline residues per $\alpha\beta$ dimer, whereas the mammoth protein (and modern elephant Hb) has only nine, implying a slightly more flexible structure. In the deoxy model, the largest deviation between the C$^\alpha$ traces of mammoth and human Hb is found in the D helix of one β/δ subunit, where Proβ51(D2) is replaced by alanine in mammoth Hb. Additionally, the mammoth protein has Aspβ/δ52(D3) close to Glu$\alpha_2$120(H3), although the aspartate makes a hydrogen bond to Hisβ/δ56(DE2). This Asp–His salt bridge is maintained in the R state and thus is unlikely to contribute to the Bohr effect, which is markedly reduced in Asian elephants relative to HbA (Yuan *et al.*, 2011). Organic anions such as inositol hexaphosphate bind to a single site near the N-termini of the β subunits of T-state Hb (Arnone, 1972). It has previously been suggested that changes at the N-termini of the β chains are responsible for the functional differences between human and bovine Hb, but there are no obvious structural differences between the human and mammoth proteins here. The mutation of Proβ5(A2) in HbA to alanine does not shift the C$^\alpha$ trace of the protein, but Hisβ2(NA2) is replaced by Asnβ/δ2, which presumably contributes to the weaker binding of organic phosphates to elephantid Hbs (Yuan *et al.*, 2011). A further example of the replacement of a proline in HbA by alanine is found at β58(E2), but this also has minimal effect on the protein structure. Compensating mutations are also found, for example the closely apposed Aspβ21(B3) and Lysβ65(E9) of HbA are replaced by Lys and Glu, respectively, in mammoth Hb. The residues forming the haem pockets of both subunit types are conserved between mammoth and human Hb, and the structures of these regions also overlap extremely well.

The allosteric mechanism of HbA involves a concerted shift in the contacts between one $\alpha\beta$ dimer and the other (Perutz *et al.*, 1998). Interactions between the α C helix and the β FG corner change substantially, while those at the β C helix and α FG corner show much more modest changes. These two regions are therefore known as the switch and flexible slide regions, respectively (Baldwin & Chothia, 1979), and interactions at these sites are largely preserved between fish and mammalian Hbs (Tame *et al.*, 1996; Yokoyama *et al.*, 2004). Since fish and mammalian lineages diverged hundreds of millions of years ago, it is unsurprising to find that mammoth Hb and HbA share the same key residues forming hydrogen bonds characteristic of the T and R states (Park *et al.*, 2006). These are listed in Table 2. Tyr$\alpha_1$42(C7) forms a hydrogen bond to Asp$\beta/\delta_2$99(G1) in the T state which is used as a marker in NMR studies to determine the quaternary state (Fung & Ho, 1975). The region around this residue overlaps extremely well with human HbA. Asp$\alpha_1$94(G1) is another highly conserved residue which hydrogen bonds to Trp$\beta_2$37(C3) in the T state, as illustrated in the electron-density map around these residues in deoxy and carbonmonoxy mammoth Hb (Fig. 2). The preservation of these core residues and their interactions shows that the allosteric mechanism is essentially the same as that of human Hb. Using the 'Tame frame' of 148 C$^\alpha$ atoms (Park *et al.*, 2006), which is similar to the BGH frame (Baldwin & Chothia, 1979), to compare different Hb molecules, we find r.m.s.d.s of 0.58 Å between mammoth and human Hb or mammoth and bovine Hb and of 0.30 Å between human and bovine Hb (Table 3), differences that are close to the expected experimental error.