Synthetic symmetrization in the crystallization and structure determination of CelA from *Thermotoga maritima*

G. Jason Forse,¹ Nina Ram,¹ D. Rey Banatao,¹ Duilio Cascio,¹ Michael R. Sawaya,¹ Heath E. Klock,² Scott A. Lesley,² and Todd O. Yeates^{1*}

¹Department of Chemistry and Biochemistry, UCLA, Los Angeles, California 90095

²Joint Center for Structural Genomics, Genomics Institute of the Novartis Research Foundation, San Diego, California 92121

Received 31 July 2010; Revised 1 November 2010; Accepted 2 November 2010

DOI: 10.1002/pro.550

Published online 16 November 2010 proteinscience.org

Abstract: Protein crystallization continues to be a major bottleneck in X-ray crystallography. Previous studies suggest that symmetric proteins, such as homodimers, might crystallize more readily than monomeric proteins or asymmetric complexes. Proteins that are naturally monomeric can be made homodimeric artificially. Our approach is to create homodimeric proteins by introducing single cysteines into the protein of interest, which are then oxidized to form a disulfide bond between the two monomers. By introducing the single cysteine at different sequence positions, one can produce a variety of synthetically dimerized versions of a protein, with each construct expected to exhibit its own crystallization behavior. In earlier work, we demonstrated the potential utility of the approach using T4 lysozyme as a model system. Here we report the successful application of the method to *Thermotoga maritima* CelA, a thermophilic endoglucanase enzyme with low sequence identity to proteins with structures previously reported in the Protein Data Bank. This protein had resisted crystallization in its natural monomeric form, despite a broad survey of crystallization conditions. The synthetic dimerization of the CelA mutant D188C yielded well-diffracting crystals with molecules in a packing arrangement that would not have occurred with native, monomeric CelA. A 2.4 Å crystal structure was determined by single anomalous dispersion using a seleno-methionine derivatized protein. The results support the notion that synthetic symmetrization can be a useful approach for enlarging the search space for crystallizing monomeric proteins or asymmetric complexes.

Keywords: macromolecular crystallization; symmetry; protein design; oligomer; lattice contact; disulfide bond

Introduction

The problem of obtaining well-ordered crystals persists as the major obstacle in macromolecular crys-

tallography. Recent data from structural genomics efforts indicate that only about one-fourth of those proteins that can be expressed and purified in soluble form go on to yield high quality crystals when operating within a high throughput pipeline.^{1,2} There is therefore considerable interest in developing generally applicable salvage methods for improving the success rates of crystallization.³

General ideas that have been pursued for improving crystallization include (1) choosing or creating variations on a macromolecule of interest to sample distinct crystallization properties, and

Abbreviations: DTT, dithiothreitol; HEPES, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; MOPS, 3-(N-morpholino)-propanesulfonic acid; MPD, 2-methyl-2,4-pentanediol; β-ME, β-mercaptoethanol

Grant sponsor: NIH; Grant number: PSI U54 GM074898; Grant sponsor: BER program (DOE office of Science)

*Correspondence to: Todd O. Yeates, Department of Chemistry and Biochemistry, UCLA, 611 Charles Young Dr. East, Los Angeles, CA 90095-1569. E-mail: yeates@mbi.ucla.edu

Figure 1. Application of synthetic symmetrization to the structure determination of CelA. (A) By introducing a single cysteine at different sequence positions (yellow), disulfide crosslinking can produce different synthetically dimerized constructs, each with its own unique opportunities for crystallization. Figure adapted from Banatao *et al.* 2006.²⁴ (B) Eight single cysteine mutants (yellow) for disulfide-based dimerization of CelA were chosen before knowledge of the crystal structure (shown). The selections were based mainly on the predicted likelihood of being surface exposed. Due to the low sequence identity of CelA to other proteins of known structure, predictions relied on sequence alignments and considerations of sequence conservation. The selected residues were at positions in the amino acid sequence that were poorly conserved and charged. (C) Two different orientations of CelA showing surface representations with residues colored by surface accessible area from DSSP.²⁷ Cysteine mutations at sites most exposed to the protein surface formed dimers with the highest yield. Notably, the sites D188C (red) and E150C (yellow) are in loops with sharp turns facilitated by a proline residue. The dimer formed by the D188C mutant was the subject of crystal structure determination.

(2) modifying the target molecule in a way that might specifically promote its ability to form good lattice contacts in a crystal. Some approaches exploit both ideas. Trying to crystallize homologues of a protein of interest is an example of varying the protein to be crystallized.^{4,5} Disordered regions, often at the N- or C-terminus of the protein⁶ can be removed either by proteolysis³ or targeted deletion based on detection of disorder.⁷ The methods of lysine methylation^{8–10} and side chain entropy reduction^{11,12} aim to alter the surface properties of a protein and thereby its crystallization behavior. They may also promote lattice contacts directly. Other examples include the use of antibody fragments,^{13–16} other binding partners,^{17–19} or fused or inserted protein domains to improve the chances of forming lattice contacts.^{20–23} Such methods have been especially important in the case of transmembrane proteins.

A few recent studies have advanced an approach to crystallization that we refer to as ‘synthetic symmetrization’, which combines the ideas noted above, namely variation and specific design of contacts.^{24–26} The essence is to build in a self-associating contact or bond to generate a dimeric (or potentially higher oligomeric) version of the target protein molecule. This approach introduces a specific interaction that can promote the formation of a crystal lattice, especially if the self-association leads to a symmetric arrangement. Furthermore, to the extent that a self-associating contact can be introduced at various places in the surface of the molecule, various independent constructs can be created [Fig. 1(A)]. Each of these constructs offers entirely distinct opportunities for crystallization.

In a proof of principle study of synthetic symmetrization, single cysteine residues were introduced in three different positions of phage T4

lysozyme—these were used to generate three distinct lysozyme dimers by disulfide cross-linking—from which six new crystal forms of that protein were obtained.²⁴ In more recent work by other groups, motifs other than disulfide bonding have been used to drive the self-association of a target protein to promote symmetrization²⁸ and crystallization.^{25,26} Here, we report on a test of disulfide-based synthetic symmetrization on a set of proteins, and report the successful crystallization and structure determination of one of them, CelA, an endoglucanase from the thermophilic bacterium *Thermotoga maritima*. The native form of the protein had resisted earlier attempts at crystallization, and only remote homologues had been structurally elucidated.

Results

Protein selection

A list of proteins for testing synthetic symmetrization as a crystallization strategy was constructed based on data from the Joint Center for Structural Genomics (JCSG, www.jcsg.org). Proteins from *Thermotoga maritima* that had entered crystallization trials, but did not yield diffracting crystals, were taken as a starting set. Additional criteria were used to narrow the list further. To reduce complications associated with removing native cysteines, proteins lacking cysteine residues in their native sequences were identified. Our goal was to show that the method can be applied to unique protein targets in the hopes of crystallizing a protein that would reveal novel structural information. Towards that end, candidate proteins were selected having the lowest sequence similarity to proteins with previously reported structures. Based on those criteria, a set of six proteins was selected for further study: TM0227, TM0252, TM0733, TM0829, TM1059, and TM1524. In each target protein, several positions were chosen for cysteine mutagenesis based on criteria discussed below (also see Materials and Methods). In initial experiments, the mutants were expressed and purified as part of the JCSG crystallization pipeline.²

In initial attempts to prepare dimers of the selected proteins by disulfide cross-linking, aggregation was a prevalent problem affecting the majority of the constructs. We anticipated that this might relate to cross-linking of proteins that were either oligomeric or heterogeneous in their assembly states, given that polyvalent polymerization is prone to formation of large assemblies such as filaments and irregular networks. Size exclusion chromatography and dynamic light scattering were used to assess oligomeric state and monodispersity. Protein TM0252 is subunit C of a heterotrimeric glutamyl-tRNA amidotransferase. Size exclusion chromatography indicated that the protein existed as soluble species of

large size, potentially reflecting a tendency to aggregate in the absence of GatA and GatB. The dimeric forms of some GatC cysteine mutants remained soluble but were not pursued owing to the higher assembly states observed. For protein TM0733, light scattering indicated the formation of soluble aggregates, and dimerization led generally to insoluble material. Protein TM0829 is a CBS domain protein that appeared to be a mixture of dimers and tetramers by size exclusion chromatography. Some of the dimerized proteins remained soluble, but did not yield well-ordered crystals. Protein TM0227 gave material with low solubility; the sequence was subsequently judged to not be a bona fide protein and was not pursued. Protein TM1029 is a sporulation coat protein, whose annotation suggests a potential for assembly. Cross-linked mutants were insoluble. Ultimately only one of the protein candidates, TM1524 (CelA), could be purified in a monodisperse, monomeric state. Subsequent stages of the synthetic symmetrization experiments focused on the monomeric protein CelA, a predicted endoglucanase with potential applications in cellulose degradation.

Design of cysteine mutants

In this work, the plan was to use cysteine mutagenesis to introduce specific sites for disulfide-based dimerization. Eight sites (K7, E31, K57, E150, D188, H201, D204, and K208) were selected for introducing single cysteine mutations into CelA [Fig. 1(B)]. The intention was to choose amino acid positions that were likely to be exposed on the surface of the protein, but also relatively well-ordered structurally. Given that the target proteins had been selected on the basis of having little or no sequence similarity to other proteins of known structure, the selection of positions with the desired properties relied mainly on information from multiple sequence alignments.

Several principles were considered in selecting amino acid positions for disulfide-based dimerization. Charged residues with low sequence conservation have a high likelihood of residing on the protein surface.^{29,30} Also, poorly conserved positions are less likely to be important for protein structure and stability. High polarity and low sequence conservation were therefore used as general selection criteria.

Two further ideas were employed in selecting cysteine mutations. We anticipated that a crosslink involving a structurally ordered region would be more likely to produce a well-ordered (i.e., less flexible) dimer with favorable chances of crystallization. Segments of Cel A predicted to be α -helical (using the program Psipred³¹) were viewed as having a high probability of being structurally ordered. The three sites in lysozyme that were used successfully to obtain novel crystal forms of that protein were in surface-exposed α -helices.²⁴ In CelA, the segment between residues 201 and 214 was predicted to be

helical. Residues H201, D204, and K208 were predicted by visual inspection to be exposed to solvent on one side of the helix based on their charge properties and spacing. With only one helix available to alter in the sequence of CelA, an additional motif was investigated, a charged residue following a proline, with the expectation that this might represent an accessible side chain in a more conformationally rigid environment. This motif could provide a sharp turn that orients a thiol well away from the rest of the protein in a relatively ordered context. In CelA, four such motifs were selected, P6/K7, P56/K57, P147/G/G/E150, and K187/D188 [Fig. 1(C)].

Synthetic Homodimerization

The eight single cysteine mutants of CelA were tested for their ability to form stable dimers in high yield upon oxidation of the cysteine residues to a disulfide bond. The proteins were dimerized with CuSO₄ after reduction by and removal of DTT (see Methods for details). The yields of the reactions were assessed on a native gel with and without reducing agent. The mutants yielding suitable amounts of dimeric protein were based on a charged residue proximal to a proline (E150C and D188C). The failure of other mutants to dimerize efficiently appears to relate to incomplete solvent accessibility (discussed below). Oxidation of the D188C mutant produced dimers with a yield of over 95%. The dimeric form of this mutant was therefore pursued most vigorously. The dimeric species was further purified from smaller molecular weight species by gel filtration and concentrated before crystallization trials. Compared to the native, monomeric protein, the dimeric protein was only half as soluble, to ~5 mg/mL. An explanation for this change in solubility was not evident to us even in a retrospective analysis.

The D188C CelA mutant was crystallized by hanging drop vapor diffusion. The diffraction data were indexed in hexagonal space group P6₃ with two monomers in the asymmetric unit. Attempts to phase the X-ray diffraction data by molecular replacement were unsuccessful, a failure that was not unexpected in view of the low sequence identity (21%) between CelA and the closest homolog of known structure (PDB 1H0B). Phasing was therefore pursued by multiple isomorphous replacement using heavy atoms and halides, and single-wavelength anomalous dispersion (SAD) with a Se-Met variant. A solution to the structure was obtained using SAD data from the Se-Met protein, initially phased to a resolution of 3.0 Å. After initial model building, a final structure was refined to a resolution of 2.4 Å. The R and R_{free} values are 17.4 and 21.5%. 526 out of a total of 528 residues in the asymmetric unit are present in the final model, with 91% of the backbone dihedral angles in the most favored regions of the Ramachandran plot. Seventy-

two ordered water molecules were included in the model. Further details are provided in the Materials and Methods section.

Structure

The crystal structure of *T. maritima* CelA revealed a β -sandwich fold, comprised of two β -sheets packed against each other, as observed in other related endoglucanases. The closest homolog of known structure, with a sequence identity of just 21.4% over 224 residues, is Cel12A from the thermophilic microbe, *Rhodothermus marinus*,³² hereafter referred to as RmCel12. The *T. maritima* structure is the second representative from an extreme thermophilic organism for this family of enzymes. A slightly more divergent structure with 18.4% sequence identity over 223 residues has been obtained from a moderately thermophilic organism, *Bacillus licheniformis* (BIXG12).³³ Structurally, CelA aligns well to those previous structures, despite the low sequence identity. The root mean square deviations (RMSD) to CelA are 2.15 Å over 224 C α atoms for RmCel12 and 2.19 Å over 223 C α atoms for BIXG12. These two structures also have divergent sequences, with 29.5% sequence identity to each other and an RMSD of 1.71 Å over 220 C α atoms. Strands B8 and B9 are unique to the structure of CelA and twist so they connect with and extend sheet A, forming its seventh and eighth strands.

A previous study compared structural features likely to contribute to the thermostability of RmCel12 and other family 12 cellulases.³² *T. maritima* CelA has higher thermostability than those proteins, with a half-life of over 48 hours at 90°C in 2.5 M NaCl.³⁴ The protein shares many of the distinguishing features of the thermophilic RmCel12. Compared to these proteins,³² *T. maritima* CelA has shorter loops and longer β -strands on average, more hydrophobic residues, more charged residues, fewer Asn/Gln residues, and no internal cavities. Unlike the other homologs, CelA is an intracellular protein³⁵ with no signaling sequence for membrane translocation, and its termini are structurally well-ordered. These features may account for the high thermostability of CelA even in the absence of disulfide bonds. Cystines are present in different locations in most family 12 endoglucanases, including thermostable homologs, and are therefore likely needed for local stabilization of their structures.³⁶

The structure of CelA reveals a deep and extended binding cleft [Fig. 2(A)]. CelA is an endoglucanase that cleaves β -1,4 linked polysaccharides.³⁴ A homologous structure for BIXG12 revealed an extensive cellulose binding site with substrate bound.³³ Similarly, the cleft of CelA spans over 35 Å across the surface of the protein, and the shape of the groove is suggestive of how branched polysaccharides could bind, though structures of the enzyme with ligands bound will be required to define

Figure 2. The binding cleft of CelA. (A) The disulfide-crosslinked CelA dimer is shown in a molecular surface representation. To emphasize surface relief, the model is colored by diffusion accessibility,^{37,38} with lowest diffusion accessibility colored blue. A HEPES buffer molecule present in a homologous structure of RmaCel12 has been modeled in the prominent cleft where CelA binds and cleaves cellulosic polymers. (B) *T. maritima* CelA (cyan) is shown aligned to the structure of the thermophilic homologue RmaCel12 (gray) in its active conformation. Amino acid residues presumed to be important in binding and catalysis, E135, W179, E117, E232, and N113, are shown in stick representation. The HEPES molecule present in the RmaCel12 structure, which is suggested to mimic a glucoside substrate, is shown in gray. Crystal structures of complexes containing bona fide substrates are not yet available for close homologues of CelA. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

substrate binding modes precisely. CelA was previously determined to have specific activities of 1785 U/mg for barley β -glucan, 302 U/mg for CM-cellulose, and 2.5 U/mg for Oat speltis xylan.³⁴ The cleft of CelA is spacious enough, between 15 and 20 Å in depth, to accommodate a substrate and clamp around it in its active conformation. The positions of key residues believed by sequence comparisons to be catalytic in CelA are the acid-base E232 and the nucleophile E135, which are present in the expected positions in the active site [Fig. 2(B)]. The catalytic residues align well to RmaCelA with an RMSD of 0.23 Å over 38 atoms, despite the differences in the shape of the CelA active site from that of other endoglucanases.

Lattice Contacts

Dimerized CelA crystallizes with two protein chains (i.e., one complete dimer) in the asymmetric unit. The two monomers have nearly identical conformations, with an RMSD of 0.27 Å over all atoms; non-crystallographic symmetry restraints were not applied during refinement. The engineered disulfide bond is well-ordered and visible in electron density maps [Fig. 3(A)]. It is remarkable that the two monomers in the dimer have practically no contact with each other, aside from the disulfide bond that connects them together. The space between the molecules is presumably filled with solvent; four ordered water molecules in the interface could be visualized [Fig. 3(B)]. Only 35 protein atoms (including atoms in the cysteine residue) from one monomer are within 5 Å (center-to-center) of

another atom in the opposing monomer. Therefore, the interaction between disulfide-bonded molecules in the crystal is unlike typical crystal contacts. Many more atoms are typically involved in a crystal contact. In the present crystal, seven other distinct sets of protein–protein contacts that form the crystal lattice have 194, 386, 103, 206, 592, 194, and 103 atoms within 5 Å of an atom in an adjacent molecule. However, the covalent disulfide bond between subunits of the synthetic CelA dimer apparently provides a critical stabilizing interaction in the lattice, despite the paucity of additional noncovalent interactions. Cleavage of the disulfide bond by soaking crystals in β -ME led to their disintegration within about 10 minutes. The slow speed of the disintegration could reflect a limited rate of diffusion due to the high viscosity of the MPD solution used for crystallization.

The two monomers in the synthetic CelA dimer are arranged in a nearly symmetric fashion. They are related to each other by a rotation of 175°, with a 2.5 Å shift along the axis of rotation. The dimer therefore comes close to obeying C2 symmetry. This is remarkable given that the dimer is a nonbiological construct, and a symmetric arrangement is not enforced by the crystal symmetry (i.e., the asymmetric unit of the crystal contains both molecules). The near-symmetric arrangement could be promoted by the discrete torsion angles for the cysteine side chain within the context of a relatively rigid segment of the protein. The dihedral angle of the disulfide bond is 98° and off by 5° from the ideal angle of 103°. A reorientation by 5° of the local rotation

Figure 3. The disulfide bond and crystal packing of the engineered CelA dimer. (A) A $2F_o - F_c$ electron density map contoured at 1.5σ . The two subunits are connected by a disulfide bond (between Cys 188 residues) having nearly ideal geometry. (B) The interface between disulfide-bonded monomers has very few atomic contacts other than those involving the engineered cysteine residue. Waters and bromide ions lining the interface are represented as spheres colored red and grey, respectively. (C) A stereoview of the crystal packing of the D188C dimer. The synthetic dimer occurs in the asymmetric unit of the crystal in a nearly symmetric C2 arrangement. The view is centered along this local axis of $\sim$ twofold symmetry axis (indicated by the dyad symbol). A second local symmetry axis is shown in black. These symmetry axes pass through the engineered disulfide bonds that link monomers together. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

axis that relates the two monomers would bring that axis in line with the crystal a axis [Fig. 3(C)], which would be a symmetry axis in the higher crystal space group $P6_322$. Instead of conforming to exact twofold symmetry, a slight shift from a purely symmetric conformation allows the subunits of dimeric CelA to make the contacts needed to form an ordered crystal lattice in space group $P6_3$.

Discussion

Creating an artificially symmetric version of a macromolecule can aid in crystallization. The present work demonstrates a successful application of that notion for a case in which the crystal structure of the protein was not already known, although structures of distantly related homologues had been elucidated. Here, we employed disulfide bonding for symmetrization, following up an earlier demonstration of the same method using lysozyme as a model test system for crystallization.²⁴ Beyond the disulfide approach implemented here, recent studies have

highlighted the utility of other methods for designing self-complementary surface patches as a crystallization strategy.^{25,26} Recent successes in designing intermolecular metal binding interfaces offer another strategy for promoting symmetric associations,^{40,41} and this may represent yet another approach for promoting crystallization in the future.

The multiple proteins and design trials undertaken here illuminate limitations of the symmetrization method as we have currently implemented it. First, proteins that are naturally oligomeric may not be easy targets for the method. In the absence of structural knowledge, designing naïve dimerization motifs into oligomeric proteins is likely to lead to extended, irregular assemblies not conducive to crystallization. For any given oligomer, a fortuitous choice for cysteine positions might lead to a single defined species comprised of two copies of the natural oligomer, but heterogeneous assemblies are likely to prevail. It is possible that suitable dimerization sites could have been obtained for some of the

Table I. Data Collection and Refinement Statistics

	Se-Met peak	Native
Data collection		
Space group	P6 ₃ (a = b = 114.51, c = 98.48)	P6 ₃ (a = b = 114.03, c = 98.82)
Beamline/generator	APS 24-ID-E	APS 24-ID-C
Resolution (Å)	99.22-3.00	98.75-2.40
Wavelength (Å)	0.9791	0.9794
Total reflections	85156	161171
Unique reflections	14767	27835
Redundancy	5.8	5.8
Completeness (%) ^a	99.8 (100)	97.0 (85.7)
$R_{\text{merge}}(\%)^{\text{a,b}}$	10.6 (51.2)	8.4 (45.0)
$I/\sigma(I)^{\text{a}}$	11.4 (2.4)	5.6 (2.0)
Refinement statistics		
Resolution (Å)	99.2-3.0	98.8-2.4
Number of atoms		
Protein		4316
Bromide		7
Water		72
$R_{\text{cryst}}(\%)^{\text{c}}$		17.4
$R_{\text{free}}(\%)^{\text{c}}$		21.5
Average B value (Å ²)		
chain A		49.7
chain B		42.8
waters		40.6
Ramachandran diagram (% in most favored region)		91%
RMSD from ideal values		
Bonds (Å)		0.009
Angles (°)		1.10

^a Numbers in parentheses represent data in the highest resolution shell (3.00–3.11 Å for the Se-Met derivative, and 2.40–2.49 Å for the native).

^b $R_{\text{merge}}(I) = \sum_{hkl} (\sum_i |I_{hkl,i} - [I_{hkl}]|) / \sum_i I_{hkl,i}$.

^c $R_{\text{cryst}} = \sum_{hkl} |F_{\text{obs}}| - |F_{\text{calc}}| / \sum_{hkl} |F_{\text{obs}}|$. R_{free} was computed identically, except that 5% of the reflections were omitted as a test set.

original protein targets in this study that appeared to be oligomeric. Further studies would be required to evaluate the general feasibility of obtaining discrete oligomeric species when dimerizing naturally oligomer proteins. This is an important limitation in view of the abundance of symmetric homooligomeric protein assemblies in nature. Screening proteins at an early stage to determine native molecular weight and to evaluate monodispersity would be important in choosing suitable targets for future applications of the synthetic symmetrization approach. Size exclusion chromatography, light scattering, and ultracentrifugation are all useful in this regard. While the common occurrence of natural oligomers presents a limitation in applying the strategy to proteins, we note that, compared to proteins, nucleic acids such as RNA molecules are much less often oligomeric, and so could benefit more readily from the artificial symmetrization approach. Methods for symmetrizing nucleic acid molecules could include the introduction of self-complimentary (palindromic) sequence segments, or the design of oligomeric nucleic acid binding proteins. Monomeric RNA binding proteins have already been employed to improve RNA crystallization.¹⁸

Predicting solvent accessible sites for cross-linking presents another challenge. Only a minority of

the sites chosen for cysteine insertion in the CelA protein were able to form cross-linked dimers in high yield. A retrospective analysis showed that the reactivity correlated strongly with the solvent accessibility at the residue position in question. It remains a challenge to determine *ab initio* which residue positions are likely to be highly exposed to solvent, yet not so disordered as to lead to a highly flexible protein dimer that might be difficult to crystallize. With regard to selecting inflexible regions, we retrospectively evaluated whether charged residues adjacent to prolines (which was the basis for some of the mutation sites chosen in this study) are less flexible than charged residues overall. In a survey of the PDB, the average temperature factor for a charged residues preceding a proline was only marginally lower (35.5 vs. 36.1 Å²), suggesting that adjacency to proline is likely to provide a minimal advantage. With regard to solvent exposure of side chains in such a motif, a survey showed that adjacency to proline neither hindered nor enhanced the average solvent exposure; in either case, about 72% of the charged side chains met a criterion of having at least 15% of their total area (including the backbone) exposed. The challenges in predicting exposed but well-ordered sites for dimerization emphasizes the value in pursuing multiple sites for dimerization

simultaneously to maximize the chances of success. The potential complication of reactive cysteines in native protein sequences presents a final limitation. Among proteins of moderate size – about 200 amino acids – a survey of known structures indicates that about 25% have no native cysteines, while an additional 20% have a single cysteine; in the earlier test of synthetic symmetrization using lysozyme, the single native cysteine was replaced by mutagenesis with favorable results.²⁴

The crystal lattice observed for the D188C CelA dimer highlights the behavior of artificially disulfide-linked protein dimers. Within the crystal lattice, the two monomers of the synthetic dimer make almost no contact with each other aside from the disulfide bond. This situation contrasts greatly with typical nonbonded interfaces in protein crystals. The connection between the two monomers of CelA appears so tenuous that we investigated whether the disulfide bond was a necessary part of the crystal. Visually, we examined the packing arrangement in the crystal to assess whether the disulfide bond was an absolute requirement for all the molecules in the crystal to be connected to each other. The requirement for connectivity between protein molecules in a crystal has been emphasized in earlier work.⁴² Interestingly, the disulfide bond is not a requirement for molecular connectivity in the CelA crystal; among the protein molecules that form nonbonded crystal contacts with monomer A and those that form nonbonded crystal contacts with monomer B, some form contacts with each other. In that sense, the disulfide bond between the two monomers might be viewed as a superfluous contact. However, we found experimentally that the disulfide bond was necessary for the stability of the crystal. That observation emphasizes the key point that the crystal form obtained using dimerized CelA could not have been obtained using the native monomeric protein. Dimerization therefore provided a new opportunity for crystallization. We note that whether dimerization of the mutant was a requirement for crystallization was not tested explicitly, for example, by performing crystallization experiments under reducing conditions to see if crystal forms other than the one formed by the dimer might be possible.

The two CelA monomers in the asymmetric unit of the crystal are not strictly related by twofold symmetry. The angle of rotation between subunits is roughly 175°. In previous experiments symmetrizing T4 lysozyme, three of the six distinct crystal forms showed artificial dimers with strict symmetry imposed by the lattice symmetry, while the other three novel crystal forms showed rotations between subunits of 176, 177, and 180°. In keeping with the observations made earlier in the lysozyme experiments, the CelA dimer deviates only modestly (i.e.,

by a 5° rotation, and 2.5 Å translation) from strict symmetry. While this tendency towards symmetry seems notable in view of the artificial nature of the dimers that were created, a relatively simple geometric consideration might explain the trend. A recent geometric modeling study examined how identical objects tend to self-associate.⁴³ In the model system employed there, when two identical molecules are held together on a perfectly flat face by a connection between identical points in the two molecules – for example, the unique cysteine residues here – such an arrangement will always obey a twofold axis of symmetry.⁴³ Though deviations from 180° symmetry are permitted when two molecules are attached at identical points that are not on perfectly flat faces of the molecules, the observation of strictly symmetric or nearly symmetric arrangements of molecules in the synthetic protein dimers studied so far is perhaps not so surprising in view of this geometric consideration.

Notwithstanding certain basic tenets, such as the adverse effect flexibility has on crystallization, it remains unclear why some proteins are so much harder to crystallize than others. It is therefore difficult to understand fully how various approaches to improving crystallization exert their effects. For the artificial symmetrization approach discussed here, the added symmetry may be an incremental advantage, though the main benefit to crystallization likely arises from the independent variations on the molecule that arise from the varied constructions. Each specific (dimerized) construct provides opportunities for an entirely distinct set of molecular contacts that could be consistent with a crystal lattice. Likewise, one can imagine that certain natural protein molecules tend to contact themselves in ways that are inconsistent with crystallographic arrangements. Different symmetrized versions of a molecule could mask such unproductive interactions and thereby favor other interactions more consistent with allowable crystallographic symmetries. Further development of methods for rapidly diversifying the surface properties or interaction modes of protein molecules should continue to improve the outlook for crystallizing recalcitrant macromolecules.

Materials and Methods

Cloning and plasmids

The original, wild-type *T. maritima* genes were cloned into vector pMH1 (encodes for a amino-terminal MGSDKIHSHHHHH expression/purification tag) as previously described.² Purified plasmid DNAs from these clones served as templates for incorporating single-site cysteines using the PIPE mutagenesis method.⁴⁴ Six proteins from *T. Maritima* were mutated at the following positions to generate constructs for synthetic symmetrization:

TM0227: R4, E9, Q12, N26, E30, E45, E50, R57
TM0252: K6, E13, R25, K30, E40, D37, D75, K78
TM0733: K17, R20, D25, Q28, D34, E41, K53,
R113
TM0829: E21, K25, Q76, N80, K83, D111, R115,
R141
TM1059: E2, K14, K28, Q41, K44, K75, K80
TM1524: K7, E31, K57, E150, D188, H201,
D204, K208

In purification experiments on CelA D188C, the protein bound weakly to a Ni-affinity column, so a new construct was generated with the 6xHis tag at the C-terminus by amplifying from the pMH1F-CelA vector with the primers:

5'-ATTTATCATATGGTGGTACTGATGACATGCC
CG-3' and

5'-TTGGATCCTCAGTGGTGGTGGTGGTGGTGT
TCTCTCACCTCCAGATCAATAG-3'.

The PCR product was cloned into pET22b. The resulting construct pET22-CelA encoded the D188C mutant with a noncleavable 6xHis tag at the C-terminus.

Purification and dimerization

The expression vector pET22-CelA was transformed into Rosetta cells (Novagen). Growth and lysis of cells expressing CelA has been described previously.³⁴ Seleno-Met derivative protein was prepared following a protocol also described previously.⁴⁵ The supernatant of the lysed cells was then filtered through a 0.22 μ m filter and loaded onto a Histrap column (GE Healthcare), washed with lysis buffer, then buffer with 70 mM Imidazole, and finally eluted using a gradient up to 0.5 M Imidazole. DTT was added to the protein to a final concentration of 10 mM, concentrated using an Amicon centrifugal filter with a 10 kDa cutoff (Millipore), then purified over the S75 10/300 size exclusion column (GE Healthcare) in batches of 500 μ L. A buffer containing 20 mM Tris pH 8.0, 100 mM NaCl was used. The fractions containing CelA were dimerized with 1 mM CuSO₄ and were incubated at 22°C for 30 minutes. This oxidation step was quenched by the addition of EDTA to a final concentration of 10 mM. Dimers were buffer exchanged into 20 mM MOPS pH 7.0 and 50 mM NaCl and concentrated using an Amicon with a 30 kDa cutoff.

Crystallization

The protein was set up in mosquito trays at 4 mg/mL with the Hampton Index screen. The protein crystallized in 0.1 M Tris HCl, pH 8.5, 0.2 M NH₄C₂H₃O₂, and 45% MPD. This condition was optimized with Linbro plates by varying the concentration of MPD, salt, and the pH. The optimal concentration found was 42% MPD, 0.2 M 2NH₄C₂H₃O₂, and 0.1 M Tris HCl, pH 8.3. Crystals of the protein in this condition were transferred to fresh mother

liquor solution, crushed and diluted with more mother liquor by a factor of 10⁵ to form a dilute seed stock. Fresh seed stock was used to nucleate growth of additional crystals, including those formed using NaBr instead of NH₄C₂H₃O₂, and those of the selenomethionine derivative.

Data collection

The selenomethionine derivative data were collected at the Advanced Photon Source on the microfocal beamline 24-ID-E at a wavelength of 0.9791 Å. The phases obtained from this dataset were applied to native diffraction data from a crystal grown in 42% MPD, 0.35 M NaBr, and 0.1 M Tris HCl, pH 8.3.

Data Processing

Reflections for the selenomethionine derivative crystal out to 3.0 Å were integrated and scaled with DENZO/SCALEPACK.⁴⁶ The space group was determined to be P6₃ based on symmetry and systematic absences. An anisotropy correction was applied to the reflections before scaling.⁴⁷ All 24 of the expected selenium sites were identified and used for phasing using Phenix.autosolve,⁴⁸ and Phenix.autobuild was used to find 90% of the protein's residue positions. The model from Phenix.autobuild provided phases that were applied to a 2.4 Å dataset from a protein without selenomethionines. The positions of remaining atoms were placed using Coot,⁴⁹ and the structure was refined initially with Phenix.refine and finally with BUSTER-TNT.⁵⁰ We judged that noncrystallographic symmetry restraints between the two molecules in the asymmetric unit were unnecessary during the final refinement, as the R_{free} value remained low at 20.8%.

Model Analysis

Structural homologues were found using DALI.⁵¹ Homologous models were aligned to CelA with TM-Align.⁵² The packing density was assessed with VOIDOO⁵³ using a 1.2 Å probe. The secondary structure and surface accessibility was defined using DSSP.²⁷

Accession numbers

The coordinates and structure factors for CelA (PDB ID 3O7O) have been deposited in the Protein Data Bank (www.pdb.org).

Acknowledgments

The authors thank Lukasz Goldschmidt for assistance with sequence analysis and prediction of surface exposure, Marcin Apostol for assistance with mass spectrometry, the staff at APS beamline 24-ID-E for X-ray data collection, the JCSG Crystallomics Core for cloning and protein expression, and Ian Wilson for a careful review of the manuscript.

References

- Slabinski L, Jaroszewski L, Rodrigues APC, Rychlewski L, Wilson IA, Lesley SA, Godzik A (2007) The challenge of protein structure determination—lessons from structural genomics. *Protein Sci* 16: 2472–2482.
- Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreuzsch A, Spraggon G, Klock HE, McMullan D, Shin T, Vincent J, Robb A, Brinen LS, Miller MD, McPhillips TM, Miller MA, Scheibe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger M, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC (2002) Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proc Natl Acad Sci USA* 99: 11664–11669.
- Dong A, Xu X, Edwards AM, Chang C, Chruszcz M, Cuff M, Cymborowski M, Di Leo R, Egorova O, Evdokimova E, Filippova E, Gu J, Guthrie J, Ignatchenko A, Joachimiak A, Klostermann N, Kim Y, Korniyenko Y, Minor W, Que Q, Savchenko A, Skarina T, Tan K, Yakunin A, Yee A, Yim V, Zhang R, Zheng H, Akutsu M, Arrowsmith C, Avvakumov GV, Bochkarev A, Dahlgren L, Dhe-Paganon S, Dimov S, Dombrowski L, Finerty P, Flodin S, Flores A, Gräslund S, Hammerström M, Herman MD, Hong B, Hui R, Johansson I, Liu Y, Nilsson M, Nedyalkova L, Nordlund P, Nyman T, Min J, Ouyang H, Park H, Qi C, Rabeh W, Shen L, Shen Y, Sukumard D, Tempel W, Tong Y, Tresagues L, Vedadi M, Walker JR, Weigelt J, Welin M, Wu H, Xiao T, Zeng H, Zhu H (2007) In situ proteolysis for protein crystallization and structure determination. *Nat Methods* 4: 1019–1021.
- Jaroszewski L, Slabinski L, Wooley J, Deacon AM, Lesley SA, Wilson IA, Godzik A (2008) Genome pool strategy for structural coverage of protein families. *Structure* 16: 1659–1667.
- Kendrew JC, Parrish RG, Marrack JR, Orlans ES (1954) The species specificity of myoglobin. *Nature* 174: 946–949.
- Thornton JM, Sibanda BL (1983) Amino and carboxy-terminal regions in globular proteins. *J Mol Biol* 167: 443–460.
- Pantazatos D, Kim JS, Klock HE, Stevens RC, Wilson IA, Lesley SA, Woods VL (2004) Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. *Proc Natl Acad Sci USA* 101: 751–756.
- Rypniewski WR, Holden HM, Rayment I (1993) Structural consequences of reductive methylation of lysine residues in hen egg white lysozyme: an X-ray analysis at 1.8-Å resolution. *Biochemistry* 32: 9851–9858.
- Walter TS, Meier C, Assenberg R, Au K, Ren J, Verma A, Nettleship JE, Owens RJ, Stuart DI, Grimes JM (2006) Lysine methylation as a routine rescue strategy for protein crystallization. *Structure* 14: 1617–1622.
- Kim Y, Quartey P, Li H, Volkart L, Hatzos C, Chang C, Nocek B, Cuff M, Osipiuk J, Tan K, Fan Y, Bigelow L, Maltseva N, Wu R, Borovilos M, Duggan E, Zhou M, Binkowski TA, Zhang R, Joachimiak A (2008) Large-scale evaluation of protein reductive methylation for improving protein crystallization. *Nat Methods* 5: 853–854.
- Longenecker KL, Garrard SM, Sheffield PJ, Derewenda ZS (2001) Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. *Acta Cryst D* 57: 679–688.
- Derewenda ZS (2010) Application of protein engineering to enhance crystallizability and improve crystal properties. *Acta Cryst D* 66: 604–615.
- Ostermeier C, Iwata S, Ludwig B, Michel H (1995) Fv fragment-mediated crystallization of the membrane protein bacterial cytochrome c oxidase. *Nat Struct Biol* 2: 842–846.
- Kovari LC, Momany C, Rossmann MG (1995) The use of antibody fragments for crystallization and structure determinations. *Structure* 3: 1291–1293.
- Lange C, Hunte C (2002) Crystal structure of the yeast cytochrome bc₁ complex with its bound substrate cytochrome c. *Proc Natl Acad Sci USA* 99: 2800–2805.
- Lee S, Lee A, Chen J, MacKinnon R (2005) Structure of the KvAP voltage-dependent K⁺ channel and its dependence on the lipid membrane. *Proc Natl Acad Sci USA* 102: 15441–15446.
- Koide S (2009) Engineering of recombinant crystallization chaperones. *Curr Opin Struct Biol* 19: 449–457.
- Ferré-D'Amaré AR, Zhou K, Doudna JA (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature* 395: 567–574.
- Sennhauser G, Grütter MG (2008) Chaperone-assisted crystallography with DARPins. *Structure* 16: 1443–1453.
- Privé GG, Verner GE, Weitzman C, Zen KH, Eisenberg D, Kaback HR (1994) Fusion proteins as tools for crystallization: the lactose permease from *Escherichia coli*. *Acta Cryst D* 50: 375–379.
- Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B (2003) Crystal structures of fusion proteins with large-affinity tags. *Protein Sci* 12: 1313–1322.
- Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi H, Yao X, Weis WI, Stevens RC, Kobilka BK (2007) GPCR engineering yields high-resolution structural insights into beta₂-adrenergic receptor function. *Science* 318: 1266–1273.
- Nauli S, Farr S, Lee Y, Kim H, Salem F, Bowie JU (2007) Polymer-driven crystallization. *Protein Sci* 16: 2542–2551.
- Banatao DR, Cascio D, Crowley CS, Fleissner MR, Tienson HL, Yeates TO (2006) An approach to crystallizing proteins by synthetic symmetrization. *Proc Natl Acad Sci USA* 103: 16230–16235.
- Yamada H, Tamada T, Kosaka M, Miyata K, Fujiki S, Tano M, Moriya M, Yamanishi M, Honjo E, Tada H, Ino T, Yamaguchi H, Futami J, Seno M, Nomoto T, Hirata T, Yoshimura M, Kuroki R (2007) Crystal lattice engineering, an approach to engineer protein crystal contacts by creating intermolecular symmetry: crystallization and structure determination of a mutant human RNase 1 with a hydrophobic interface of leucines. *Protein Sci* 16: 1389–1397.
- Guo M, Shapiro R, Schimmel P, Yang X (2010) Introduction of a leucine half-zipper engenders multiple high-quality crystals of a recalcitrant tRNA synthetase. *Acta Cryst D* 66: 243–250.
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637.
- Grueninger D, Treiber N, Ziegler MOP, Koetter JWA, Schulze M, Schulz GE (2008) Designed protein-protein association. *Science* 319: 206–209.
- Derewenda ZS (2004) Rational protein crystallization by mutational surface engineering. *Structure* 12: 529–535.
- Baud F, Karlin S (1999) Measures of residue density in protein structures. *Proc Natl Acad Sci USA* 96: 12494–12499.
- Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292: 195–202.

32. Crennell SJ, Hreggvidsson GO, Nordberg Karlsson E (2002) The structure of *Rhodothermus marinus* Cel12A, a highly thermostable family 12 endoglucanase, at 1.8 Å resolution. *J Mol Biol* 320: 883–897.
33. Gloster TM, Ibatullin FM, Macauley K, Eklöf JM, Roberts S, Turkenburg JP, Bjørnvad ME, Jørgensen PL, Danielsen S, Johansen KS, Borchert TV, Wilson KS, Brumer H, Davies GJ (2007) Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12. *J Biol Chem* 282: 19177–19189.
34. Liebl W, Ruile P, Bronnenmeier K, Riedel K, Lottspeich F, Greif I (1996) Analysis of a *Thermotoga maritima* DNA fragment encoding two similar thermostable cellulases, CelA and CelB, and characterization of the recombinant enzymes. *Microbiol* 142: 2533–2542.
35. Bauer MW, Driskill LE, Kelly RM (1998) Glycosyl hydrolases from hyperthermophilic microorganisms. *Curr Opin Biotechnol* 9: 141–145.
36. Sandgren M, Shaw A, Ropp TH, Wu S, Bott R, Cameron AD, Ståhlberg J, Mitchinson C, Jones TA (2001) The X-ray crystal structure of the *Trichoderma reesei* family 12 endoglucanase 3, Cel12A, at 1.9 Å resolution. *J Mol Biol* 308: 295–310.
37. Yeates TO (1995) Algorithms for evaluating the long-range accessibility of protein surfaces. *J Mol Biol* 249: 804–815.
38. Tsai Y (2009) Analysis of carboxysome shell proteins and crystal defects, Ph.D dissertation. University of California, Los Angeles, CA, USA.
39. Dani VS, Ramakrishnan C, Varadarajan R (2003) MODIP revisited: re-evaluation and refinement of an automated procedure for modeling of disulfide bonds in proteins. *Protein Eng* 16: 187–193.
40. Salgado EN, Ambroggio XI, Brodin JD, Lewis RA, Kuhlman B, Tezcan FA (2010) Metal templated design of protein interfaces. *Proc Natl Acad Sci USA* 107: 1827–1832.
41. Salgado EN, Radford RJ, Tezcan FA (2010) Metal-directed protein self-assembly. *Acc Chem Res* 43: 661–672.
42. Wukovitz SW, Yeates TO (1995) Why protein crystals favour some space-groups over others. *Nat Struct Biol* 2: 1062–1067.
43. Schulz GE (2010) The dominance of symmetry in the evolution of homo-oligomeric proteins. *J Mol Biol* 395: 834–843.
44. Klock HE, Koesema EJ, Knuth MW, Lesley SA (2008) Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins* 71: 982–994.
45. Van Duyne GD, Standaert RF, Karplus PA, Schreiber SL, Clardy J (1993) Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J Mol Biol* 229: 105–124.
46. Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276: 307–325.
47. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D (2006) Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 103: 8060–8065.
48. Adams PD, Mustyakimov M, Afonine PV, Langan P (2009) Generalized X-ray and neutron crystallographic analysis: more accurate and complete structures for biological macromolecules. *Acta Cryst D* 65: 567–573.
49. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Cryst D* 66: 486–501.
50. Bricogne G, Blanc E, Brandl M, Flensburg C, Keller P, Paciorek P, Roversi P, Sharff A, Smart O, Vonnrhein C, Womack T (2009). BUSTER version 2.8.0. Cambridge, United Kingdom: Global Phasing Ltd.
51. Holm L, Kääriäinen S, Rosenström P, Schenkel A (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics* 24: 2780–2781.
52. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33: 2302–2309.
53. Kleywegt GJ, Jones TA (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Cryst D* 50: 178–185.