Title

Ab initio structure solution of a proteolytic fragment using ARCIMBOLDO

Journal

Acta Crystallographica Section F: Structural Biology Communications, 74(9)

ISSN

2053-230X

Authors

Abendroth, Jan

Sankaran, Banumathi

Myler, Peter J

et al.

Publication Date

2018-09-01

DOI

10.1107/s2053230x18010063

Peer reviewed

STRUCTURAL BIOLOGY COMMUNICATIONS

ISSN 2053-230X

Received 15 May 2018
Accepted 12 July 2018

Edited by R. Sankaranarayanan, Centre for Cellular and Molecular Biology, Hyderabad, India

Keywords: proteolytic fragments; *ARCIMBOLDO*; *ab initio* structure determination; structural genomics; infectious diseases; *Mycobacterium smegmatis*.

PDB reference: proteolytic fragment of C-terminal dimerization domain of *M. smegmatis* target, 6cum

Supporting information: this article has supporting information at journals.iucr.org/f

Ab initio structure solution of a proteolytic fragment using *ARCIMBOLDO*

Jan Abendroth,[a,b] Banumathi Sankaran,[c] Peter J. Myler,[a,d] Donald D. Lorimer[a,b] and Thomas E. Edwards[a,b]*

[a]Seattle Structural Genomics Center for Infectious Disease (SSGCID), Seattle, Washington, USA, [b]Beryllium Discovery Corporation, Bainbridge Island, WA 98110, USA, [c]Molecular Biophysics and Integrated Bioimaging, Berkeley Center for Structural Biology, Lawrence Berkeley National Laboratory, Berkeley, California, USA, and [d]Center for Infectious Disease Research, formerly Seattle Biomedical Research Institute, 307 Westlake Avenue North Suite 500, Seattle, WA 98109, USA. *Correspondence e-mail: tedwards@be4.com

Crystal structure determination requires solving the phase problem. This can be accomplished using *ab initio* direct methods for small molecules and macromolecules at resolutions higher than 1.2 Å, whereas macromolecular structure determination at lower resolution requires either molecular replacement using a homologous structure or experimental phases using a derivative such as covalent labeling (for example selenomethionine or mercury derivatization) or heavy-atom soaking (for example iodide ions). Here, a case is presented in which crystals were obtained from a 30.8 kDa protein sample and yielded a 1.6 Å resolution data set with a unit cell that could accommodate approximately 8 kDa of protein. Thus, it was unclear what had been crystallized. Molecular replacement with pieces of homologous proteins and attempts at iodide ion soaking failed to yield a solution. The crystals could not be reproduced. Sequence-independent molecular replacement using the structures available in the Protein Data Bank also failed to yield a solution. Ultimately, *ab initio* structure solution proved successful using the program *ARCIMBOLDO*, which identified two α-helical elements and yielded interpretable maps. The structure was the C-terminal dimerization domain of the intended target from *Mycobacterium smegmatis*. This structure is presented as a user-friendly test case in which an unknown protein fragment could be determined using *ARCIMBOLDO*.

1. Introduction

Small-molecule and macromolecular crystal structures obtained at resolutions higher than 1.2 Å can be determined using direct methods. However, macromolecular structures are seldom obtained at such resolution, with less than 2% of the entries in the Protein Data Bank reported at 1.2 Å resolution or better. Structures at lower resolution are generally determined using either experimental phases (Adams *et al.*, 2009) obtained using several different techniques [single-wavelength anomalous dispersion (SAD), multi-wavelength anomalous dispersion (MAD), single isomorphous replacement (SIR), multiple isomorphous replacement (MIR) and others] or molecular replacement (MR) with a homologous structure (generally >25–30% sequence identity and <1.5 Å C^{α} r.m.s.d. values; Rossmann, 1990; Scapin, 2013).

For nucleic acids, a method has been reported for the *de novo* determination of novel, folded RNA structures using secondary-structure prediction and model A-form double-stranded RNA helices as search models in molecular replacement (Robertson *et al.*, 2010; Robertson & Scott, 2008). This technique has been successful in practical applications such as the crystal structures of the L1 ligase ribozyme

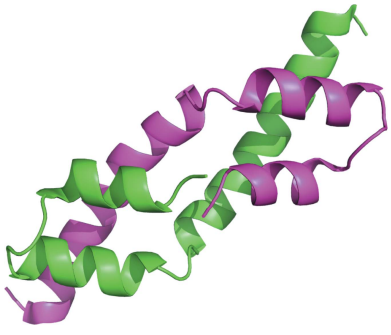


Table 1
Macromolecule-production information.

Source organism	*M. smegmatis* (strain ATCC 700084/mc^{2}155)
Gene	MSMEG_4869
DNA source	Genomic DNA
Cloning and expression vector	pAVA0421
Expression host	*E. coli* BL21(DE3) R3 Rosetta
Complete amino-acid sequence of the construct produced	MAHHHHHHMGTLEAQTQ/GPGSMPHGVVDV PELITVARGGSMRAVGRLLTLVESDRRGE VLAALGPATPRVIGVTGPPGAGKSTTVGA MVGAYRERGLRVAVLAVDPSSPYSGGALL GDRIRMAAHINDPDVLIRSMAARGHLGGL AAAVPAAIRLLAALSYDLIVLETVGVGQS EIEIAAIADPTVVILNPGAGDAVQAAKAG VLEVADLVVVNKADRDGADQTVRDLRAET DVPVLKLVAAQGDGLHELIEAIEAHQRAD TPERRRARARSQILSLAQTLLRNHADLDR LSAAVADGSSDAYTAAERLFAGSVD

Table 2
Crystallization.

Method	Sitting drop, vapor diffusion
Plate type	96-well Compact Jr, Rigaku Reagents
Temperature (K)	287
Protein concentration (mg ml^{-1})	60.56
Buffer composition of protein solution	20 m*M* HEPES pH 7.0, 300 m*M* NaCl, 5%(*v*/*v*) glycerol, 1 m*M* TCEP
Composition of reservoir solution	15%(*w*/*v*) PEG 3350, 0.1 *M* succinic acid pH 7.0
Volume and ratio of drop	0.4 µl protein solution:0.4 µl reservoir solution
Volume of reservoir (µl)	80

(Robertson & Scott, 2007) and the preQ1 riboswitch (Klein *et al.*, 2009). In an analogous approach for proteins, the program *ARCIMBOLDO* can determine crystal structures *ab initio via* the generation of model polyalanine α-helices and β-strands in combination with molecular replacement (Rodríguez *et al.*, 2009). This program has been successful in several model and novel test cases (Rodríguez *et al.*, 2009, 2012).

Crystal structures have recently been reported for several mycobacterial MeaB- and MMAA-like GTPases (Edwards *et al.*, 2015). Attempts to obtain a crystal structure of one of these proteins from *Mycobacterium smegmatis* (gene MSMEG_4869) resulted in crystals which diffracted to high resolution but with a unit cell that was far too small to contain the entire protein molecule. All attempts at molecular replacement failed and, owing to a limited number of crystals that were not reproducible, we were unable to obtain experimental phases. Ultimately, the phases were solved readily using *ab initio* structure determination in *ARCIMBOLDO* with two model polyalanine α-helices, which revealed the C-terminal dimerization domain of the intended target. Given the ease with which the structure could be determined, we present this as a model test case for crystallographers interested in sequence-independent structure determination using *ARCIMBOLDO*.

2. Materials and methods

2.1. Macromolecule production

The MSMEG_4869 gene from *M. smegmatis* encodes a 294-amino-acid protein (UniProt ID A0R1T8). The gene was cloned from genomic DNA into the pAVA0421 vector, which encodes an N-terminal hexahistidine affinity tag followed by the 3C protease cleavage sequence (Table 1). The protein was expressed and purified following standard protocols described previously (Bryan *et al.*, 2011; Choi *et al.*, 2011; Serbzhinskiy *et al.*, 2015). Briefly, the plasmid was transformed into *Escherichia coli* BL21(DE3) R3 Rosetta competent cells and the protein was expressed using auto-induction medium (Studier, 2005) in a LEX Bioreactor (Epiphyte Three Inc.). The protein was purified by immobilized metal (Ni^{2+}) affinity chromatography (IMAC), followed by cleavage of the expression tag with 3C protease, subtractive IMAC to remove the 3C protease and noncleaved protein, and size-exclusion chromatography (SEC) on a HiLoad 26/600 Superdex 75 (GE Healthcare) using a mobile phase consisting of 300 m*M* NaCl, 20 m*M* HEPES pH 7.0, 5% glycerol, 1 m*M* TCEP. The peak fractions were pooled and analysed by SDS–PAGE. The protein was concentrated to 60.56 mg ml^{-1} using an Amicon purification system (Millipore) and 200 µl aliquots were flash-frozen in liquid nitrogen and stored at −80°C until use in crystallization experiments. The protein was shown to be >95% pure by SDS–PAGE analysis. The SSGCID target identifier is MysmA.00200.a, the expression clone identifier is MysmA.00200.a.A1 and the protein batch identifier is MysmA.00200.a.A1.PS00535. The plasmid and the protein are available from SSGCID (http://www.ssgcid.org/available-materials).

2.2. Crystallization

The protein at 60.56 mg ml^{-1} was crystallized using the sitting-drop vapor-diffusion method against the JCSG+ and PACT screens (Newman *et al.*, 2005). Initially, only poorly diffracting crystals were obtained. After four months, JCSG+ screen condition G7 yielded sizable crystals (Table 2). Crystals were cryoprotected with 20%(*v*/*v*) ethylene glycol prior to vitrification in liquid nitrogen for X-ray data collection.

2.3. Data collection and processing

Data-collection and processing information is reported in Table 3. The data were integrated with *XDS* and reduced with *XSCALE* (Kabsch, 2010). The X-ray diffraction images were deposited with the Integrated Resource for Reproducibility in Macromolecular Crystallography and are available for PDB entry 6cum at https://www.proteindiffraction.org/ (Grabowski *et al.*, 2016).

2.4. Structure solution and refinement

The structure was determined *de novo* using *ARCIMBOLDO* (Rodríguez *et al.*, 2009), followed by automated model building in *ARP*/*wARP* (Langer *et al.*, 2008). The structure was refined in *PHENIX* (Adams *et al.*, 2010) with manual model building in *Coot* (Emsley & Cowtan, 2004). The quality of the structure was assessed with *MolProbity* (Headd *et al.*, 2009). Structure-solution and refinement information is

Table 3
Data collection and processing.

Values in parentheses are for the outer shell.

Diffraction source	ALS beamline 5.0.2
Wavelength (Å)	0.99990
Temperature (K)	100
Detector	ADSC Q315r CCD
Crystal-to-detector distance (mm)	176
Rotation range per image (°)	1.0
Total rotation range (°)	80
Space group	$P3_121$
a, b, c (Å)	48.53, 48.53, 54.82
α, β, γ (°)	90, 90, 120
Mosaicity (°)	0.27–0.32
Resolution range (Å)	50.00–1.60 (1.64–1.60)
Total No. of reflections	47035 (3567)
No. of unique reflections	10171 (762)
Completeness (%)	99.1 (99.7)
Multiplicity	4.6 (4.7)
Wilson B factor (Å^2)	22
$\langle I/\sigma(I)\rangle$	24.59 (2.75)
$R_{\rm merge}$	0.033 (0.468)
$CC_{1/2}$	100 (93.7)

Table 4
Structure solution and refinement.

Values in parentheses are for the outer shell.

Resolution range (Å)	50.0–1.60 (1.64–1.60)
Completeness (%)	98.7 (99.7)
No. of reflections, working set	10103 (1288)
No. of reflections, test set	979 (139)
Final $R_{\rm cryst}$	0.177 (0.243)
Final $R_{\rm free}$	0.208 (0.272)
No. of non-H atoms	
Protein	414
Water	67
Ethylene glycol	8
Total	489
R.m.s. deviations	
Bonds (Å)	0.007
Angles (°)	0.9
Average B factors (Å^2)	
Protein	30
Water	45
Ethylene glycol	41
Ramachandran plot	
Most favored (%)	100
Allowed (%)	0
PDB code	6cum

reported in Table 4. Figures were generated using *PyMOL* and *CCP*4*mg*.

3. Results and discussion

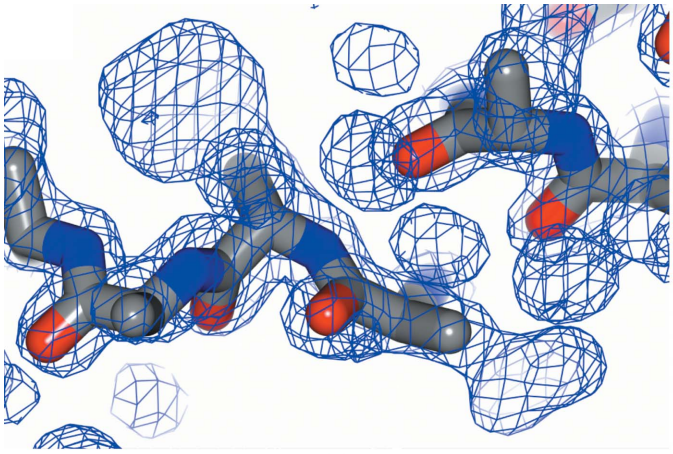
The Seattle Structural Genomics Center for Infectious Disease (SSGCID) is a structural genomics center funded by the National Institute of Allergy and Infectious Diseases (NIAID) with the mission to determine the crystal structures of potential drug targets from NIAID priority organisms (Myler *et al.*, 2009; Stacy *et al.*, 2011). One of those organisms is *M. tuberculosis* (*Mtb*), the causative agent of tuberculosis. To maximize our chances of obtaining a structure of a specific target as well as our general understanding of the target, SSGCID has adopted an ortholog approach in which closely related species within the same genus are also targeted (Baugh *et al.*, 2015; Edwards *et al.*, 2012). The *Mtb* target Rv1496 was originally misannotated as a lysine/arginine/ornithine transport system ATPase (LAO/AO ATPase), and was later shown to be an MeaB- and MMAA-like GTPase (Edwards *et al.*, 2015). In addition to *Mtb* Rv1496, the orthologs from *M. abscessus*, *M. avium*, *M. leprae*, *M. marinum*, *M. paratuberculosis*, *M. smegmatis*, *M. thermoresistible* and *M. ulcerans* were submitted to the SSGCID structure-determination pipeline. These efforts resulted in recombinant protein production of *Mtb* Rv1496 and the *M. marinum* (gene MMAR_4787), *M. smegmatis* (genes MSMEG_3160 and MSMEG_4869) and *M. thermoresistible* (gene KEK_00260) orthologs, and crystal structures of *Mtb* Rv1496 as well as the *M. smegmatis* (MSMEG_3160) and *M. thermoresistible* (KEK_00260) orthologs (Edwards *et al.*, 2015).

For the *M. smegmatis* ortholog from the MSMEG_4869 gene, we expressed and purified the 30.8 kDa recombinant protein to >95% purity as determined by SDS–PAGE analysis (data not shown). Crystals appeared after four months (Table 2). These crystals resulted in a 1.6 Å resolution data set in a trigonal point group (321) with unit-cell dimensions $a = b = 48.53$, $c = 54.82$ Å (Table 3). Curiously, this unit cell can only accommodate approximately 8 kDa of protein mass, with a projected solvent content of 47%, whereas the target *M. smegmatis* protein from the MSMEG_4869 gene is approximately 30.8 kDa and cannot fit into the asymmetric unit. Therefore, we reduced the data in lower symmetry point groups (*e.g.* *C*2) to attempt structure determination *via* molecular replacement with Rv1496 or orthologs; this approach failed. We theorized that the crystals could contain the C-terminal dimerization domain (three α-helices), yet attempts to solve the structure by MR using the C-terminal dimerization domains of Rv1496 or its orthologs also failed. This was not too surprising since the protein sequence from the MSMEG_4869 gene deviates significantly in its primary sequence for this region relative to orthologs (<20% sequence identity). Only a few crystals were obtained, and thus attempts to determine experimental phases using soaking experiments were limited to iodide ion soaking using a previously developed protocol (Abendroth *et al.*, 2011). This approach also failed, causing us to question whether these crystals were indeed the correct target or something else entirely. In addition to specific model-based molecular-replacement searches in *Phaser* (McCoy *et al.*, 2007) and *MOLREP* (Vagin & Teplyakov, 2010), PDB-wide molecular MR using *MoRDa* as well as sequence-independent MR using *SIMBAD* (Simpkin *et al.*, 2018), both of which are from the *CCP*4 suite (Winn *et al.*, 2010), also failed to produce a solution.

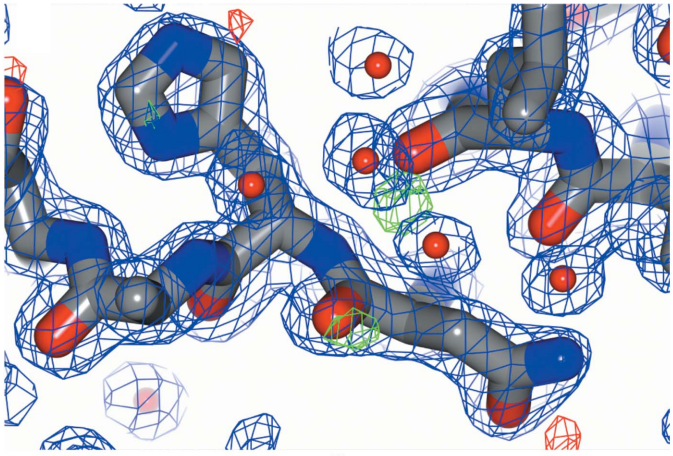
Use of the program *ARCIMBOLDO* has been described for *ab initio* structure determination using model polyalanine α-helices and β-strands for data sets below atomic resolution (Rodríguez *et al.*, 2009, 2012). *ARCIMBOLDO* combines the location of model fragments with *Phaser* (Read & McCoy, 2016) with density modification and autotracing with *SHELXE* (Usón & Sheldrick, 2018) in a multisolution frame. This approach has been successful for test cases as well as for a

novel protein (Rodríguez *et al.*, 2009), for DNA-binding proteins (Pröpper *et al.*, 2014) and for coiled coils (Caballero *et al.*, 2018). Because this approach appears to work well at <2 Å resolution for proteins with a significant α-helical content, it was applied to the 1.6 Å resolution data set. As a starting point, we searched for two helices with ten residues each. Within the *ARCIMBOLDO_LITE* pipeline (Sammito *et al.*, 2015), *Phaser MR* (McCoy *et al.*, 2007) found a solution from two fragments with LLGs of 57 and 185, respectively, and *Z*-scores of 9.3 and 15.6, respectively, within 30 min using a IntelCore i7 3.6 GHz CPU. The model then was extended to 51 amino acids consisting of three connected helices using *SHELXE* (Usón & Sheldrick, 2018) as part of the *ARCIMBOLDO* pipeline (Fig. 1*a*). The polyalanine model from *ARCIMBOLDO* was used for MR in *Phaser* (McCoy *et al.*, 2007), and the density was improved with *Parrot* (Cowtan, 2010) to yield improved electron-density maps. An initial sequence was postulated based on the electron density. A sequence search against the SSGCID targets revealed that the sequence matched the C-terminal region of the intended target (MysmA.00200.a). The model was built with *ARP/wARP* (Langer *et al.*, 2008) using the correct sequence. The final model was refined with *PHENIX* and spans ordered residues Asp240–Ala290 of the 294-amino-acid protein (Fig. 1*b*). Thus, the crystals contain approximately 6 kDa protein and 61% solvent content, with a Matthews coefficient of 3.1 Å^3 Da^{-1}, which yields large solvent channels (Fig. 1*c*). Several proteolysis sites could result in this 6 kDa fragment. For example, 3C protease was used to remove the N-terminal affinity tag, and aberrant cleavage could have occurred at residue Gln237. Alternatively, cleavage could have occurred at Arg238 by the trypsin-like protease II of *E. coli* (Strongin *et al.*, 1979) if this protease were retained in tiny amounts after purification.

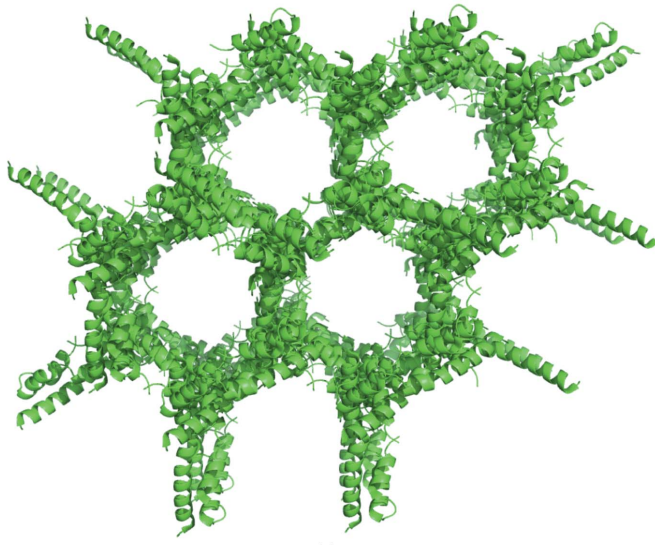
Although the asymmetric unit only contains one copy of the proteolytic fragment, the protein crystallized as a dimer *via* crystallographic symmetry. In fact, the C-terminal domain dimerized in a similar fashion to other mycobacterial MeaB- and MMAA-like GTPases (Fig. 2). N-terminal to the first ordered residue of the dimerization domain is a short linker at residue Arg238 followed by the last α-helix of the GTPase domain, *via* analogy to other family members. As described above, Gln237 and Arg238 are possible proteolytic sites for the fragment crystallized here. The C-terminal dimerization domain of MysmA.00200.a contains different length α-helices to those of the related proteins MysmA.00200.b (MSMEG_3160 gene) and MytuD.00200.a (*Mtb* Rv1496), reflecting their low sequence identity (<20%). As a result, the C$^\alpha$ r.m.s.d. values for these protein domains are >5 Å. Using the *PDBeFOLD* secondary-structure matching server (Krissinel & Henrick, 2004), the nearest structural homologs are all dimerization domains from human deoxynucleotidyltransferase terminal-interacting protein 1 (DNTTIP1, 47 C$^\alpha$ atoms; r.m.s.d. 2.69 Å; PDB entry 4d6k; Itoh *et al.*, 2015), *Zika virus* (39 C$^\alpha$ atoms; r.m.s.d. 2.92 Å; PDB entry 5h37; Zhang *et al.*, 2016), *Japanese encephalitis virus* (40 C$^\alpha$ atoms; r.m.s.d. 3.24 Å; PDB entry 5ywp; Qiu *et al.*, 2018) and *Dengue virus* (42 C$^\alpha$ atoms; r.m.s.d. 3.25 Å; PDB entry 3j27; Zhang *et al.*,



(*a*)



(*b*)



(*c*)

Figure 1
(*a*) Initial polyalanine model and electron-density maps obtained after *ab initio* structure determination in *ARCIMBOLDO*. (*b*) Final refined model of MysmA.00200.a (MSMEG_4869 gene) and electron-density maps. (*c*) Crystal lattice, showing large solvent channels.

2013). Thus, it is not surprising that the structure obtained here crystallized as a dimer.

Analysis using the *PSIPRED* server (Buchan *et al.*, 2013; McGuffin *et al.*, 2000) revealed that the C-terminal domain of MysmA.00200.a should contain three helices spanning residues Pro242–Arg262, Leu267–Ala275 and Ala28–Phe289. The annotated PDB entry of this structure contains three helices spanning residues Thr241–Asn263, Asp266–Asp276 and Asp280–Leu288, which closely mirror those predicted by *PSIPRED*. To better understand the performance of *ARCIMBOLDO*, we performed a retrospective analysis in which we ran additional searches with (i) one 22-residue helix, (ii) three ten-residue helices and (iii) one ten-residue helix. All three searches yielded virtually identical models with 50–51 residues and C^{α} r.m.s.d. values between 0.4 and 0.6 Å, in which most of the difference can be attributed to the orientation of the C-terminal residue. In an additional retrospective analysis, the same set of searches was performed with *Fragon* (Jenkins, 2018), which uses a similar approach to *ARCIMBOLDO* with a different set of programs. *Fragon* uses *Phaser* (McCoy *et al.*, 2007) to place idealized secondary-structure elements; phases are then improved using *ACORN* (Yao *et al.*, 2006), followed by model building with *ARP*/*wARP* (Langer *et al.*, 2008). Searches starting from one, two and three ten-residue helices each yielded a complete model.

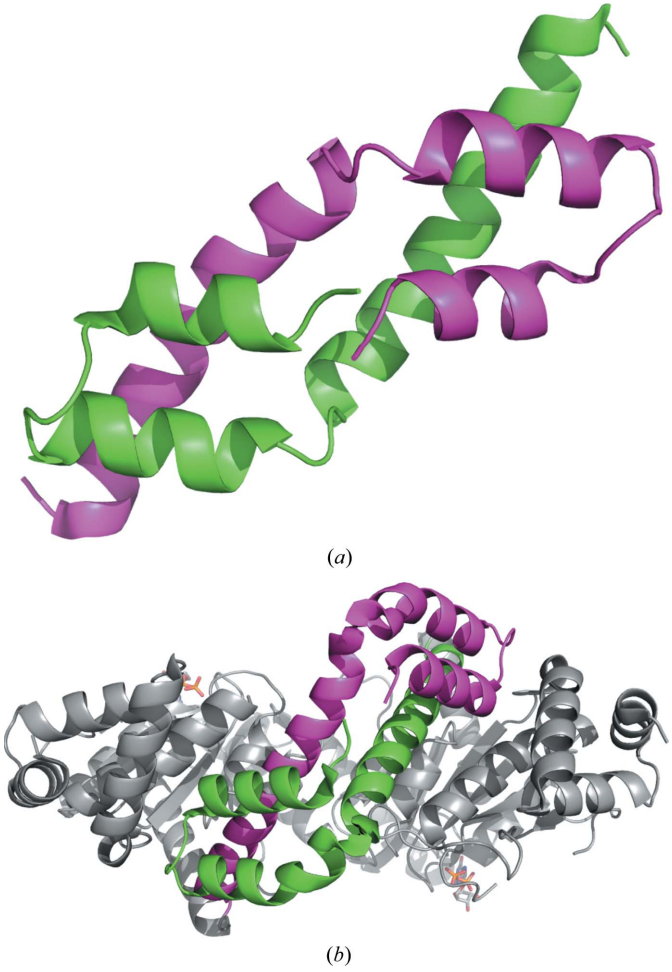


Figure 2
(*a*) Crystal structure of the C-terminal dimerization domain of the *M. smegmatis* protein from the MSMEG_4869 gene, showing the biologically relevant dimer. The asymmetric unit contains a single chain (shown in green ribbons); the second protomer of the dimer (magenta ribbons) is generated *via* crystallographic symmetry. (*b*) Crystal structure of the ortholog from *M. thermoresistibile* (PDB entry 3tk1), which contained a dimer in the asymmetric unit. The GTPase domains are shown as gray ribbons with GDP shown in stick representation. The dimerization domain is shown with one molecule in green ribbons and the other molecule in magenta ribbons.

4. Conclusions

Despite not knowing exactly what was present in our 1.6 Å resolution data set, we determined the structure of a proteolytic fragment of the C-terminal dimerization domain of the intended *M. smegmatis* target using the *ab initio* structure-determination program *ARCIMBOLDO*. This example serves as a robust test case for individuals interested in learning to use *ARCIMBOLDO* or complementary software such as *Fragon*.

Acknowledgements

We thank the SSGCID cloning and protein production groups at the Center for Infectious Disease Research.

Funding information

This work was supported by National Institutes of Health/National Institute of Allergy and Infectious Diseases (contract Nos. HHSN272201700059C, HHSN272201200025C and HHSN272200700057C to PJM). The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health, National Institute of General Medical Sciences and the Howard Hughes Medical Institute. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences of the US Department of Energy under Contract No. DE-AC02-05CH11231.

References

Abendroth, J., Gardberg, A. S., Robinson, J. I., Christensen, J. S., Staker, B. L., Myler, P. J., Stewart, L. J. & Edwards, T. E. (2011). *J. Struct. Funct. Genomics*, **12**, 83–95.
Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Adams, P. D., Afonine, P. V., Grosse-Kunstleve, R. W., Read, R. J., Richardson, J. S., Richardson, D. C. & Terwilliger, T. C. (2009). *Curr. Opin. Struct. Biol.* **19**, 566–572.
Baugh, L. *et al.* (2015). *Tuberculosis*, **95**, 142–148.
Bryan, C. M., Bhandari, J., Napuli, A. J., Leibly, D. J., Choi, R., Kelley, A., Van Voorhis, W. C., Edwards, T. E. & Stewart, L. J. (2011). *Acta Cryst.* F**67**, 1010–1014.
Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. (2013). *Nucleic Acids Res.* **41**, W349–W357.
Caballero, I., Sammito, M., Millán, C., Lebedev, A., Soler, N. & Usón, I. (2018). *Acta Cryst.* D**74**, 194–204.
Choi, R., Kelley, A., Leibly, D., Nakazawa Hewitt, S., Napuli, A. & Van Voorhis, W. (2011). *Acta Cryst.* F**67**, 998–1005.
Cowtan, K. (2010). *Acta Cryst.* D**66**, 470–478.
Edwards, T. E., Baugh, L., Bullen, J., Baydo, R. O., Witte, P., Thompkins, K., Phan, I. Q., Abendroth, J., Clifton, M. C., Sankaran,

B., Van Voorhis, W. C., Myler, P. J., Staker, B. L., Grundner, C. & Lorimer, D. D. (2015). *J. Struct. Funct. Genomics*, **16**, 91–99.
Edwards, T. E., Liao, R., Phan, I., Myler, P. J. & Grundner, C. (2012). *Protein Sci.* **21**, 1093–1096.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Grabowski, M., Langner, K. M., Cymborowski, M., Porebski, P. J., Sroka, P., Zheng, H., Cooper, D. R., Zimmerman, M. D., Elsliger, M.-A., Burley, S. K. & Minor, W. (2016). *Acta Cryst.* D**72**, 1181–1193.
Headd, J. J., Immormino, R. M., Keedy, D. A., Emsley, P., Richardson, D. C. & Richardson, J. S. (2009). *J. Struct. Funct. Genomics*, **10**, 83–93.
Itoh, T., Fairall, L., Muskett, F. W., Milano, C. P., Watson, P. J., Arnaudo, N., Saleh, A., Millard, C. J., El-Mezgueldi, M., Martino, F. & Schwabe, J. W. (2015). *Nucleic Acids Res.* **43**, 2033–2044.
Jenkins, H. T. (2018). *Acta Cryst.* D**74**, 205–214.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Klein, D. J., Edwards, T. E. & Ferré-D'Amaré, A. R. (2009). *Nature Struct. Mol. Biol.* **16**, 343–344.
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.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). *Bioinformatics*, **16**, 404–405.
Myler, P. J., Stacy, R., Stewart, L., Staker, B. L., Van Voorhis, W. C., Varani, G. & Buchko, G. W. (2009). *Infect. Disord. Drug Targets*, **9**, 493–506.
Newman, J., Egan, D., Walter, T. S., Meged, R., Berry, I., Ben Jelloul, M., Sussman, J. L., Stuart, D. I. & Perrakis, A. (2005). *Acta Cryst.* D**61**, 1426–1431.
Pröpper, K., Meindl, K., Sammito, M., Dittrich, B., Sheldrick, G. M., Pohl, E. & Usón, I. (2014). *Acta Cryst.* D**70**, 1743–1757.
Qiu, X. *et al.* (2018). *Nature Microbiol.* **3**, 287–294.
Read, R. J. & McCoy, A. J. (2016). *Acta Cryst.* D**72**, 375–387.
Robertson, M. P., Chi, Y.-I. & Scott, W. G. (2010). *Methods*, **52**, 168–172.
Robertson, M. P. & Scott, W. G. (2007). *Science*, **315**, 1549–1553.
Robertson, M. P. & Scott, W. G. (2008). *Acta Cryst.* D**64**, 738–744.
Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nature Methods*, **6**, 651–653.
Rodríguez, D., Sammito, M., Meindl, K., de Ilarduya, I. M., Potratz, M., Sheldrick, G. M. & Usón, I. (2012). *Acta Cryst.* D**68**, 336–343.
Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.
Sammito, M., Millán, C., Frieske, D., Rodríguez-Freire, E., Borges, R. J. & Usón, I. (2015). *Acta Cryst.* D**71**, 1921–1930.
Scapin, G. (2013). *Acta Cryst.* D**69**, 2266–2275.
Serbzhinskiy, D. A., Clifton, M. C., Sankaran, B., Staker, B. L., Edwards, T. E. & Myler, P. J. (2015). *Acta Cryst.* F**71**, 594–599.
Simpkin, A. J. *et al.* (2018). *Acta Cryst.* D**74**, 595–605.
Stacy, R., Begley, D. W., Phan, I., Staker, B. L., Van Voorhis, W. C., Varani, G., Buchko, G. W., Stewart, L. J. & Myler, P. J. (2011). *Acta Cryst.* F**67**, 979–984.
Strongin, A. Y., Gorodetsky, D. I. & Stepanov, V. M. (1979). *J. Gen. Microbiol.* **110**, 443–451.
Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.
Usón, I. & Sheldrick, G. M. (2018). *Acta Cryst.* D**74**, 106–116.
Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Yao, J. X., Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006). *Acta Cryst.* D**62**, 901–908.
Zhang, S., Kostyuchenko, V. A., Ng, T.-S., Lim, X.-N., Ooi, J. S. G., Lambert, S., Tan, T. Y., Widman, D. G., Shi, J., Baric, R. S. & Lok, S. M. (2016). *Nature Commun.* **7**, 13679.
Zhang, X., Ge, P., Yu, X., Brannan, J. M., Bi, G., Zhang, Q., Schein, S. & Zhou, Z. H. (2013). *Nature Struct. Mol. Biol.* **20**, 105–110.c