Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Crystallization and preliminary crystallographic study of b0220, an 'ORFan' protein of unknown function from *Escherichia coli*

Chantal Abergel,* Vincent Monchois, Sabine Chenivesse, Sandra Jeudy and Jean-Michel Claverie

Information Génétique et Structurale, UMR1889 CNRS-AVENTIS, 31 Chemin Joseph Aiguier, 13402 Marseille CEDEX 20, France

Correspondence e-mail: chantal@igs.cnrs-mrs.fr

Received 12 October 2000
Accepted 25 October 2000

Newly sequenced microbial genomes continue to reveal up to 50% functionally uncharacterized 'anonymous' genes. A significant fraction of these anonymous ORFs does not exhibit any sequence similarity to any protein in the databases and constitutes a set of unique sequences, denoted 'ORFans'. The structure determination of ORFan proteins is both of evolutionary and functional interest. Here, the first crystallization of an *Escherichia coli* ORFan gene product, the 157 amino-acid b0220 protein, is reported. The crystals belong to the trigonal space group $P3$ or $P3_1$, with unit-cell parameters $a = b = 47.2$, $c = 88.4$ Å. There are two molecules in the asymetric unit. Frozen crystals diffract to 1.6 Å resolution using synchrotron radiation. Phasing was performed using multiwavelength anomalous dispersion (MAD) on the selenomethionine-substituted b0220 protein.

1. Introduction

Despite the accumulation of sequence information from a large diversity of species and phyla, newly sequenced bacterial genomes continue to reveal a high proportion of genes of unknown function (see, for instance, Stover *et al.*, 2000), including a significant subset of 'ORFans' (Fischer & Eisenberg, 1999), putative open reading frames (ORFs) without significant similarity to previously encountered protein (and conceptual translation) sequences. Most genes found in databases have only been predicted by computer methods and have never been experimentally validated. It is thus expected that some annotated ORFs, in particular among the ORFans, might not correspond to real genes. In a previous study, we verified the existence of cognate transcripts for 25 *E. coli* ORFans with a surprising rate of success (92%; Alimi *et al.*, 2000). Given that most ORFans appear to be transcribed, we have now initiated a systematic structure-determination program for the proteins encoded by these unique genes. Since structures evolve and change more slowly than amino-acid sequences, the knowledge of their three-dimensional structure might allow the relation of some ORFans to previously described gene families, hence providing some functional hints. Alternatively, targeting ORFans is also a suitable strategy to optimize the discovery of original protein folds, one of the goals of structural genomics.

The first *E. coli* ORFan protein for which crystals have been obtained corresponds to b0220, a 474-nucleotide ORF annotated in the *E. coli* database (ftp://ftp.genetics.wisc.edu; Blattner *et al.*, 1997). It is part of a single-gene operon and was found to exhibit a high level of expression during the exponential and stationary phases of *E. coli* growth (Alimi *et al.*, 2000). The presence of a signal peptide indicates that the protein is likely to be periplasmic. We report here the first successful crystallization to our knowledge of an *E. coli* ORFan gene product.

2. Results and discussion

2.1. Expression of the b0220 gene product

The b0220 encoding gene was PCR amplified from *E. coli* K-12 MG1655 genomic DNA and cloned into pQE-60 vector (Qiagen) to be fused with a C-terminal His_6 tag (Monchois *et al.*, in preparation). *E. coli* XL1-blue carrying the relevant plasmid pQE-0220 was cultured in LB medium supplemented with 100 µg ml^{-1} ampicillin. After initial growth at 310 K, the temperature was set to 303 K when the OD_{600} reached 0.4. Expression of b0220his protein was induced by adding 1 m*M* isopropylthio-β-D-galactoside. Cells were harvested at an OD_{600} of around 2–2.5, resuspended in buffer *A* (sodium phosphate 20 m*M* pH 8.0, 300 m*M* NaCl) containing 1.5% Triton X-100, 1.5% glycerol and total proteins were extracted by sonication. Selenomethionine derivatives were obtained by culturing XL1-blue(pQE-0220) in M9 minimal medium at 303 K supplemented with 150 mg ml^{-1} lysine, phenylalanine and threonine, 75 mg ml^{-1} leucine, isoleucine and valine, and 50 mg ml^{-1} selenomethionine prior to induction with 1 m*M* IPTG when the OD_{600}

reached 0.5. After 24 h incubation, total proteins were extracted as indicated below.

2.2. Purification

The clarified lysate was applied to a 5 ml HiTrap chelating column (Pharmacia) charged with Ni^{2+} and equilibrated with buffer *A* (50 m*M* sodium phosphate pH 8.0, 300 m*M* NaCl). The column was washed with ten column volumes of buffer *A*, ten column volumes of buffer *A* containing 25 m*M* imidazole and five column volumes of buffer *A* containing 70 m*M* imidazole at a flow rate of 1 ml min^{-1}. Elution was performed with a linear gradient over eight column volumes from 70 to 500 m*M* imidazole. The eluates were collected and the presence and purity of the b0220his in the different fractions were assessed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE). The fractions corresponding to the elution of b0220his with 150–200 m*M* imidazole were run on a desalting column (Fast Desalting Column HR 10/10, Pharmacia). The fractions contained at least 98% pure b0220his in 20 m*M* Tris buffer pH 8.0, 50 m*M* NaCl. Isoelectrofocalization revealed a single band around 7.0. The same protocol was applied to the selenomethionine-substituted b0200his and argon was applied to the fractions as they were collected. The b0220his protein produced was analyzed by mass spectroscopy and by N-terminal sequencing to confirm the selenomethionine incorporation. The sequence corresponds to the mature b0220his protein without signal peptide; mass spectroscopy confirmed the selenomethionine substitution.

2.3. Crystallization

The b0220his protein was concentrated to 5 mg ml^{-1} in 20 m*M* Tris pH 8.0, 50 m*M* NaCl using a centrifugal filter device (Ultrafree Biomax 5K, Millipore, Bedford MA, USA) under argon for the selenomethionine-substituted b0220his. Precipitation experiments were carried out on the b0220his protein using various precipitating agents (ammonium sulfate, PEG, NaCl, MPD, ethanol) at various pHs (5, 6, 7, 8, 9). The solutions of the Hampton Research Crystal Screen I and II were then used and crystallization trials were performed at 293 K by hanging-drop vapour diffusion using 24-well culture plates. Each hanging drop was prepared by mixing 0.5 µl of the 5 mg ml^{-1} b0220his with an equal volume of the reservoir solution. The hanging drop on the cover glass was vapour-equilibrated against 500 µl of the reservoir solution in each well of the tissue-culture plate. Promising results were obtained after one week from solution 28 of Crystal Screen II containing 1.6 *M* sodium citrate pH 6.5. Crystallization conditions were optimized using the *SamBA* software (Audic *et al.*, 1997) and the best crystals were obtained at 1.4 *M* sodium citrate pH 6.75, 8%(*w*/*v*) PEG 4000. Subsequently, 5% glycerol was added to the crystallization medium as a cryoprotectant agent.

2.4. Data collection and processing

One crystal was collected in a 0.5 mm^3 Hampton Research loop, flash-frozen to 105 K in a cold nitrogen-gas stream and subjected to X-ray diffraction. This data set was collected on a MAR Research imaging-plate detector at the ESRF radiation synchrotron facility (BM30) at a wavelength of 0.9794 Å. Data collection was carried out with oscillation angles of 1.0° and a crystal-to-detector distance of 225 mm. The total oscillation range collected was 120°. Space-group determination was performed using the autoindexing option in *DENZO* (Otwinowski, 1993). The crystals belong to the trigonal space group $P3$ or $P3_1$, with unit-cell parameters $a = b = 47.19$, $c = 88.41$. The packing density for two monomers of b0220 (15 kDa) in the asymmetric unit of these crystals (volume 170 556 $Å^3$) is 1.92 $Å^3$ Da^{-1}, a reasonable value for globular proteins, indicating an approximate solvent content of 36% (Matthews, 1968).

The data set was processed using the *MOSFLM* package (Leslie, 1990; Kabsch, 1993; Campbell, 1995; Steller *et al.*, 1998); the *SCALA* program from the *CCP*4 package (Collaborative Computational Project, 1994) was used for the scaling and data reduction of the native data set. The crystal diffracted to 1.58 Å and 264 737 reflections were measured in the resolution range 1.58–20.43 Å. This was reduced to a data set of 28 414 unique reflections with an R_{sym} value of 5.6. This represents a completeness of 94.1%, with a multiplicity of 3.6 and an average $I/\sigma(I)$ of 8.8. For the highest resolution shell, 9113 reflections were measured in the resolution range 1.64–1.58 Å, corresponding to 2921 unique *hkl*, an R_{sym} value of 31.5, an average $I/\sigma(I)$ of 2.4, a completeness of 94.1% and a multiplicity of 3.1.

In order to solve the b0220 structure, three wavelength data sets were collected on the selenomethionine-substituted b0220 crystals. The three data sets were collected at the ESRF synchrotron-radiation facility (BM30) in order to use the MAD method (Hendrickson *et al.*, 1990). They were processed using the *DENZO* package (Otwinowski, 1993) and programs from the *CCP*4 package (Collaborative Computational Project, 1994) were used for scaling. This work is currently in progress.

We thank Dr C. Cambillau for access to X-ray diffraction equipment and helpful discussions, Dr Michel Roth for expert assistance on the ESRF BM30 beamline and Sophie Guasco for technical assistance.

References

Alimi, J. P., Poirot, O., Lopez, F. & Claverie, J.-M. (2000). *Genome Res.* **10**, 959–966.
Audic, S., Lopez, F., Claverie, J.-M., Poirot, O. & Abergel, C. (1997). *Proteins*, **29**, 251–256.
Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). *Science*, **277**, 1453–1474.
Campbell, J. W. (1995). *J. Appl. Cryst.* **28**, 236–242.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Fischer, D. & Eisenberg, D. (1999). *Bioinformatics*, **15**, 759–762.
Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *EMBO J.* **9**, 1665–1672.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
Leslie, A. G. W. (1990). *Crystallographic Computing.* Oxford University Press.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Otwinowski, Z. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.
Steller, I., Bolotovsky, R. & Rossmann, M. (1998). *J. Appl. Cryst.* **30**, 1036–1040.
Stover, C. K, Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S. L., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., Garber, R. L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L. L., Coulter, S. N., Folger, K. R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G. K., Wu, Z. & Paulsen, I. T. (2000). *Nature (London)*, **406**, 959–964.