Determining the DUF55 domain structure of human thymocyte nuclear protein 1 from crystals partially twinned by tetartohedry

Feng Yu,[ae] Aixin Song,[b] Chunyan Xu,[a] Lihua Sun,[ae] Jian Li,[ae] Lin Tang,[a] Minmin Yu,[c] Todd O. Yeates,[d] Hongyu Hu[b] and Jianhua He[a*]

[a]Shanghai Institute of Applied Physics, Chinese Academy of Sciences, People's Republic of China, [b]State Key Laboratory of molecular Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, People's Republic of China, [c]Physical Biosciences Division, Lawrence Berkeley National Laboratory, USA, and [d]Department of Chemistry and Biochemistry, University of California, USA, [e]Graduate School, Chinese Academy of Sciences, People's Republic of China. E-mail: hejh@sinap.ac.cn

Synopsis This paper reports the structure of the DUF55 domain from human thymocyte nuclear protein 1, which was determined from partial tetartohedrally twinned crystals.

Abstract Human thymocyte nuclear protein 1 (hTHYN1) contains a unique DUF55 domain of 167 residues (55 - 221), but its cellular function is unclear. Crystals of DUF55 belong to the trigonal space group $P3_1$, but twinning causes the data to approach an apparent 622 symmetry. Two datasets to 2.3 Å resolution were collected. Statistical analysis confirmed that both datasets were partially twinned by tetartohedry. Tetartohedral twin fractions were estimated. After the structure was determined, only one twofold axis of rotational pseudosymmetry was found in the crystal structure. Using the *DALI* program, a YTH domain, which is a potential RNA binding domain from human YTH domain-containing protein 2, was identified to have the most similar three-dimensional fold to DUF55. It is implied that DUF55 might be a potential RNA-related domain.

Keywords: THYN1; HSPC144; DUF55; tetartohedral twinning; YTH domain

1. Introduction

Owing partly to the large-scale use of synchrotron radiation and improvements in software tools, thousands of new protein structures are being determined each year. However, not all protein crystals can be determined by routine procedures; crystal twinning leads to many difficult cases (Yeates, 1997; Parsons, 2003). The causes of crystal twinning are not fully understood, but more than half of the reported cases occur in combination with pseudosymmetry (Lebedev *et al.*, 2006). In recent years, several interesting and complicated

cases have been determined successfully (*e.g.* (Gayathri *et al.*, 2007; MacRae & Doudna, 2007; Anand *et al.*, 2007) and reviewed in (Dauter *et al.*, 2005)). In the case of merohedral twinning, the multiple twin domains within a single crystal specimen usually cannot be distinguished by visual microscopic examination. Furthermore, because the reciprocal lattices of the twin domains overlap exactly in merohedral twinning, the observed diffraction pattern is typically unremarkable. Therefore, in order to detect twinning, the statistics of the intensity data must be examined (Stanley, 1972; Rees, 1980; Yeates, 1988; Padilla & Yeates, 2003).

To determine structures from twinned crystals, it is essential to identify the true space group, determine the underlying twin operation(s) (i.e. the symmetry operation(s) relating the twin domains), and estimate the twin fractions (the volume ratios of the component twin domains). Once the true space group has been confirmed, data reduction can be carried out as with ordinary diffraction data. Usually, twinning does not prevent structure determination by molecular replacement (Chandra *et al.*, 1999; Yeates, 1997). Phasing highly twinned data by MIR/MAD/SAD methods tends to be more challenging, however, this has also been accomplished successfully in several cases (Barends *et al.*, 2005; Yang *et al.*, 2000; Dauter, 2003; Yeates & Rees, 1987; Sultana *et al.*, 2007; Larsen & Harrison, 2004; Toms *et al.*, 2004; MacRae & Doudna, 2007; Rudolph *et al.*, 2003). In cases of twinning, atomic refinement can be carried out in different ways. For example, the effects of twinning can be applied to the model intensities, or the observed intensities can be corrected or 'detwinned' to obtain estimates of the true crystallographic intensities for comparison to the model intensities. Most reported cases of twinning in macromolecular crystals involve just two twin domain orientations (hemihedral twinning), but tetartohedral twinning involving four twin domains (*e.g. P*3 twinned towards apparent *P*622) is also possible (Rosendal *et al.*, 2004; Barends *et al.*, 2005; Gayathri *et al.*, 2007; Anand *et al.*, 2007), while an even higher category, ogdohedral twinning with eight twin domains (*e.g. P*3 twinned towards apparent *P*6/*mmm*) is also possible, though only for achiral or racemic mixtures of molecules. In the case of tetartohedral twinning, four twin domains are related by three twin operations, and the observed intensity of every reflection is a weighted sum of four twin-related reflection intensities:

$$J_{obs} = \alpha_1 I_1 + \alpha_2 I_2 + \alpha_3 I_3 + \alpha_4 I_4 \qquad \text{(Eq. 1)},$$

where J_{obs} is an observed (twinned) intensity, the I_k are true reflection intensities, and the α_k are the four tetartohedral twin fractions (which must sum to unity). This paper reports the structure of the DUF55 domain from human thymocyte nuclear protein 1 (hTHYN1), as determined from partial tetartohedrally twinned crystals using molecular replacement and detwinned data.

hTHYN1 was first identified in human $CD34^+$ hematopoietic stem/progenitor cells, as formerly named HSPC144 (Zhang *et al.*, 2000), but the biological function is unknown. It contains 225 amino-acid residues, and is highly conserved among a wide range of species including yeast, plants, and vertebrates. The previous studies of chicken and mouse THYN1 proteins indicate that THYN1 plays a role in apoptosis, but the mechanism is still unclear (Compton *et al.*, 2001; Jiang, Toyota, Yoshimoto *et al.*, 2003). There are some differences in the tissue distribution between chicken and mouse THYN1 (Miyaji *et al.*, 2002), however, the THYN1 protein exists mainly in the nucleus (Jiang, Toyota, Takada *et al.*, 2003). The recombinant full-length hTHYN1 protein is unstable at high concentration and is therefore not well suited for structural analysis. Using limited proteolysis and mass spectrometry, two stable domains, HSPC144-P (residues 44-225) and DUF55 (54-221) (formerly DUF589) have been identified (Song *et al.*, 2005). DUF55 is a unique domain of undefined function according to the Pfam database (Finn *et al.*, 2006), which exhibits highly conserved sequence in eukaryotes. Elucidating the domain structure of hTHYN1 may help us understand the function of these proteins in apoptosis.

2. Materials and methods

2.1. Protein expression and purification

DUF55 was expressed and purified as described previously (Song *et al.*, 2005). Briefly, the DUF55 DNA sequence (residues 55 - 221) was cloned into the plasmid pET22b, which produces a C-terminally His-tagged protein. DUF55 was purified using metal chelating chromatography, SP cation-exchange chromatography (Amersham) and gel filtration on Amersham HiLoad Superdex 75. Finally, the protein was concentrated in a buffer of 20 mM HEPES (pH7.2), 150 mM NaCl, 5% glycerol, and 2 mM DTT.

2.2. Protein crystallization

DUF55 at 10 $mg \cdot ml^{-1}$ was crystallized by the hanging drop vapor diffusion method in 291 K or 277 K. After screening numerous conditions, the following condition was identified: a hanging drop containing 1 µl of protein solution, 1 µl of reservoir solution (0.1 M NaAc pH 4.0 - 4.8, 28% PEG2000MME, 200 mM $(NH_4)_2SO_4$), and 0.2 µl 30% 1,6-diaminohexane as additive (1,6-diaminohexane is basic reagent, so actual pH is about 10.9). After a week, fan-like crystals appeared. The sizes of crystals obtained at 291 K (0.45 mm × 0.15 mm × 0.08 mm) were notably bigger than those obtained at 277 K (0.30 mm × 0.07 mm × 0.02 mm). Crystals were soaked in a cryoprotectant solution (0.1 M NaAc pH 4.0 - 4.8, 42%

PEG2000MME, 75 mM $(NH_4)_2SO_4$, 75 mM NaCl, 15% Glycerol, 3% 1,6-diaminohexane) and flash-frozen in liquid nitrogen.

2.3. Data collection

The diffraction of DUF55 crystals was too weak to collect usable datasets using a rotating-anode X-ray generator. Two datasets (named dataset 1 and dataset 2) to 2.3 Å resolution were collected on beamline 3W1A at the Beijing Synchrotron Radiation Facility and processed with *MOSFLM* (Leslie, 2006) and *SCALA* (Collaborative Computational Project, Number 4, 1994) (Bailey, 1994).

2.4. Data analysis and molecular replacement

Initially, the datasets were processed as $P6_{2/4}22$ according to self-rotation function plots (Figure 1) and detected systematic absences. However, crystal twinning was indicated by an impossibly low value for the Matthews coefficients in that symmetry (1.16 $Å^3 \cdot Da^{-1}$ for one molecule per asymmetric unit). The presence of twinning was confirmed by examination of the cumulative intensity distribution (Figure 2) and Padilla-Yeates local intensity statistics (Figure 3). Plausible lower symmetry space groups included $P3_{1/2}$, $P3_{1/2}12$, $P3_{1/2}21$ or $P6_{2/4}$. Molecular replacement searches were therefore carried out under all possible symmetries using the program *PHASER* (Mccoy et al., 2007), and dataset 1. The search model was PDB 2ar1 with 43% identity (Arakaki *et al.*, 2006). Although potential molecular replacement solutions were found under $P3_112$, $P3_121$, and $P6_4$, they were judged to be unreliable because of low Log Likelihood Gain (LLG) values (13, 49 and 5, respectively). A more reliable solution was found under $P3_1$ with an LLG value of 182. We tested out all 4 solutions in *CNS* (Brunger et al., 1998), and after one cycle of hemihedral/tetartohedral twinning refinement, only the solution in $P3_1$ led to improved R_{work} and R_{free} values simultaneously (Table 1), while the others got worse R_{free} though R_{work} were clearly decreased. The true space group was therefore assigned as $P3_1$, implying that the DUF55 crystals were tetartohedrally twinned. Similar results were obtained using dataset 2 (data not shown).

2.5. Estimation of tetartohedral twin fractions

Overall intensity statistics were evaluated in order to test for high or perfect twinning (Yeates, 1997), and the results (Table 2) confirmed severe twinning. In order to identify the twin operator and estimate the twin fraction, the data (reduced in $P3_1$) were evaluated using the *H*-test (Yeates, 1988) under the three possible twin operators in $P3_1$. The results (Table 2) suggested that dataset 1 was partially twinned by tetartohedry and that dataset 2 might be

nearly perfectly twinned by tetartohedry. We therefore attempted to refine the structure against dataset 2 using *CNS* and available scripts (Barends *et al.*, 2005). But it was not possible to obtain a satisfactory refinement. One possible explanation was that dataset 2 was also partially (but not perfectly) tetartohedrally twinned, and that the correct twin fractions therefore needed to be incorporated into the refinement. To this end, a new method for estimating tetartohedral twin fractions (Yeates & Yu, 2008) was applied. Two unique and equally plausible solutions for the four twin fraction values were obtained from this method; the correct solution had to be distinguished by further analysis. The potential solutions for the twin fractions for datasets 1 and 2 are shown in Table 2.

In the case of tetartohedral twinning, the observed intensities (J_k) are related to the true crystallographic intensities (I_k) as follows:

$$\begin{cases} \alpha_1 I_1 + \alpha_2 I_2 + \alpha_3 I_3 + \alpha_4 I_4 = J_1 \\ \alpha_1 I_2 + \alpha_2 I_1 + \alpha_3 I_4 + \alpha_4 I_3 = J_2 \\ \alpha_1 I_3 + \alpha_2 I_4 + \alpha_3 I_1 + \alpha_4 I_2 = J_3 \\ \alpha_1 I_4 + \alpha_2 I_3 + \alpha_3 I_2 + \alpha_4 I_1 = J_4 \end{cases} \qquad \text{(Eq. 2).}$$

The assignment of subscripts to twin operations is arbitrary, so we can say that I_1 refers to I(h,k,l), I_2 refers to I(k,h,-l), I_3 refers to I(-k,-h,-l) and I_4 refers to I(-h,-k,l), according to the three underlying twin operations. If the twin fractions (α_1, α_2, α_3, α_4) have been estimated, then the intensities can be detwinned (i.e. one can solve the equations above, given the observed intensities, to obtain the true intensities, setting $I_{true}=I_1$ arbitrarily). If the data have been detwinned successfully, then the resulting intensities should obey exponential statistics (Wilson, 1949). On the other hand, if the incorrect solution for the twin fractions is chosen, then the detwinned intensities may not follow the correct distribution. In this way it is possible to distinguish between the correct and incorrect solutions for the twin fractions (Yeates & Yu, 2008). This approach was used to analyze the potential twin fraction solutions for datasets 1 and 2 (Figure 4). In the case of dataset 1, the detwinned data calculated under solution 1 followed a somewhat more ideal distribution than those calculated under solution 2. In the case of dataset 2, solution 2 was much better.

It is important to note here that for each solution for the twin fraction values, there are four different orderings of the four twin fractions which are equally correct (Yeates & Yu, 2008); different orderings simply correspond to exchanging the assignments of I_1, I_2, I_3, and I_4 in equation 2. Which ordering of the twin fractions is chosen is therefore arbitrary, unless it is necessary to obtain agreement with a previously defined set of intensities. That was the case here, as a molecular replacement model (and its calculated intensities) had already been obtained (i.e. prior to detwinning). In this work therefore, detwinning under all four allowed permutations of the twin fraction solution was performed and the correct solution was

decided by the behavior of atomic refinement and by inspection of electron density maps. For completeness, the four orderings of the alternate (less plausible) solution for the twin fractions were also tested. The same procedure was applied to dataset 2. Thus, eight separate preliminary refinements were conducted using each of the two datasets (results for dataset 1 are shown in Table 3). Solution 1 for dataset 1 was (0.424, 0.300, 0.134, 0.142), with alternate orderings (0.300, 0.424, 0.142, 0.134), (0.134, 0.142, 0.424, 0.300), and (0.142, 0.134, 0.300, 0.424) being equally possible. It was verified that solution 1 of dataset 1 (0.424, 0.300, 0.134, 0.142) and solution 2 of dataset 2 (0.291, 0.276, 0.151, 0.282) were correct. This was consistent with the previous examination of detwinned intensity statistics (Figure 4).

2.6. Refinement under partial tetartohedral twinning

Because all four twin-related reflections must be observed for detwinning to be carried out (Eq. 2), data completeness is important for detwinning. We therefore refined the structure using dataset 1. An atomic model was rebuilt using *COOT* (Emsley & Cowtan, 2004), and then simulated annealing and *B*-factor refinement was performed, employing non-crystallographic symmetry restraints using *CNS*. After about ten cycles of refinement, most of the amino-acid residues could be traced, including four residues which were introduced by plasmid construction. However, the $R_{\rm detwin}$ and free $R_{\rm detwin}$ values remained relatively high (29.84% and 34.92%, respectively). It was surmised that this was likely due in part to the magnification of measurement errors caused by detwinning (*i.e.* by inverting the linear equations in Eq. 2). Consequently, after this point an alternate refinement strategy was employed. Instead of detwinning based on observed intensities alone, detwinning was performed using proportionality rules applied to the current values of the calculated model structure factors,

$$F_{\det win} = \sqrt{\frac{F_{obs}^2 + 0.424k(F_{calc} + F_{bulk})^2 - 0.300kT_1(F_{calc} + F_{bulk})^2 - 0.134kT_2(F_{calc} + F_{bulk})^2 - 0.142kT_3(F_{calc} + F_{bulk})^2}{2}} \quad \text{(Eq. 3)},$$

where k is a scale factor, T_1, T_2 and T_3 indicate the twin operations that apply to the structure factors, $F_{\rm calc}$ is the structure-factor array for macromolecular model including ordered solvent and $F_{\rm bulk}$ is the structure-factor array for an appropriate model of disordered solvent. In this way, detwinned (true) intensities can be obtained without the error magnification caused by inverting equation 2, although some model bias is likely introduced as a trade off. After every round of refinement and model building, $F_{\rm detwin}$ would be recalculated for obtaining more precise values. Following several cycles of recalculation of $F_{\rm detwin}$, refinement and model building, the final $R_{\rm detwin}$ and free $R_{\rm detwin}$ were 18.23% and 23.78%, respectively. After this structure had been determined, the final structure was refined against dataset 2. The final $R_{\rm detwin}$ and free $R_{\rm detwin}$ of the second structure were 21.67% and 25.63%, respectively.

The model quality was analyzed using *PROCHECK* (Laskowski et al., 1993). The pictures were drawn using *PyMOL*.

3. Results and Discussion

3.1. Overall Structure

The DUF55 crystal (dataset 1) was in the $P3_1$ space group with unit cell parameters 51.26 Å, 51.26 Å, 122.40 Å. Detailed crystallographic data statistics for DUF55 are shown in Table 4. There are 350 residues, including the His-tag, in an asymmetric unit, and 328 residues of them can be traced excluding the first residues of both chains, the region 64-71 of chain B, and the His-tag, since no clear electron density for these residues was observed. The model is therefore virtually complete. The amino-acid sequence numbering of DUF55 was chosen to match that of the full-length hTHYN1. Although 1,6-diaminohexane is important for crystallization, it does not appear in the electron density maps. In the refined structure, there are 2719 non-H protein atoms, 37 water molecules and 4 sulfate ions. 82.9% of all residues are located in the most allowed regions of a Ramachandran diagram. The R_{detwin} and free R_{detwin} factors are 18.23% and 23.78%, respectively. The R_{twin} and free R_{twin} values (14.81% and 19.12%, respectively), which were calculated after refinement, are lower than R_{detwin} and free R_{detwin} as expected due to the effects of statistical averaging in twinned intensities (Redinbo & Yeates, 1993). In order to avoid bias, quadruplets of twin-related reflections were kept together in either the test set or the refinement set throughout refinement. The detailed refinement statistics and quality of the model are shown in Table 5.

The DUF55 structure consists of six α-helices and six β-strands that were well defined (Figure 5A). The α-helices A, B and β-strands 1-6 form a notable surface cleft. There are five homologous structures known, of which four were determined by X-ray diffraction methods (PDB code: 2ar1, 2eve, 2g2x, 1zce). The r.m.s. deviation between DUF55 and the four homologous structures are 1.35 Å, 1.25 Å, 1.30 Å and 1.42 Å, respectively. Three of the homologous structures mentioned above contain ligands in this cleft, which are apparently carried along from the crystallization conditions or cryoprotectant solutions. 2ar1 contains a glycerol molecule in this cleft, while a MOPS molecule and sulfate ions exist in the same location of 2eve and 2g2x, respectively. Sulfate ions are also present in this cleft in the present structure of DUF55, but the locations of the sulfate ions are different. In DUF55, two of them are close to the location of sulfonic acid group in the MOPS molecule in 2eve, whereas the other two are close to Arg200 and Arg202 (Figure 5B). Unexpectedly, a well resolved intermolecular disulfide bond was found in both structures (Figure 5B) involving Cys118 of both chains (Figure 6).

There are two molecules in an asymmetric unit, which form a dimeric state through an intermolecular disulfide bond, while gel filtration experiments show the protein exists as a monomer in solution (Figure 7). No intermolecular disulfide bond is observed in the other structures though homologous Cys118 also exists in 2eve and 2g2x. In three of the four structures, only one monomer exists in an asymmetric unit (2ar1, 2eve and 1zce). 2g2x is the exception, where an asymmetric unit contains three molecules without apparent non-crystallographic point group symmetry; interactions between different pairs of molecules are distinct. Thus, the dimeric structure observed in the crystalline state in the present work may not be biologically significant.

3.2. Structure similarity to the YTH domain in YTH domain-containing protein 2

A structural similarity search was performed using the *DALI* program (Holm & Sander, 1996). Most of the hits have no function annotations. The highest Z score (8.4) of those, which have function annotations, is from the YTH domain in the human YTH domain-containing protein 2 (PDB code: 2yu6). YTH is a potential RNA-binding domain (Stoilov *et al.*, 2002). After superimposing DUF55 with the YTH domain using secondary structure matching, most of the α-helices and β-strands could be superimposed; the r.m.s. deviation of Cα atoms is 3.4 Å. Sequence alignment shows that Lys60, Ser61 and Trp90 are highly conserved in the homologous DUF55 domains and the YTH domain. The spatial locations of these three residues are also conserved in 2yu6 and the DUF55. It is implied that DUF55 might be a potential RNA-related domain. Other structural neighbours of known function identified here had been reported previously based on a search for folds similar to 2ar1 (Arakaki *et al.*, 2006). The highest Z score among those is 4.7 for the N-terminal domain of *E. coli* Lon protease (PDB code: 2ane).

3.3. Crystal packing and twinning

Analyzing the DUF55 crystal packing, it is seen that the NCS twofold is nearly parallel to [100]. The NCS rotation matrix calculated using *CNS* is:

$$\begin{pmatrix} 0.99998 & 0.00599 & -0.00221 \\ 0.00608 & -0.99915 & 0.04086 \\ -0.00197 & -0.04087 & -0.99916 \end{pmatrix} \approx \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

This has been referred to as rotational pseudosymmetry (RPS) (Zwart *et al.*, 2008). The presence of RPS apparently promotes the growth of twinned crystals due to layering and low steric hindrance of molecules at the twin domain interface. It is however not a requirement for twin formation.

To our knowledge, four tetartohedrally twinned protein crystal structures have been determined before this work (PDB code: 1qzw, 2pi8, 2h6r and 2pk2). Three of them contain 222 rotational pseudosymmetry (1qzw, 2pi8, 2h6r), occurring with non-crystallographic symmetry (NCS) (Rosendal *et al.*, 2004; Barends *et al.*, 2005; Gayathri *et al.*, 2007). 2pk2 also has three mutual perpendicular NCS twofold symmetry, which nearby parallels to three twin operators, respectively (Anand *et al.*, 2007). Because of their fortunate orientations, layering would also occur resulting in low steric hindrance at the twin domain interface. Hence the NCS two symmetry of 2pk2 has similar function to 222 rotational pseudosymmetry. However, in the crystal of DUF55 there is no additional NCS twofold axis that would be required to generate local 222 symmetry. Layering does not seem to occur along the [001] and [120] directions, so a layering effect does not appear to promote twinning in the present case. It is easier to interpret the causes of twinning in the other four known tetartohedral twinning cases owing to their 222 symmetry. DUF55 provides an example with only one axis of rotational pseudosymmetry. This suggests that a tetartohedrally twinned case without any NCS rotational pseudosymmetry might be found in the future.

Acknowledgements We would like to thank Prof. Zongxiang Xia, Prof. Bauke W. Dijkstra and Karin van Straaten for the help in the structure refinement. Part of the diffraction data used in this study were collected at University of Science and Technology of China, we are grateful to Prof. Maikun Teng and Xiao Zhang. This work was supported by Shanghai Natural Science Foundation (Grant No. 07JC14062), and National Natural Science Fundation (Grant No. 10774155). This work was also supported the by Department of Energy under Contract DE-AC02-05CH11231.

References

Anand, K., Schulte, A., Fujinaga, K., Scheffzek, K. & Geyer, M. (2007). *J Mol Biol* **370**, 826-836.

Arakaki, T., Le Trong, I., Phizicky, E., Quartley, E., DeTitta, G., Luft, J., Lauricella, A., Anderson, L., Kalyuzhniy, O., Worthey, E., Myler, P. J., Kim, D., Baker, D., Hol, W. G. J. & Merritt, E. A. (2006). *Acta Cryst.* **F62**, 175-179.

Bailey, S. (1994). *Acta Cryst.* **D50**, 760-763.

Barends, T. R. M., de Jong, R. M., van Straaten, K. E., Thunnissen, A. M. W. H. & Dijkstra, B. W. (2005). *Acta Cryst.* **D61**, 613-621.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* **D54**, 905-921.

Chandra, N., Acharya, K. R. & Moody, P. C. E. (1999). *Acta Cryst.* **D55**, 1750-1758.

Compton, M. M., Thomson, J. M. & Icard, A. H. (2001). *Apoptosis* **6**, 299-314.

Dauter, Z. (2003). *Acta Cryst.* **D59**, 2004-2016.

Dauter, Z., Botos, I., LaRonde-Leblanc, N. & Wlodawer, A. (2005). *Acta Cryst.* **D61**, 967-975.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* **D60**, 2126-2132.

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L. & Bateman, A. (2006). *Nucl. Acids Res.* **34**, D247-D251.

Gayathri, P., Banerjee, M., Vijayalakshmi, A., Azeez, S., Balaram, H., Balaram, P. & Murthy, M. R. N. (2007). *Acta Cryst.* **D63**, 206-220.

Holm, L. & Sander, C. (1996). *Science* **273**, 595-602.

Jiang, X. Z., Toyota, H., Takada, E., Yoshimoto, T., Kitamura, T., Yamada, J. & Mizuguchi, J. (2003). *Tissue Cell* **35**, 471-478.

Jiang, X. Z., Toyota, H., Yoshimoto, T., Takada, E., Asakura, H. & Mizuguchi, J. (2003). *Apoptosis* **8**, 509-519.

Larsen, N. A. & Harrison, S. C. (2004). **344**, 885-892.

Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283-291.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Acta Cryst.* **D62**, 83-95.

Leslie, A. G. W. (2006). *Acta Cryst.* **D62**, 48-57.

MacRae, I. J. & Doudna, J. A. (2007). *Acta Cryst.* **D63**, 993-999.

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.

Miyaji, H., Yoshimoto, T., Asakura, H., Komachi, A., Kamiya, S., Takasaki, M. & Mizuguchi, J. (2002). *Gene* **297**, 189-196.

Padilla, J. E. & Yeates, T. O. (2003). *Acta Cryst.* **D59**, 1124-1130.

Parsons, S. (2003). *Acta Cryst.* **D59**, 1995-2003.

Redinbo, M. R. & Yeates, T. O. (1993). *Acta Cryst.* **D49**, 375-380.

Rees, D. (1980). *Acta Cryst.* **A36**, 578-581.

Rosendal, K. R., Sinning, I. & Wild, K. (2004). *Acta Cryst.* **D60**, 140-143.

Rudolph, M. G., Kelker, M. S., Schneider, T. R., Yeates, T. O., Oseroff, V., Heidary, D. K., Jennings, P. A. & Wilson, I. A. (2003). *Acta Cryst.* **D59**, 290-298.

Song, A. X., Chang, Y. G., Gao, Y. G., Lin, X. J., Shi, Y. H., Lin, D. H., Hang, Q. H. & Hu, H. Y. (2005). *Protein Expres Purif* **42**, 146-152.

Stanley, E. (1972). *J. Appl. Cryst.* **5**, 191-194.

Stoilov, P., Rafalska, I. & Stamm, S. (2002). *Trends Biochem Sci* **27**, 495-497.

Sultana, A., Alexeev, I., Kursula, I., Mantsala, P., Niemi, J. & Schneider, G. (2007). *Acta Cryst.* **D63**, 149-159.

Toms, A. V., Kinsland, C., McCloskey, D. E., Pegg, A. E. & Ealick, S. E. (2004). *J Biol Chem* **279**, 33837-33846.

Wilson, A. (1949). *Acta Cryst.* **2**, 318-321.

Yang, F., Dauter, Z. & Wlodawer, A. (2000). *Acta Cryst.* **D56**, 959-964.

Yeates, T. (1988). *Acta Cryst.* **A44**, 142-144.
Yeates, T. O. (1997). *Method Enzymol* **276**, 344-358.
Yeates, T. O. & Rees, D. C. (1987). *Acta Cryst.* **A43**, 30-36.
Yeates, T. O. & Yu, F. (2008). *Acta Cryst.* **D64**, 1158-1164.
Zhang, Q. H., Ye, M., Wu, X. Y., Ren, S. X., Zhao, M., Zhao, C. J., Fu, G., Shen, Y., Fan, H. Y., Lu, G., Zhong, M., Xu, X. R., Han, Z. G., Zhang, J. W., Tao, J., Huang, Q. H., Zhou, J., Hu, G. X., Gu, J., Chen, S. J. & Chen, Z. (2000). *Genome Res* **10**, 1546-1560.
Zwart, P. H., Grosse-Kunstleve, R. W., Lebedev, A. A., Murshudov, G. N. & Adams, P. D. (2008). *Acta Cryst.* **D64**, 99-107.

Figure 1 Self-rotation function plots of dataset 1 ($\kappa = 180^{o}$).

Figure 2 Cumulative Wilson distributions for acentric-reflection intensities ($z, z = I / \sum f_i^2$, f_i is scattering factors of atoms). For DUF55 data (dataset 1), the distribution is strongly sigmoidal, which is indicative of twinning.

Figure 3 Analysis of Padilla-Yeates local intensity statistics (Padilla & Yeates, 2003). Theoretical distributions for untwinned acentric data are shown by the red straight line, those for perfectly hemihedrally twinned (acentric) data are shown by the red curve, and the distribution for DUF55 acentric data (dataset 1) is shown by the blue curve.

Figure 4 The comparison of cumulative Wilson distributions calculated under two possible solutions for the tetartohedral twin fractions. (a) dataset 1. The detwinned data calculated under solution 1 are better than those calculated under solution 2. (b) dataset 2. Solution 2 is much better.

Figure 5 (a) A ribbon diagram of DUF55 showing secondary-structure elements with labels (the α-helices are lettered and the β-strands are numbered). (b) Dimer structure of DUF55. The intermolecular disulfide bond is colored blue, and four sulfate ions are also shown.

Figure 6 Stereo view of the intermolecular disulfide bond, which is formed by Cys 118 residues from both chains. The $2F_o$-F_c map is set to 1 σ. (white: C; blue: N; red: O; yellow: S)

Figure 7 Analytical gel filtration profile for DUF55. The blocks are the standard markers, and the line is a linear fitting curve. The theoretical molecular weight of DUF55 with his-tag is 20.5kDa, so DUF55 exists in a monomeric state in solution. The globular marker proteins are RNase A (13.7 kDa), Carbonic Anhydrase (29.0 kDa), Ovalbumin (44.3 kDa) and Conalbumin (75.0 kDa).

Table 1 Results of molecular replacement of dataset 1

	$P3_1$	$P3_121$	$P3_112$	$P6_4$
LLG	182	49	13	5
Before refinement				
R_{twin}†	0.3083	0.4569	0.4515	0.4475
Free R_{twin}	0.3140	0.4937	0.4633	0.4302
After twinning refinement‡				
R_{twin}	0.2400	0.3815	0.3892	0.3620
Free R_{twin}	0.2984	0.5077	0.4743	0.4384

†In this paper, R_{twin} for hemihedral twinning was calculated as,

$$R_{twin} = \sum \left| |F_{obs}| - \left[(1-\alpha)|F_{calc} + F_{bulk}|^2 + \alpha T |F_{calc} + F_{bulk}|^2 \right]^{1/2} \right| \Big/ \sum |F_{obs}| \quad ;$$

R_{twin} for tetartohedral twinning was calculated as,

$$R_{twin} = \frac{\sum \left| |F_{obs}| - \left[\alpha_1 |F_{calc} + F_{bulk}|^2 + \alpha_2 T_1 |F_{calc} + F_{bulk}|^2 + \alpha_3 T_2 |F_{calc} + F_{bulk}|^2 + \alpha_4 T_3 |F_{calc} + F_{bulk}|^2 \right]^{1/2} \right|}{\sum |F_{obs}|} \quad ,$$

where T_1, T_2 and T_3 indicate the twin operations that apply to the structure factors, F_{calc} is the structure-factor array for macromolecular model including ordered solvent and F_{bulk} is the structure-factor array for an appropriate model of disordered solvent. (Barends *et al.*, 2005).

‡ In all of cases, there were assumed as perfect twinning. For hemihedral twinning, refinements were carried out with *CNS* and anneal_twin.inp/bindividual_twin.inp. For tetartohedral twinning, refinement scripts for perfect tetartohedral twinning were developed by Barends et al. (Barends *et al.*, 2005)

Table 2 Twinning tests

	Dataset 1	Dataset 2
Perfect twinning test		
$<\|I\|^2>/(<\|I\|>)^{2\dagger}$	1.5665	1.4530
$(<\|F\|>)^2/<\|F\|^2>^{\ddagger}$	0.8818	0.9022
H-test[$]		
-h,-k,l	0.327	0.447
k,h,-l	0.460	0.449
-k,-h,-l	0.327	0.439
Estimated tetartohedral twin fractions (Yeates & Yu, 2008)		
Solution 1(α_1, α_2, α_3, α_4)	0.424, 0.300, 0.134, 0.142	0.349, 0.218, 0.209, 0.224
Solution 2(α_1, α_2, α_3, α_4)	0.076, 0.200, 0.366, 0.358	0.151, 0.282, 0.291, 0.276

† 2.0 for untwinned, 1.5 for perfectly hemihedrally twinned, 1.25 for perfectly tetartohedrally twinned. (Stanley, 1972)

$<|I|^2>/(<|I|>)^2$ for perfect tetartohedral twinning was calculated as,

$$\int_0^{+\infty} z^2 \times \frac{4^4}{\Gamma(4)} z^3 \exp(-4z)dz = \int_0^{+\infty} \frac{128}{3} z^5 \exp(-4z)dz = 1.25 \cdot$$

‡ 0.785 for untwinned, 0.885 for perfectly hemihedrally twinned, 0.940 for perfectly tetartohedrally twinned. (Stanley, 1972)

Wilson ratio for perfect tetartohedral twinning was calculated as,

$$\left(\int_0^{+\infty} z^{1/2} \times \frac{4^4}{\Gamma(4)} z^3 \exp(-4z)dz \right)^2 = \left(\int_0^{+\infty} \frac{128}{3} z^{7/2} \exp(-4z)dz \right)^2 = 0.940 \cdot$$

$ The values listed represent estimates of the hemihedral twin fraction, α, assuming hemihedral twinning about the specified operator. These values do not translate directly to tetartohedral twin fractions.

Table 3 Comparison of partial refinements under the eight possible twin fraction solutions for dataset 1†

Twin fractions	R_{detwin}/free R_{detwin}		R_{twin}/free R_{twin}	
Solution 1	Before refinement	After refinement	Before refinement	After refinement
0.424, 0.300, 0.134, 0.142	**46.19%/44.82%**	**37.50%/42.76%**	**31.45%/32.07%**	**25.88%/28.60%**
0.300, 0.424, 0.142, 0.134	46.97%/45.78%	41.38%/47.98%	31.43%/32.05%	27.67%/31.65%
0.134, 0.142, 0.424, 0.300	50.84%/51.12%	46.54%/52.89%	32.75%/33.14%	30.80%/32.28%
0.142, 0.134, 0.300, 0.424	50.46%/51.51%	46.48%/51.64%	32.72%/33.09%	30.31%/33.81%
Solution 2				
0.076, 0.200, 0.366, 0.358	50.13%/51.35%	48.17%/53.17%	33.04%/33.27%	31.37%/34.34%
0.200, 0.076, 0.358, 0.366	49.15%/48.75%	43.78%/47.68%	32.77%/33.13%	29.62%/31.45%
0.366, 0.358, 0.076, 0.200	45.91%/45.39%	38.51%/45.95%	31.67%/32.36%	26.74%/30.42%
0.358, 0.366, 0.200, 0.076	46.08%/44.74%	39.11%/45.13%	31.53%/32.20%	26.65%/ 30.28%

†Refinements were carried out with detwinned data. All twin related reflections were kept together in either the test set or the refinement set. The reported values are from preliminary refinements based on detwinning without the aid of calculated model structure factors (see text). Final refinement values are given in Table 5.

Table 4 Crystallographic data statistics for DUF55

Values in parentheses are for the highest resolution shell.

	Dataset 1	Dataset 2
Space group	$P3_1$	$P3_1$
Unit cell parameters (Å)	a=b=51.26, c=122.40	a=b=51.36, c=122.75
Resolution limit (Å)	44.41-2.30(2.42-2.30)	44.50-2.30(2.42-2.30)
R_{merge} (%)	5.6(29.1)	5.4(24.9)
Total number of observations	77043(10467)	122732(15115)
Total number unique	15931(2326)	15329(2094)
<I/sigma(I)>	22.5(5.4)	27.9(6.7)
Completeness (%)	99.4(98.8)	95.1(88.8)
Multiplicity	4.8(4.5)	8.0(7.2)
Molecules per AU	2	2

Table 5 Refinement statistics and quality of the model.

Values in parentheses are for the highest resolution shell.

R_{detwin}	18.23(25.13)
Free R_{detwin}	23.78(31.86)
R_{twin}	14.78(21.84)
Free R_{twin}	19.10(25.63)
Model quality	
No. of atoms	
Protein atoms	2719
Sulfate ions	4
Waters	37
Average *B* factors (Å^2)	
Mean *B* value	36.24
B value from Wilson plot	40.06
R.m.s. deviations from ideal values	
Bond lengths (Å)	0.0060
Bond angles (°)	1.1706
Residues in Ramachandran plot (%)	
Most allowed region	82.9
Allowed region	17.1
Generously allowed region	0.0
Disallowed region	0.0

FIGURE 1

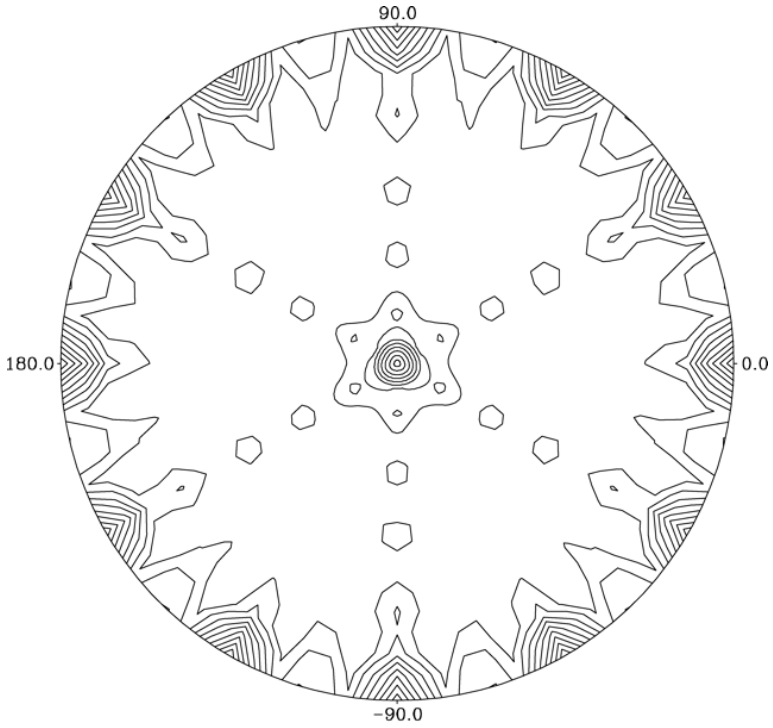


FIGURE 2

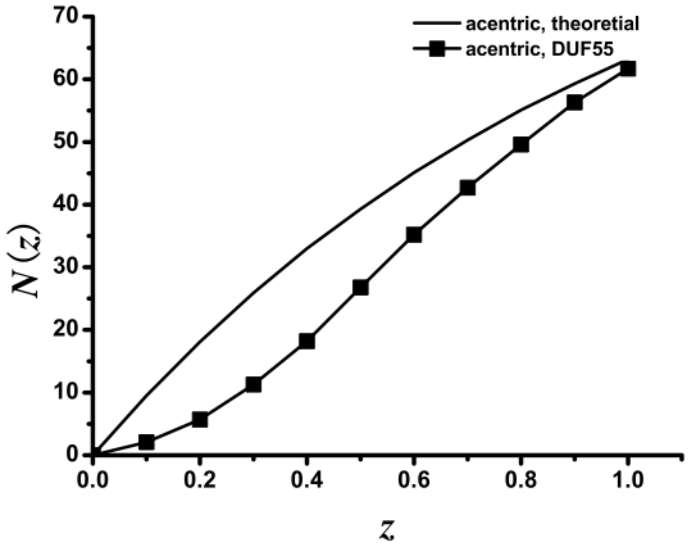

acentric, theoretial
acentric, DUF55
70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
N(z)
z

FIGURE 3

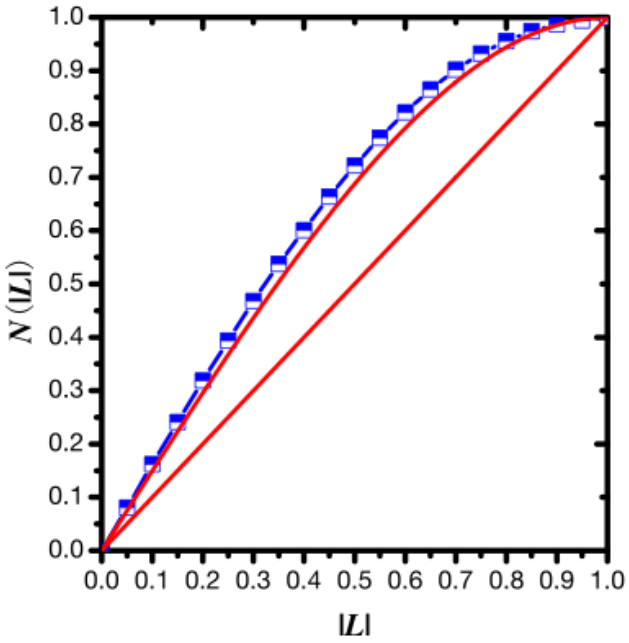
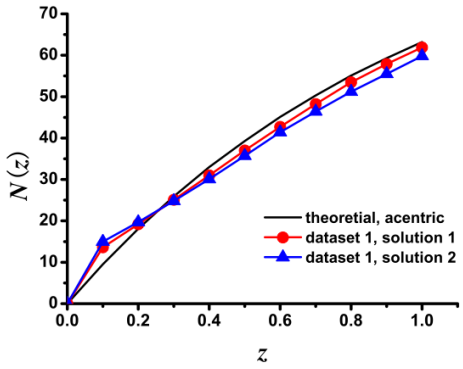


FIGURE 4

A


70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
N(z)
z
theoretial, acentric
dataset 1, solution 1
dataset 1, solution 2

B

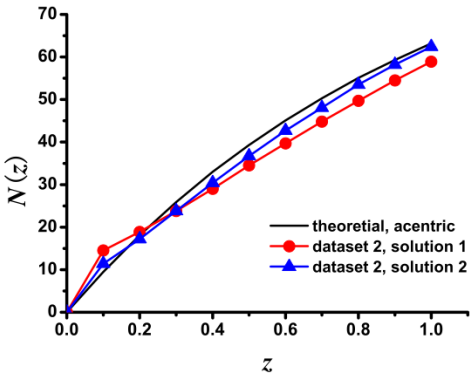
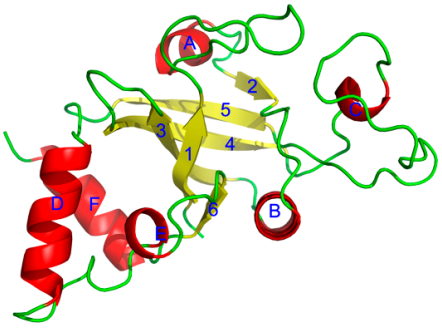

70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
N(z)
z
theoretial, acentric
dataset 2, solution 1
dataset 2, solution 2

FIGURE 5

A



B

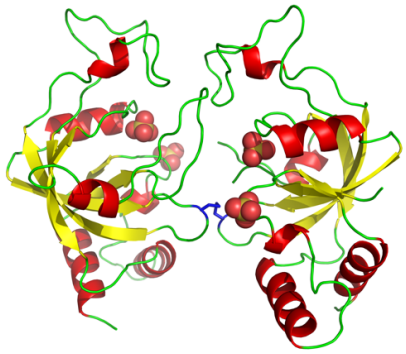


FIGURE 6

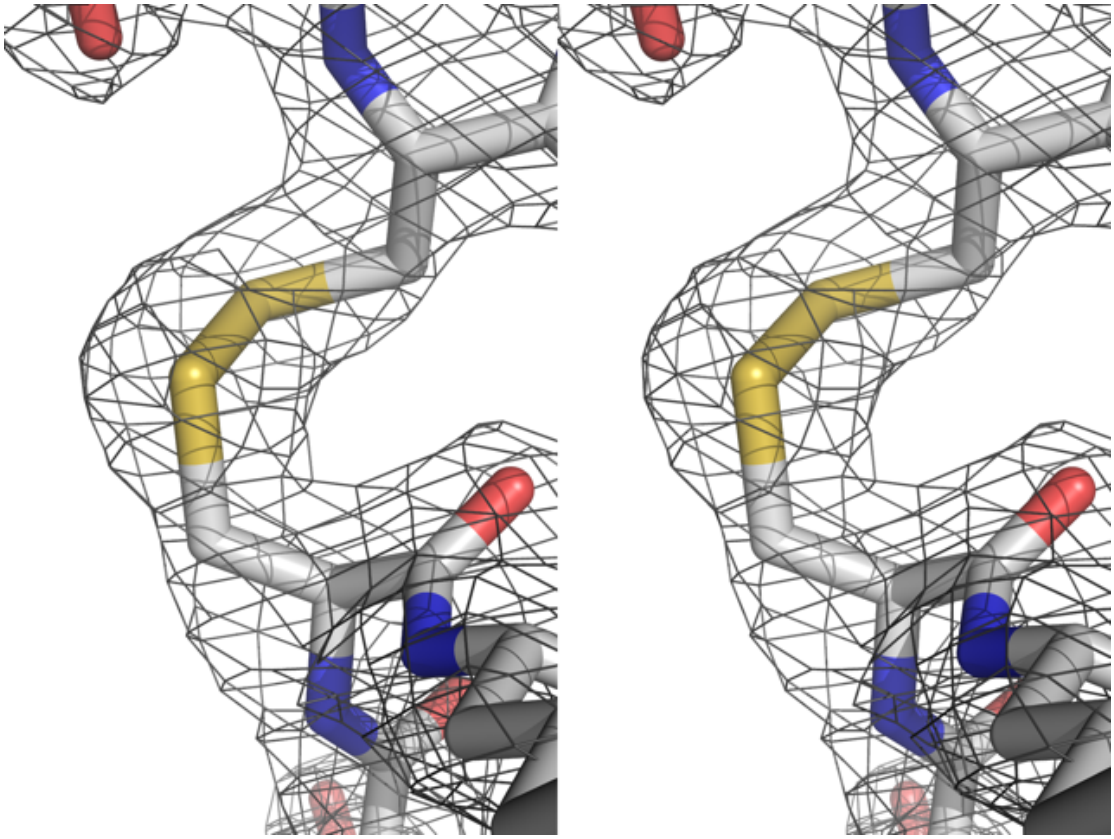
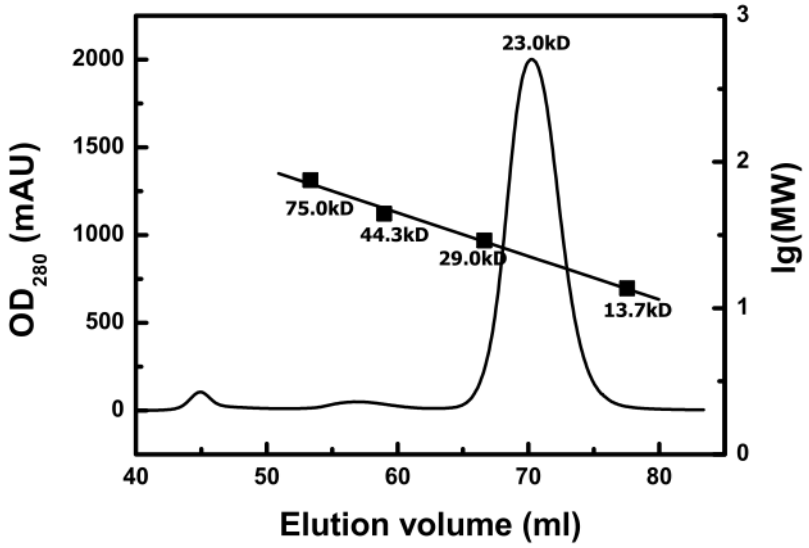


FIGURE 7


23.0kD
75.0kD
44.3kD
29.0kD
13.7kD
OD280 (mAU)
lg(MW)
Elution volume (ml)
2000
1500
1000
500
0
3
2
1
0
40
50
60
70
80