Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

Crystallization and preliminary X-ray crystallographic analysis of Aart, a designed six-finger zinc-finger peptide, bound to DNA

Justin W. Crotty,[a] Christopher Etzkorn,[b] Carlos F. Barbas III,[c] David J. Segal[d] and Nancy C. Horton[e]*

[a]Department of Chemistry, University of Arizona, Tucson, AZ 85721, USA, [b]Albert Einstein College of Medicine, New York, NY, USA, [c]Department of Molecular Biology, Scripps Research Institute, La Jolla, CA 92037, USA, [d]Department of Pharmacology and Toxicology, University of Arizona, Tucson, AZ 85721, USA, and [e]Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 85721, USA

Correspondence e-mail: nhorton@u.arizona.edu

Received 12 March 2005
Accepted 5 May 2005
Online 1 June 2005

Crystals of a designed six-finger zinc-finger protein, Aart, bound to a 22-base-pair duplex DNA containing a consensus binding site have been obtained. Crystals grew by hanging-drop vapor diffusion from solutions containing polyethylene glycol 4000 as the precipitating agent. The irregularly shaped crystals belong to space group $P1$, with unit-cell parameters $a = 41.95$, $b = 71.76$, $c = 74.73$ Å, $\alpha = 100.87$, $\beta = 96.22$, $\gamma = 106.33^\circ$. There are most likely to be two protein–DNA complexes in the asymmetric unit. A complete native data set has been collected from a high-energy synchrotron source to a resolution of 2.95 Å at 100 K, with an $R_{\rm merge}$ of 9.3%.

1. Introduction

Zinc fingers are protein domains of approximately 30 residues composed of an α-helix and two β-strands stabilized in part by the ligation of a Zn^{2+} ion. These domains are found in all kingdoms of life and are one of the most common domains identifiable by sequence in the human genome. Zinc fingers mediate interactions with several classes of biological macromolecules, including DNA, protein and lipids. There are several different classes of zinc fingers and of these the Cys2-His2 zinc finger is the best characterized as a DNA-binding domain. Each finger typically recognizes 3–4 base pairs of DNA. Therefore, extended DNA sequences can be recognized by tandem arrays of zinc fingers.

The seemingly simple relationship between zinc-finger sequence and DNA-target sequence has inspired several attempts to design or select zinc fingers that can bind to any desired DNA sequence. These largely successful efforts have led to the production of custom zinc-finger peptides, which are currently being used for a number of medical and biochemical applications (Blancafort *et al.*, 2004; Corbi *et al.*, 2004; Jamieson *et al.*, 2003). However, these studies have also revealed several unanticipated difficulties in the engineering of custom zinc-finger peptides, including the design of zinc-finger sequences appropriate for recognizing triplet nucleotides beginning with C, T or A and the optimal linker length between fingers.

One of the most common and successful methods for zinc-finger design utilizes selection *via* phage-display techniques. While zinc fingers recognizing triplets beginning with G nucleotides can be obtained by these methods fairly readily, specific recognition of C, T or A as the first nucleotide in the triplet has been more challenging (Dreier *et al.*, 2001). Furthermore, there are very few examples of structural studies with zinc-finger proteins that are capable of achieving such interactions (Wolfe *et al.*, 2001).

The composition of the linker between fingers is known to have an effect on DNA-binding affinity (Kim & Pabo, 1998; Moore, Choo *et al.*, 2001; Moore, Klug *et al.*, 2001). The five-amino-acid 'canonical' linker TGEKP is found between roughly 50% of all naturally occurring zinc-finger domains (Jacobs, 1992). This high degree of conservation suggests that there are functional roles for all amino acids. However, recent observations by several groups suggest that the use of this linker results in decreased affinity, especially in arrays of four or more domains (Kim & Pabo, 1998; Laity *et al.*, 2000; Liu *et al.*, 1997; Moore, Choo *et al.*, 2001; Nagaoka, Kaji *et al.*, 2001; Nagaoka, Nomura *et al.*, 2001). Unique recognition of a particular DNA sequence in the human genome might be useful for medical

applications in humans, such as gene therapy. This would require the recognition of 16–18 nucleotides, which could be accomplished with a protein containing six zinc fingers. It is hypothesized that the canonical linker positions the contacting residues of adjacent fingers slightly out of register with the DNA bases. One consequence of this awkward alignment is that the protein exhibits lower affinity because binding energy is consumed contorting the DNA or is lost owing to missing DNA contacts. Although many successful six-zinc-finger artificial transcription factors with canonical linkers have been reported (Beerli *et al.*, 2000; Guan *et al.*, 2002; Blancafort *et al.*, 2003, Segal *et al.*, 2004), several studies have found that using longer linkers in various arrangements can produce proteins of higher affinity (Moore, Klug *et al.*, 2001).

In order to address these issues, we have crystallized a designed six-finger zinc-finger protein referred to as Aart (Fig. 1, molecular weight 21 439.85 Da). Aart was designed and constructed (Dreier *et al.*, 2001) based on the use of zinc-finger domains of predetermined specificity to bind the sequence 5′-ATG TAG AGA AAA ACC AGG-3′. However, the DNA sequence preferred by Aart was determined (Segal *et al.*, 2003) by cyclic amplification and selection of targets (CAST) analysis to be (the consensus site) 5′-ATG TAG GGA AAA GCC CGG G-3′. Note that two triplets begin with A and one each begins with C or T. We have obtained usable crystals of Aart bound to DNA containing the consensus site.

The structure of Aart bound to the consensus DNA is expected to shed light on the issues of optimal linker length and sequence specificity for triplets beginning with C, T or A. These results will be applied towards our long-term goal of trying to derive a comprehensive set of recognition rules that should enable improved design of custom zinc-finger proteins, as well as potentially being able to predict the binding site of uncharacterized natural proteins. These rules would certainly involve the direct contacts between side chains and nucleotides, including the 'non-modular' contacts where sequence specificity at one triplet is aided by contacts from an adjacent finger. Importantly, the rules will most likely also involve additional components such as the sequence-dependent energetics of DNA structure, as well as contacts within the protein and between the protein and DNA that position the finger relative to the DNA.

2. Purification, crystallization and diffraction data collection

The Aart sequence (Fig. 1) was designed and constructed based on the use of zinc-finger domains of predetermined specificity with the canonical TGEPK linker between adjacent fingers (Dreier *et al.*, 2001). Aart protein was expressed in *Escherichia coli* as a C-terminal fusion to the maltose-binding protein (MBP purification system, NEB). The fusion protein contains a factor Xa protease site for cleavage of the MBP tag. A total of 16 l cell culture [*E. coli* strain BL21 Star (DE3), Invitrogen] expressing the MBP-Aart fusion protein was prepared by inoculating 200 ml culture grown overnight in LB media containing 50 µg ml^{-1} ampicillin and 90 µ*M* $ZnCl_2$ into 1 l LB containing 50 µg ml^{-1} ampicillin and 90 µ*M* $ZnCl_2$. The cultures were grown at 310 K until they reached an OD of 0.7 measured at 600 nm (each culture was monitored independently) and expression was induced by adding 0.3 m*M* IPTG (final concentration). The cultures were then grown for an additional 21 h at 310 K and harvested by centrifugation at 5200 rev min^{-1} on a Sorvall HG-4L rotor for 25 min. The cell pellets were stored at 193 K until needed.

```
1   |   10    |    20     |   30     |   40
ISEFGSSSSVAQAALE PGEKP YACPECGKSFSR SDHLAEHQRTH
                 |    50    |   60     |   70
                 TGEKP YKCPECGKSFSD KKDLTRHQRTH
                   |    80    |    90    |  100
                 TGEKP YKCPECGKSFSQ RANLRAHQRTH
                      |   110    |   120    |
                 TGEKP YACPECGKSFSQ LAHLRAHQRTH
                130     |  140    |   150    |
                 TGEKP YKCPECGKSFSR EDNLHTHQRTH
                  160     |  170     |  180     |  190
                 TGEKP YKCPECGKSFSR RDALNVHQRTH TGKKTS
                 ______ __________ -1 123456 ______
                 linker    β-sheet     α-helix
```

Figure 1
Aart sequence.

A cell pellet from 4 l growth culture was resuspended in an adequate volume (3 ml per gram of wet cells; 60 ml in this case) of cold ZBA (20 m*M* Tris pH 7.5, 90 µ*M* $ZnCl_2$, 1 m*M* $MgCl_2$, 90 m*M* KCl) and sonicated at 277 K with 50% output for 20 1 s pulses. A 20 µl sample was removed, spun to remove cell debris and the absorbance at 280 nm measured. Sonication continued in this way until the absorbance at 280 nm remained constant. A total of 20 sonication sessions was necessary to completely disrupt the cells. The cell lysate was centrifuged for 60 min at 12 500*g* and then filtered using a 0.8 µm filter followed by an additional filtration using a 0.2 µm filter. The filtrate was then diluted 1:1 with ZBA and loaded onto a 115 ml amylose resin (New England Biolabs) column. After washing the column with ten column volumes of ZBA, the protein was eluted with 50 ml 10 m*M* maltose. A single peak eluted and the pooled fractions were concentrated in a YM-10 Centricon (Amicon Inc.). The concentrated fraction pool was diluted with cleavage buffer (20 m*M* Tris–HCl pH 8.0, 20 m*M* NaCl, 2 m*M* $CaCl_2$, 90 µ*M* $ZnCl_2$) and factor Xa protease was added at a concentration of 0.001 mg per millilitre of fraction volume and agitated overnight at room temperature. The cleavage of the MBP tag from Aart was verified by SDS–PAGE and the cleavage reaction was diluted 1:4 in ZBA and loaded onto a heparin column pre-equilibrated in ZBA.

Protein was eluted with a 2 *M* KCl gradient and fractions containing Aart were identified by SDS–PAGE. Aart was further purified from co-purifying DNA by passage over a MonoQ column (Amersham Inc.) equilibrated in ZBA. The flowthrough was dialyzed into crystallization buffer (10 m*M* HEPES pH 7.5, 400 m*M* NaCl, 5 m*M* DTT, 90 µ*M* $ZnCl_2$, 1 m*M* $MgCl_2$) containing 50%(*w*/*v*) glycerol, aliquoted, flash-frozen in liquid nitrogen and stored at 193 K. Prior to crystallization, protein was dialyzed extensively against crystallization buffer. The small amount of precipitate that formed was removed by centrifugation and the protein was concentrated to 23 mg ml^{-1} using a YM-10 Centricon (Amicon Inc.).

Unmodified as well as 5-bromouracil-containing DNA was obtained from commercial sources of synthetic oligonucleotides purified by reverse-phase HPLC (Aggarwal, 1990). Annealing of

```
22       C AG ATG TAG GGA AAA GCC CGG G
              TC TAC ATC CCT TTT CGG GCC C G

23       AGG ATG TAG GGA AAA GCC CGG G CC
           CC TAC ATC CCT TTT CGG GCC C GG T

25     G AGG ATG TAG GGA AAA GCC CGG G CCT
          TCC TAC ATC CCT TTT CGG GCC C GGA C

27    TG AGG ATG TAG GGA AAA GCC CGG G CCT C
       C TCC TAC ATC CCT TTT CGG GCC C GGA G A
```

Figure 2
Sequences screened for cocrystallization.

complementary strands was performed in sterile water or TE (10 m*M* Tris–HCl, 1 m*M* EDTA pH 8) by heating to 368 K in a heating block and slowly cooling to room temperature overnight. Aart was mixed with the various oligonucleotides shown in Fig. 2 at a 1:2 molar ratio in buffer containing 50 m*M* HEPES pH 7.5, 40 µ*M* zinc acetate, 1 m*M* DTT and 400–600 m*M* NaCl. Further concentration of the protein–DNA complex was often necessary to bring the concentration of Aart to within 10–20 mg ml^{-1}. Crystallization conditions were screened using the hanging-drop vapor-diffusion method with solutions containing 5–25%(*w*/*v*) PEG 4000, 0.15 *M* NaCl and 0.1 *M* buffers of varying pH (sodium acetate pH 4.5, sodium citrate pH 5.5, imidazole pH 6.5, HEPES pH 7.5, Tris pH 8.5) at 290 K. Drops composed of 1.5 µl Aart–DNA solution and 1.5 µl crystallization solution were placed over a 1 ml well of crystallization solution. Crystalline material appeared in the drops within one week. The most well formed crystals were obtained with the 22-base-pair oligonucleotide (Fig. 3, molecular weight 13 142.6 Da). The best crystallization conditions contained 50 m*M* citrate pH 5.5, 100 m*M* ammonium acetate, 200 m*M* NaCl and 22.5%(*w*/*v*) PEG 4000 in the well compartment of the crystallization tray. The drop solution contained 1.5 µl Aart–DNA (in 50 m*M* HEPES pH 7.5, 40 µ*M* zinc acetate, 1 m*M* DTT, 400 m*M* NaCl) and 1.5 µl crystallization solution without NaCl [50 m*M* citrate pH 5.5, 100 m*M* ammonium acetate and 22.5%(*w*/*v*) PEG 4000]. NaCl was excluded from the crystallization solution added to the drop in order to compensate for the relatively high NaCl concentration (400 m*M*) necessary to keep the Aart–DNA complex soluble. Crystals formed in 1 or 2 d and grew to full size between one and two weeks. These crystals did not grow without the presence of both the Aart protein and the DNA. Crystals grown in the presence of DNA containing 3–6 substitutions of 5-bromouracil at the thymine positions show very strong fluorescence signals at both the bromine (0.92 Å) as well as the zinc (1.28 Å) absorbance edges (data not shown), indicating the presence of both Aart protein and DNA within the crystals.

Crystals of Aart–DNA were exchanged into a cryoprotectant consisting of 25%(*w*/*v*) PEG 4000, 0.3 *M* NaCl, 0.05 *M* sodium citrate pH 5.5, 40 µ*M* zinc acetate and 30%(*w*/*v*) glycerol. A crystal was flash-frozen in liquid nitrogen and mounted at the Argonne National Laboratory Advanced Photon Source BL-14BMC. The native data set was found to be 97% complete to 2.95 Å (Table 1). The calculated Matthews coefficient of 2.91 Å^3 Da^{-1} (with a solvent content of 58%) suggests the presence of two Aart–DNA complexes in the asymmetric unit. The self-rotation function shows large (6–8σ) peaks in the $\kappa = 90^\circ$ and $\kappa = 180^\circ$ sections, indicating the presence of non-crystallographic symmetry. Fluorescence scans of 5-bromouracil-containing crystals were performed at Stanford Synchrotron Light Source BL 9-1 (bromine) and Argonne National Laboratory Advanced Photon Source BL-19BM (bromine, zinc).

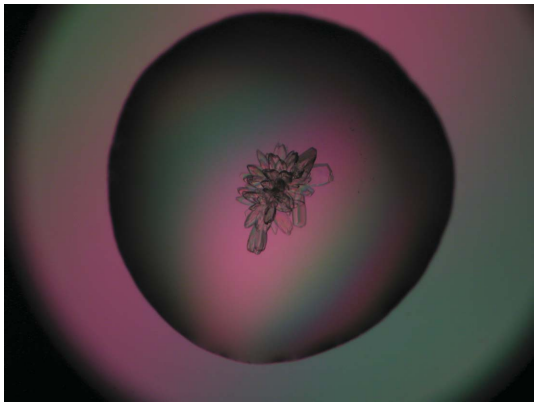

Figure 3
Crystals of Aart complexed with a 22-base-pair oligomer of DNA.

Table 1
Diffraction data summary.

Values in parentheses are for the highest resolution shell.

X-ray source	APS BL-14BMC
Wavelength (Å)	0.9
Detector	ADSC
Temperature (K)	100
Camera distance (mm)	250
Total oscillation angle (°)	180
Rotation angle per frame (°)	1
Space group	*P*1
Unit-cell parameters (Å, °)	a = 41.95, b = 71.76, c = 74.73, α = 100.87, β = 96.22, γ = 106.33
Resolution range (Å)	100–2.95
$R_{\rm merge}$† (%)	9.3 (54.9)
$\langle I/\sigma(I)\rangle$	13.6 (2.2)
Total reflections	30382
Unique reflections	15754
Completeness (%)	96.9 (98.2)

† $R_{\rm merge} = \sum_{hkl} |\langle I_{hkl}\rangle - I_{hkl}| / \sum_{hkl} I_{hkl}$, where $\langle I_{hkl}\rangle$ is the average intensity over symmetry-related and equivalent reflections and I_{hkl} is the observed intensity for reflection *hkl*.

We thank Dr Axel Warsinke for his assistance in the development of the protein-purification protocol. CFB and AW were supported by The Skaggs Institute for Chemical Biology and CA86258. JWC was supported in part by NSF IGERT 9870659. Portions of this research were carried out at the Advanced Photon Source, a national user facility operated by Argonne National Laboratory, as well as at Stanford Synchrotron Radiation Laboratory, a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program and the National Institute of General Medical Sciences.

References

Aggarwal, A. K. (1990). *Methods*, **1**, 83–90.
Beerli, R. R., Dreier, B. & Barbas, C. F. III (2000). *Proc. Natl Acad. Sci. USA*, **97**, 1495–1500.
Blancafort, P., Magnenat, L. & Barbas, C. F. III (2003). *Nature Biotechnol.* **21**, 269–274.
Blancafort, P., Segal, D. J. & Barbas, C. F. III (2004). *Mol. Pharmacol.* **66**, 1361–1371.
Corbi, N., Libri, V., Onori, A. & Passananti, C. (2004). *Biochem. Cell Biol.* **82**, 428–436.
Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D. & Barbas, C. F. III (2001). *J. Biol. Chem.* **276**, 29466–29478.
Guan, X., Stege, J., Kim, M., Dahmani, Z., Fan, N., Heifetz, P., Barbas, C. F. III & Briggs, S. P. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 13296–13301.
Jacobs, G. H. (1992). *EMBO J.* **11**, 4507–4517.
Jamieson, A. C., Miller, J. C. & Pabo, C. O. (2003). *Nature Rev. Drug Discov.* **2**, 361–368.
Kim, J. S. & Pabo, C. O. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 2812–2817.
Laity, J. H., Dyson, H. J. & Wright, P. E. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 11932–11935.
Liu, Q., Segal, D. J., Ghiara, J. B. & Barbas, C. F. III (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5525–5530.
Moore, M., Choo, Y. & Klug, A. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 1432–1436.
Moore, M., Klug, A. & Choo, Y. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 1437–1441.

Nagaoka, M., Kaji, T., Imanishi, M., Hori, Y., Nomura, W. & Sugiura, Y. (2001). *Biochemistry*, **40**, 2932–2941.
Nagaoka, M., Nomura, W., Shiraishi, Y. & Sugiura, Y. (2001). *Biochem. Biophys. Res. Commun.* **282**, 1001–1007.
Segal, D. J., Beerli, R. R., Blancafort, P., Dreier, B., Effertz, K., Huber, A., Koksch, B., Lund, C. V., Magnenat, L., Valente, D. & Barbas, C. F. III (2003). *Biochemistry*, **42**, 2137–2148.
Segal, D. J., Goncalves, J., Eberhardy, S., Swan, C. H., Torbett, B. E., Li, X. & Barbas, C. F. III (2004). *J. Biol. Chem.* **279**, 14509–14519.
Wolfe, S. A., Grant, R. A., Elrod-Erickson, M. & Pabo, C. O. (2001). *Structure*, **9**, 717–723.