Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Crystallization and preliminary X-ray analysis of a conserved hypothetical protein PAE2754 from *Pyrobaculum aerophilum* and of a double Leu→Met mutant engineered for MAD phasing

Kristina Bäckbro,‡ Annette Roos,‡ Edward N. Baker and Vickery L. Arcus*

School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand

‡ Present address: Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24 Uppsala, Sweden.

Correspondence e-mail: v.arcus@auckland.ac.nz

Received 18 December 2003
Accepted 27 January 2004

Structural genomics offers a potential route to the discovery of protein function. As part of a structural genomics project focused on the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*, a conserved hypothetical protein, PAE2754, has been expressed in *Escherichia coli*, purified and crystallized. Because of the difficulties of preparing interpretable heavy-atom derivatives with limited resolution and 8–12 molecules in the asymmetric unit, two leucine residues were selected for mutation to methionine. The double mutant L65M/L80M was created, expressed incorporating SeMet and crystallized. The crystals are monoclinic, space group $P2_1$, with unit-cell parameters a = 56.4, b = 193.3, c = 60.5 Å, β = 94.6° and eight molecules (two tetramers) in the asymmetric unit. The crystals diffract to 2.75 Å resolution and are suitable for MAD phasing.

1. Introduction

The availability of many fully sequenced genomes offers new opportunities to discover the functions of 'conserved hypothetical' proteins. These are proteins that are currently of unknown function but which are conserved across many different biological species and presumably fill important but not yet understood roles. For proteins such as these, structural genomics initiatives (Adams *et al.*, 2003; Goulding *et al.*, 2002; Stevens *et al.*, 2001; Teichmann *et al.*, 2001) offer a possible route towards the discovery of function, since protein three-dimensional structure can reveal possible catalytic or binding sites or phylogenetic relationships that are undetected at the sequence level (Lo Conte *et al.*, 2000; Teichmann *et al.*, 2001).

The work described here has its origins in a pilot structural genomics project that began by focusing on a hyperthermophilic crenarchaeon, *Pyrobaculum aerophilum*, and then developed into a major international project based on the *Mycobacterium tuberculosis* genome (The International TB Structural Genomics Consortium; http://www.doe-mbi.ucla.edu/TB; Goulding *et al.*, 2002). Both organisms colonize extreme environments, albeit of very different types. *P. aerophilum* has an optimum growth temperature of 373 K, although (unlike most other hyperthermophiles) it also has a limited tolerance for oxygen. *M. tuberculosis* is a mesophilic bacterium, but has the unusual property of being able to switch to a dormant or 'persistent' state when engulfed by macrophages; it can remain in this state, marked by low oxygen and low nutrient supply, for many years (Hatfull & Jacobs, 2000).

In a search for protein targets that might be associated with adaptation to conditions of stress, we performed a whole-genome comparison between *P. aerophilum* and *M. tuberculosis*. We identified 250 orthologues and chose a number of conserved hypothetical open reading frames from *P. aerophilum* as targets for a structural genomics approach to identifying the functions of these proteins. The open reading frame PAE2754 was one of our structural genomics targets and was of particular interest because no fewer than four homologous proteins are encoded in the genome of *P. aerophilum* (PAE0151, PAE0285, PAE0337 and PAE2754) and another four in the genome of *M. tuberculosis* (Rv0065, Rv0549, Rv0960 and Rv1720). These have subsequently all been clustered at NCBI into a single COG (cluster of orthologous groups), *i.e.* COG4113. The taxonomy of this group is relatively limited, with its protein members being derived from archaea, cyanobacteria, actinobacteria and α-proteobacteria.

Here, we describe the cloning, expression, purification and crystallization of the 133-residue protein encoded by the open reading frame PAE2754 (referred to here simply as PAE2754). A particular point of interest is that the native crystals posed difficulties for phasing the diffraction data and we therefore successfully designed, expressed and crystallized a double mutant that inserted two methionine residues into the sequence, to make possible phasing by multiwavelength anomalous diffraction (MAD) methods.

DOI: 10.1107/S0907444904002021

2. Materials and methods

2.1. Cloning, expression and purification

The predicted open reading frame for PAE2754 was amplified from genomic DNA using standard PCR protocols. The gene was subcloned into the expression vector pProEX (Life Technologies/Invitrogen) and transformed into *Escherichia coli* BL21 (DE3) cells and expressed as an N-terminal His_6-tagged protein. Following overnight induction, cells were lysed and then incubated for 40 min at 353 K to denature a large fraction of the *E. coli* proteins. The supernatant was filtered and PAE2754 was purified by Ni^{2+}-affinity chromatography using a 5 ml Hi-trap column (Amersham Biosciences). A second purification step by passage through a Superdex 200 size-exclusion chromatography column (Amersham Biosciences) gave protein as a single peak that was sufficiently pure for crystallization trials. The protein was concentrated and dynamic light-scattering data (DynaPro 200, Protein Solutions) indicated a monodisperse solution of tetrameric protein. The molecular weight calculated from the hydrodynamic radius was 76.6 kDa, compared with the theoretical tetramer molecular weight of 71.9 kDa.

2.2. Design and preparation of Leu→Met mutants

Methionine mutants of PAE2754 were made in order to solve the structure by MAD methods. Two sites for leucine-to-methionine mutations were chosen on the basis of multiple sequence alignments across COG4113, coupled with secondary-structure prediction and the prediction of amphipathic helices. The two positions chosen, at Leu65 and Leu80, were each predicted to be on the hydrophobic face of an amphipathic helix; both were almost exclusively hydrophobic across the sequences of other members of the COG and both were found to be naturally substituted by Met in at least one family member. The protein was known to contain a significant proportion of helical secondary structure by circular dichroism.

Mutations to the PAE2754 gene sequence were made using the QuikChange site-directed mutagenesis kit (Stratagene). The two single-mutant proteins L65M and L80M were first made individually and expressed as for the wild-type protein in order to check for soluble expression and thermostability. The double mutant L65M/L80M (PAE2754_MM) was then made by introducing the L80M mutation onto the background of the L65M mutant. This protein was also expressed and purified as for the wild-type protein. The double mutant was stable at 353 K, suggesting that the structure was not significantly destabilized by these mutations. The plasmid encoding the double mutant was then transformed into the methionine auxotroph DL41(DE3), which also contained a rare-codon plasmid pRP encoding rare tRNAs for arginine and proline. Selenomethionine (SeMet) was introduced into the protein by overexpression in minimal media with SeMet as the only methionine source and this SeMet-substituted protein was then purified as above.

Table 1
Data-collection statistics for native PAE2754 and SeMet-substituted PAE2754_MM.

Values in parentheses refer to the outer resolution shell.

	Native	SeMet (λ = 0.9794 Å)	SeMet (λ = 0.8321 Å)
Crystal data			
Space group	$P2_12_12_1$	$P2_1$	$P2_1$
Unit-cell parameters			
a (Å)	60.6	56.4	56.5
b (Å)	165.2	193.3	193.5
c (Å)	203.4	60.5	60.5
α (°)	90	90	90
β (°)	90	94.6	94.6
γ (°)	90	90	90
Data collection			
Resolution (Å)	50–2.50 (2.59–2.50)	40–2.75 (2.85–2.75)	40–2.75 (2.85–2.75)
Measured reflections	879180	286429	245116
Unique reflections	69531	32550	33900
Completeness (%)	97.5 (91.4)	96.1 (72.7)	91.9 (51.6)
Mosaicity	0.42	0.39	0.36
$R_{\rm merge}$† (%)	6.4 (41.8)	10.3 (31.9)	11.7 (43.6)
$I/\sigma(I)$	21.9 (3.2)	10.7 (1.9)	8.8 (1.4)

† $R_{\rm merge} = \sum |I_{\rm obs} - \langle I \rangle| / \sum I_{\rm obs}$.

2.3. Crystallization

All crystallization trials were carried out at 291 K using the hanging-drop vapour-diffusion method. The initial search for crystallization conditions made use of commercially available crystallization screens (Hampton Research) and our own in-house screens formulated using orthogonal arrays (Kingston *et al.*, 1994). The initial conditions were then optimized by fine variation of successful conditions. Crystals of wild-type PAE2754 were grown by mixing 2 µl of protein solution (8 mg ml^{-1} protein, 50 m*M* Tris–HCl pH 9.2, 50 m*M* NaCl) with 2 µl of precipitant solution (100 m*M* Tris–HCl pH 9.2, 30% 2-propanol, 18% PEG 4000, 10 m*M* trimethylamine–HCl). Crystals of the native and SeMet-substituted double mutant PAE2754_MM were obtained by slight variation of these conditions, using protein solutions of 5–10 mg ml^{-1} in 50 m*M* Tris–HCl pH 9.2, 50 m*M* NaCl and precipitant solutions comprising 100 m*M* Tris–HCl pH 9.2, 25–30% 2-propanol, 17–20% PEG 4000, 10 m*M* trimethylamine–HCl.

2.4. Data collection

All crystals were flash-frozen for data collection by soaking in cryoprotectant (mother liquor plus 10% glycerol) immediately prior to placement in a stream of cold N_2 gas (110 K). Native data for PAE2754 were collected at the National Synchrotron Light Source (NSLS), Brookhaven on beamline X8C at a wavelength of 1.0000 Å. Data for the SeMet-substituted PAE2754_MM crystals were collected at two wavelengths (λ = 0.97941 and 0.83208 Å) on beamline 9-1 at the Stanford Synchrotron Radiation Laboratory (SSRL). The two wavelengths were chosen to optimize the time required for MAD data collection (Gonzalez, 2003) and data at each wavelength were collected sequentially. In each case, the raw data were processed using *DENZO* and subsequently scaled using *SCALEPACK* (Otwinowski & Minor, 1997). Data-collection and scaling statistics are given in Table 1.

3. Results and discussion

Small crystals of PAE2754 typically grew over a period of 5 d as rectangular plates with tapered ends (Fig. 1). X-ray diffraction measurements showed that the wild-type PAE2754 crystals were orthorhombic, with unit-cell parameters *a* = 60.6, *b* = 165.2, *c* = 203.4 Å and probable space group $P2_12_12_1$. Assuming a monomer molecular weight of 17.98 kDa, this is consistent with the presence of 8–16 molecules (2–4 tetra-

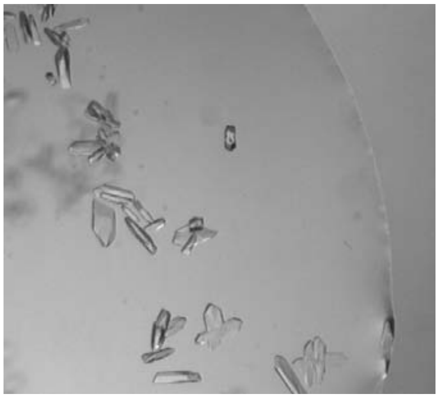

Figure 1
Crystals of native PAE2754.

mers) in the asymmetric unit, assuming a Matthews coefficient in the range 2.1–4.2 Å^3 Da^{-1}. Diffraction was seen to ~3.0 Å resolution on a home (rotating-anode) source and to 2.5 Å at NSLS, although most crystals diffracted more poorly than this. Crystals of the SeMet-substituted double mutant, on the other hand, were monoclinic, space group $P2_1$, with unit-cell parameters a = 54.4, b = 193.3, c = 60.5 Å, β = 94.6°. These crystals diffracted more poorly than the best native crystals, giving measurable X-ray data to 2.75 Å resolution at SSRL. The most likely asymmetric unit contents are eight molecules (two tetramers), which would give a Matthews coefficient of 2.64 Å^3 Da^{-1} and a solvent content of 51%.

Attempts to prepare interpretable heavy-atom derivatives of the native PAE2754 crystals by conventional soaking procedures were unsuccessful, probably because of the large number of molecules in the asymmetric unit and the relatively poor diffraction of the crystals. The only methionine in the wild-type sequence is the N-terminal residue, which was considered to be unlikely to be well ordered. We therefore decided to introduce a small number of methionine residues by site-directed mutagenesis. Analyses of substitution patterns in sequences and of the steric effects of site-specific mutations have indicated that the best candidate residue for substitution by methionine, given an appropriate environment, is leucine (Gassner & Matthews, 1999). Leucine side chains have an equivalent volume and the methionine side chain can adapt conformationally to a leucine site. On the basis of multiple sequence alignments and secondary-structure predictions we chose to mutate two leucine residues, Leu65 and Leu80, that were predicted to be on the hydrophobic face of amphipathic helices. The selection of predicted amphipathic helices was in an effort to maximize the chances that the introduced methionines would be in the hydrophobic cores of the structure and thus be well ordered. We expressed the two single mutant proteins L65M and L80M separately as well as the double mutant L65M/L80M and showed that in each case the heat stability was similar to that of the wild type.

The crystals of the SeMet-substituted double mutant belong to a different space group, although they crystallized under essentially identical conditions. An X-ray fluorescence spectrum showed a good selenium signal and data at two wavelengths were collected sequentially at SSRL (Gonzalez, 2003). Following scaling, it became apparent that the data collected at the second wavelength (λ = 0.8321 Å) was of significantly poorer quality. This is reflected in the reduction in both the completeness and $I/\sigma(I)$ in the outer resolution shell for this second data set (see Table 1) and suggests that structure solutions may need be sought by both MAD and SAD methods.

In summary, we have used site-directed mutagenesis to introduce two methionine residues into PAE2754. This has allowed us to incorporate selenomethionine into our crystals with a view to using MAD (or SAD) methods to solve the phase problem. Several factors informed our decision about where in the sequence to make the mutations. Firstly, we looked for leucine residues in the sequence, as these have been shown to be good sites for the introduction of methionine *via* mutagenesis (Gassner & Matthews, 1999). We then used a multiple sequence alignment and combined this with secondary-structure prediction to look for amphipathic helices. We reasoned that the hydrophobic face of an amphipathic helix is more likely to be buried in a hydrophobic core and thus methionines introduced at these sites would be more likely to be well ordered. Finally, we were confident that the introduced methionines had not significantly disturbed the structure as the double-mutant protein remained thermostable and crystallized under essentially the same conditions as the wild type. Efforts to solve the structure of PAE2754_MM using MAD and/or SAD methods are currently under way.

We thank Li-Wei Hung for the data collection at the National Sychrotron Light Source, Brookhaven under the auspices of the International *Mycobacterium tuberculosis* Structural Genomics Consortium. We thank Clyde Smith for MAD data collection at the SSRL. This work was supported by grants from the Marsden Fund and the Health Research Council of New Zealand.

References

Adams, M. W., Dailey, H. A., DeLucas, L. J., Luo, M., Prestegard, J. H., Rose, J. P. & Wang, B.-C. (2003). *Acc. Chem. Res.* **36**, 191–198.

Gassner, N. C. & Matthews, B. W. (1999). *Acta Cryst.* D**55**, 1967–1970.

Gonzalez, A. (2003). *Acta Cryst.* D**59**, 315–322.

Goulding, C. W. *et al.* (2002). *Curr. Drug Targets Infect. Disord.* **2**, 121–141.

Hatfull, G. F. & Jacobs, W. R. Jr (2000). Editors. *Molecular Genetics of Mycobacteria.* Washington DC: ASM Press.

Kingston, R. L., Baker, H. M. & Baker, E. N. (1994). *Acta Cryst.* D**50**, 429–440.

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). *Nucleic Acids Res.* **28**, 257–259.

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

Stevens, R. C., Yokoyama, S. & Wilson, I. A. (2001). *Science*, **294**, 89–92.

Teichmann, S. A., Murzin, A. G. & Chothia, C. (2001). *Curr. Opin. Struct. Biol.* **11**, 354–363.