*Protein Engineering, Design & Selection*, 2015, vol. 28 no. 11, pp. 491–499
doi: 10.1093/protein/gzv035
Advance Access Publication Date: 4 August 2015
Original Article

OXFORD

Original Article

# On the predictability of the orientation of protein domains joined by a spanning alpha-helical linker

**Yen-Ting Lai[1,4], Lin Jiang[2], Wuyang Chen[1,5], and Todd O. Yeates[1,3,*]**

[1]UCLA-DOE Institute for Genomics and Proteomics, [2]Department of Neurology, UCLA-Easton Center of Alzheimer's Disease Research, and [3]Department of Chemistry and Biochemistry, UCLA, 611 Charles Young Dr. East, Los Angeles, CA 90095-1569, USA

*To whom correspondence should be addressed. E-mail: yeates@mbi.ucla.edu
[4]Present address: The Biodesign Institute, Arizona State University, USA.
[5]Present address: Center for Theoretical Biological Physics, Rice University, USA.

Edited by David Eisenberg

Received 29 June 2015; Revised 29 June 2015; Accepted 8 July 2015

## Abstract

Connecting proteins together in prescribed geometric arrangements is an important element in new areas of biomolecular design. In this study, we characterize the degree of three-dimensional orientational control that can be achieved when two protein domains that have alpha-helical termini are joined using an alpha-helical linker. A fusion between naturally oligomeric protein domains was designed in this fashion with the intent of creating a self-assembling 12-subunit tetrahedral protein cage. While the designed fusion protein failed to assemble into a tetrahedral cage in high yield, a series of crystal structures showed that the two fused components were indeed bridged by an intact alpha helix, although the fusion protein was distorted from the intended ideal configuration by bending of the helix, ranging from 7 to 35°. That range of deviation in orientation creates challenges for designing large, perfectly symmetric protein assemblies, although it should offer useful outcomes for other less geometrically demanding applications in synthetic biology.

**Key words:** flexibility, protein design, protein fusion, symmetry, synthetic biology

## Introduction

Being able to place protein molecules in specific spatial arrangements opens up possibilities for varied applications in the broad area of synthetic biology (Grunberg and Serrano, 2010; Good *et al.*, 2011; Chen and Silver, 2012; Lee *et al.*, 2012). In some applications, advantages are gained by bringing multiple functionalities, such as sequential enzyme activities, into proximity (Dueber *et al.*, 2009; Delebecque *et al.*, 2011). In other lines of investigation, functional dependencies (such as mutually exclusive folding) are created through careful arrangement of protein subunits and their termini (Ha *et al.*, 2012, 2013). In other applications, complex architectures with interior chambers can be obtained from the assembly of many copies of protein or polypeptide building blocks (Ni and Tezcan, 2010; Worsdorfer *et al.*, 2011; Fletcher *et al.*, 2013; King *et al.*, 2014; Lai *et al.*, 2014).

In applications aimed at creating well-ordered supramolecular architectures, the requirement for precise geometric control over the orientations of multiple protein subunits is particularly acute. Two basic strategies have emerged to satisfy the requirements for joining protein molecules in specific orientations. The first relies on genetic fusion of two protein domains, each of which bears an alpha-helical terminus, using a short alpha-helical linker between them (Padilla *et al.*, 2001; Lai *et al.*, 2012a, 2014). If such a fusion protein folds correctly with an unbroken helix spanning the two original components, then the relative orientation between them can be calculated readily. The second approach relies on the design of a new interface between components using computational methods to suggest amino acid changes in the protein surface or surfaces (Grueninger *et al.*, 2008; Der and Kuhlman, 2013; King *et al.*, 2014). Interfacial metal ions are sometimes also included in such designs (Salgado *et al.*, 2010; Brodin *et al.*, 2012; Der *et al.*, 2012). If such designs produce correctly folded proteins with sufficiently complementary surfaces, then the prescribed orientations can be achieved. Using either of these strategies, surprisingly complex self-assembling architectures can be obtained by choosing naturally oligomeric proteins (Schulz, 2010) (e.g. dimers and trimers) as the individual components to be joined together—by helix fusion or interface design. By combining two symmetries under specific geometric rules, a wide range of highly symmetric architectures are possible (Padilla *et al.*, 2001; Sinclair *et al.*, 2011; Lai *et al.*, 2012b; King and Lai, 2013), with cubic cages or shells comprising one type.

Between the two available strategies for geometrically specific attachment of proteins, the helical fusion approach is less demanding computationally, but recent studies have emphasized that helix flexibility allows for deviations from the intended geometry (Lai *et al.*, 2013). A 12-subunit tetrahedral cage 160 Å in diameter assembled as intended, but deviated in the best case by 7.1 Å root-mean-square deviation (RMSD) (over 5280 $C_\alpha$ atoms) compared with the designed structure (Lai *et al.*, 2012a, 2013; manuscript in preparation). A designed 24-subunit cubic cage structure 230 Å in diameter came within 1.2 Å of the intended design (over 6480 $C_\alpha$ atoms), but polymorphic assemblies including 12-mers and 18-mers were also formed in solution (Lai *et al.*, 2014). In both cases, therefore, crystal structures of large cages designed by the helix fusion method could be obtained, but the effects of helix flexibility were evident.

Our previous crystallographic analyses of helix fusion designs have examined helix flexibility in the context of large assemblies, where many subunits are held together by multiple helical connections, i.e. in highly coupled systems (Lai *et al.*, 2012a, 2013, 2014). In the present study, a protein comprised of a fusion between a dimeric protein and a trimeric protein was intended to form a 12-subunit tetrahedral structure, but failed to do so in high yield. However, multiple crystal structures of the protein in states of partial assembly provided an opportunity to examine the conformational properties of protein domains joined by an artificial alpha-helical linker, with an ultimate aim to improve available strategies for connecting protein components in specific arrangements.

## Materials and methods

### Selection of candidate domains for genetic fusion

All protein dimers and trimers with available structure were downloaded from the PISA database (Krissinel and Henrick, 2007). For practical considerations in gene construction, the chain length for the component proteins was limited to 250 residues. Dimeric and trimeric structures with terminal helix segments longer than 10 residues (with either end of the helix segment located within the terminal 5 residues) were selected as fusion candidates. The secondary structures of all dimeric and trimeric structures were assigned using the STRIDE algorithm (Heinig and Frishman, 2004). At the time of the study, there were 42 trimers with a C-terminal helix and 648 dimers with an N-terminal helix (as required for an N′-trimer-linker-dimer-C′ fusion). For connections in the other order, there were 481 dimers with a C-terminal helix and 51 trimers with an N-terminal helix (for N′-dimer-linker-trimer-C′ fusions). For every pair of candidate trimer and dimer, a helix linker of up to 15 residues was inserted computationally to connect the terminal helix segments for geometry assessment. To identify fusion molecules that would be expected to form 12-mer, tetrahedrally symmetric cages, we searched for fusions that had their dimeric and trimeric symmetry axes nearly intersecting with each other (within 1 Å distance at their closest approach) and forming an angle within ±5° of 54.7° (the angle between the principle direction and the body diagonal of a cube).

### Gene construction, site-directed mutagenesis and protein overexpression and purification

A gene corresponding to the initial 2ARH-3-3KAW design was assembled by recursive PCR. Primer fragments of ~50 nucleotides were designed and codon-optimized by using the DNAWorks server (Hoover and Lubkowski, 2002) and were ordered from IDT technology. After PCR assembly, the gene fragment was inserted into the pET-22b vector through the NdeI and XhoI cutting sites. Gene fragments for the three designs were verified by DNA sequencing. The plasmids for the three designs were then transformed into *Escherichia coli* strain BL21(DE3). A 10 ml aliquot of overnight seed culture was inoculated into 1 l of LB medium supplemented with 100 μg/ml ampicillin, incubated at 37°C for 2 h. The temperature was decreased to 18°C and incubated for 1 h before adding IPTG to a concentration of 0.2 mM for induction. The culture was incubated overnight. The bacterial culture was harvested by centrifugation at 6000 *g* for 15 min. The cell pellet was re-suspended in 50 mM phosphate buffer (pH 8.0), 300 mM NaCl and 10 mM imidazole and lysed by a sonicator (Sonics vibra-cell VCX500). The cell lysate was centrifuged at 16 000 *g* for 30 min and the supernatant was filtered and applied to a His-trap column (GE healthcare). The column was washed with lysis buffer supplemented with 100 mM of imidazole and the protein was then eluted with 300 mM imidazole. The fractions containing the target protein were combined for crystallization experiments.

### Protein crystallization, data collection and data processing

To crystallize the 2ARH-3-3KAW construct, the protein concentration was adjusted to 1.2 mg/ml. The protein sample was then dialyzed in 50 mM Tris pH 8.0, 300 mM NaCl, 10 mM imidazole and 10 mM β-mercaptoethanol at room temperature overnight; this procedure was repeated twice. Protein samples were filtered to remove some precipitate that appeared during the dialysis procedure. The crystal of the 2ARH-3-3KAW design was grown in 0.9 M ammonium tartrate dibasic (pH 7.0) as the sole precipitant. The hanging-drop technique was used to grow the crystals at room temperature and the ratio between the protein sample and reservoir was 2–1 μl, with a total reservoir volume of 500 μl. A solution of 4 M trimethylamine *N*-oxide (TMAO) was used as a cryo-protectant.

The mutated protein variants were expressed and purified in the same way as the original construct, but with differences in the crystallization. The 2ARH-3-3KAW-2.0 protein was adjusted to 3.2 mg/ml by addition of 50 mM Tris (pH 8.0), 300 mM NaCl and 35 mM imidazole after purification; no dialysis was performed before crystallization. Crystals were grown in 1.0 M ammonium phosphate monobasic as the sole precipitant. The hanging-drop technique was used to grow the crystals at room temperature and the ratio between the protein sample and reservoir was 1–1 μl, with a total reservoir volume of 500 μl. A solution of 30% dimethyl sulfoxide was used as a cryo-protectant.

The concentration of the 2ARH-3-3KAW-3.0 protein variant was not adjusted after purification and before crystallization. This protein sample was in the elution buffer (50 mM phosphate buffer, 300 mM NaCl and 300 mM imidazole) and the concentration was ~3 mg/ml. Crystals were grown in 0.1 M Bis–Tris (pH 5.5) and 1.5 M ammonium sulfate. The hanging-drop technique was used to grow the crystals at room temperature and the ratio between the protein sample and reservoir was 2–1 µl, with a total reservoir volume of 500 µl. A solution of 4 M TMAO was used as a cryo-protectant.

Diffraction data for the 2ARH-3-3KAW and 2ARH-3-3KAW-2.0 crystals were collected at the Advanced Photon Source, and the data for 2ARH-3-3KAW-3.0 were collected in house (Rigaku FRE+ with HTC detector). All data were processed by the XDS package (Kabsch, 2010).

### Molecular replacement and structure refinement

Molecular replacement was used to solve the structures of all three crystal forms. The monomers of 2ARH and 3KAW crystal structures were used as searching models. In one case, the 2ARH-3-3KAW-2.0 crystal required the use of the 3KAW dimer as a search model to identify a successful solution. The molecular replacement program Phaser (McCoy *et al.*, 2007) was used to identify and position models in the unit cell. A rigid body refinement was carried out after the molecular replacement solution was identified. Helix linkers, in the form of a standard α-helix, were used to connect the 2ARH and 3KAW domains. Owing to the limited resolutions, the crystal structures of 2ARH-3-3KAW and 2ARH-3-3KAW-3.0 were briefly refined by restrained coordinate refinement (without atomic displacement parameter refinement). The 2ARH-3-3KAW-2.0 crystal diffracted to 2.2 Å and allowed a full refinement. The 2ARH-3-3KAW-2.0 crystal structure was used as a reference model (for local network restraints) during the refinement of 2ARH-3-3KAW-3.0 and led to a superior model. Refinement was carried out in PHENIX (Adams *et al.*, 2010), Refmac (Murshudov *et al.*, 1997) and BUSTER (Bricogne *et al.*, 2011).

### Redesign of surface residues near the helix linker by Rosetta

The Rosetta suite of programs (Leaver-Fay *et al.*, 2011) was used for amino acid sequence redesign, focusing on two regions: residues 196–202 and residues 287–296. The design procedure built side-chain rotamers of all amino acids onto the backbone of the selected regions, and the optimal set of rotamers was identified as those that minimized a full-atom energy function. The final energy was then evaluated as each mutated amino acid was reverted back to its native sequence. We only kept those mutations that made a significant contribution to the overall protein stability. Based on this protocol, three mutations were indicated in the final optimized mutant 2ARH-3-3KAW-3.0. Two mutations locate at the helix linker: A199I introduces a bulky isoleucine side-chain to fill a cavity at the linker region; E200Y avoids the burial of a charged glutamic acid and also introduces a bulky tyrosine side-chain to improve hydrophobic packing. Another mutation A293H occurs near the C-terminus and allows a hydrogen bond with Arg 163.

## Results

### Design of the tetrahedral cage

In this work, the intention was to design a 12-subunit tetrahedral cage by fusing a dimeric protein domain to a trimeric domain with their symmetry axes intersecting at nearly 54.7°. A tetrahedral (12-subunit) architecture is the smallest (lowest order) symmetry within the possible cubic symmetries for forming cages, which made a tetrahedron the natural target for our earlier work on designing self-assembling proteins. We designed and characterized a tetrahedral cage using these principles in earlier work (Padilla *et al.*, 2001) when the PDB was much smaller and offered fewer candidates for achieving ideal fusion geometries. The best candidate we obtained at that time had an angle of 51.7° (a 3° deviation from 54.7°) and a closest distance of 6 Å between the two axes. Structural studies on that construct and its variations not only revealed essentially correct tetrahedral assemblies, but also distortions from the intended design in the range of ~5–20%. In the present study, we sought to construct a more perfect cage with a design more closely matching the ideal geometric requirements by exploiting the larger database of known protein structures now available as building blocks. All available structures of protein dimers and trimers were retrieved from the PISA database (Krissinel and Henrick, 2007). Protein structures without helical termini were first filtered out from this dataset. A computational procedure was then used to combine the dimeric and trimeric components pairwise, joining their ends with an alpha-helical linker up to 15 residues in length. Among the many candidate fusion constructs that had symmetry axes nearly satisfying the required criterion, we identified an interesting target for experimental testing. This target has an N-terminal trimeric domain (PDB ID: 2ARH) and a C-terminal dimeric domain (PDB ID: 3KAW), connected by a three residues helical linker. For clarity, this fusion protein is referred to here as 2ARH-3-3KAW ('PDB ID of N-terminal domain'-'linker length'-'PDB ID of C-terminal domain') (Fig. 1).

The trimeric domain 2ARH is a protein with unknown function from *Aquifex aeolicus* and the dimeric domain 3KAW is a protein with unknown function from *Pseudomonas aeruginosa* (see Supplementary Fig. S1 for their structures). The structure of the dimeric domain is a simple four-helix bundle, which can be easily distinguished from the trimeric domain (Fig. 1B). In the initial design, the linker was designed to be Glu-Glu-Ala (see the 'Materials and methods' section). According to the computed model of the fusion construct, the angle and distance between the symmetry axes of the trimeric and dimeric domains were 57.7° and 0.02 Å (Fig. 1), respectively, making it very close to the ideal 54.7° angle of intersection. A 6-histidine tag (6xHis-tag) was added to the N-terminus to facilitate purification by immobilized metal affinity chromatography. Five residues (three in the trimeric domain and two in the dimeric domain) were mutated to alanine to avoid potential steric clashes in the context of the fusion construct. In some cases, an alanine appeared to offer a potentially stabilizing hydrophobic interaction. The amino acid sequence of this construct is shown in Fig. 2.

The protein was expressed and purified from *E.coli* and analyzed by size exclusion chromatography (SEC) and native polyacrylamide gel electrophoresis. The theoretical molecular weight of the designed monomer was 35.2 kDa, which would make a 12-mer assembly ~422 kDa. The purified protein exhibited polydispersity in assembly as judged by the molecular weight deduced by the SEC chromatogram (Fig. 1C), including a large species consistent with the design (470 kDa) along with a smaller species (155 kDa, which is close to the molecular weight of a tetramer). The two peaks were collected separately and analyzed by SEC again, and this resulted in chromatograms similar to the initial one, suggesting the two populations re-equilibrate in solution. Although discrete assembly states could not be maintained in solution, we turned to crystallization as a potential route for separation, with the hope that one (or more) of the assemblies could be withdrawn from solution into a crystalline state.

### Crystal structure of 2ARH-3-3KAW

One crystal form, in space group $P6_322$, was obtained for the initial design of 2ARH-3-3KAW (Table I). Two fusion protein molecules

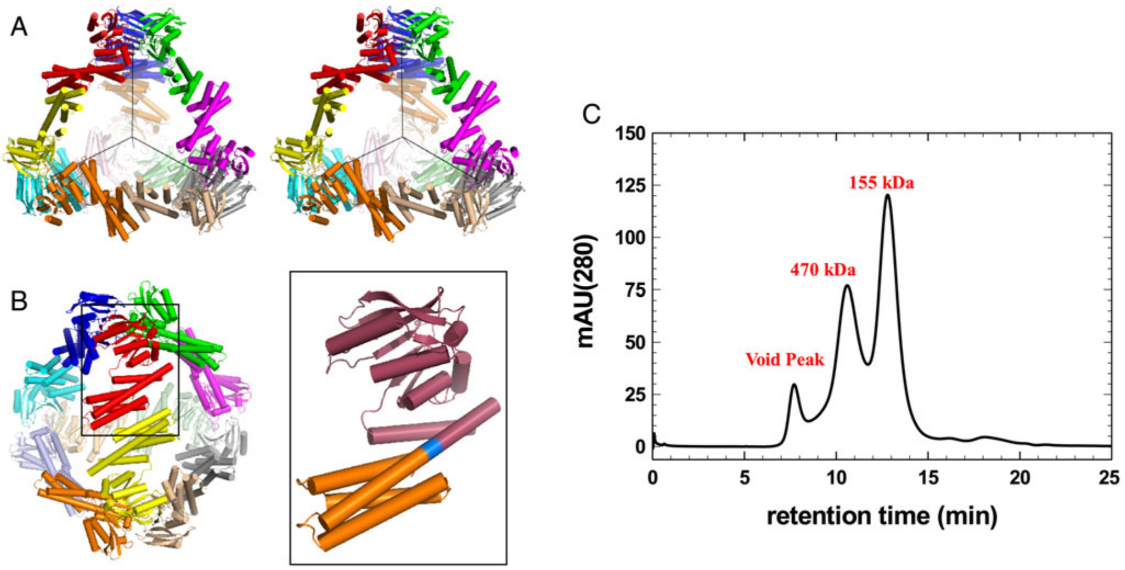

**Fig. 1** (**A**) A stereoview of the intended 12-subunit designed assembly shown roughly along a 3-fold axis of symmetry. The four 3-fold axes are shown as thin black lines inside the cage. The 12 chains forming the cage are shaded differently. (**B**) The cage shown along a 2-fold axis of symmetry (left). One protein subunit is enlarged on the right, where the trimeric domain, the linker and the dimeric domain are shaded differently. (**C**) SEC chromatogram of purified 2ARH-3-3KAW and the calculated molecular weight of the peaks. The theoretical molecular weight of a 12-mer is 422 kDa.

```
WT 2ARH/3KAW    ---------MVKYEELLKTLENGINSEEGEIRLVRKSQGRFKEEFNFDLSLGSKPLLTLK
Design1.0       MHHHHHHDAMVKYEELLKTLENGINSEEGEIRLVRKSQGRFKEEFNFDLSLGSKPLLTLK
Design2.0       ---------MVKYEELLKTLENGINSEEGEIRLVRKSQGRFKEEFNFDLSLGSKPLLTLK
Design3.0       ---------MVKYEELLKTLENGINSEEGEIRLVRKSQGRFKEEFNFDLSLGSKPLLTLK

WT 2ARH/3KAW    VFLGRKPYWQPWVEVFGVNPNLRNVFFGSEAERKLYEFLSEHFGRIFVEYFEDKETTYEL
Design1.0       VFLGRKPYWQPWVEVFGVNPNLRNVFFGSEAERKLYEFLSEHFGRIFVEYFEDKETTYEL
Design2.0       VFLGRKPYWQPWVEVFGVNPNLRNVFFGSEAERKLYEFLSEHFGRIFVEYFEDKETTYEL
Design3.0       VFLGRKPYWQPWVEVFGVNPNLRNVFFGSEAERKLYEFLSEHFGRIFVEYFEDKETTYEL

WT 2ARH/3KAW    QKGVPPALSRLGFELLKLGYTYFRDWFIPEGLMEGGHKIQAEKPKTEEAKKRHLENLKKE
Design1.0       QKGVPPALSRLGFELLKLGYTYFRDWFIPEGLMEGGHKIQAEKPKTAEAKARHLANLKKE
Design2.0       QKGVPPALSRLGFELLKLGYTYFRDWFIPEGLMEGGHKIQAEKPKTAEAKARHLANLKKE
Design3.0       QKGVPPALSRLGFELLKLGYTYFRDWFIPEGLMEGGHKIQAEKPKTAEAKARHLANLKKE

WT 2ARH/3KAW    FEEFIGKCEDEGLIKKVKERYNFLE---EERCRLAAQACIRACERYLALCTESSREQRQH
Design1.0       FEEFIGKCEDEGLIKKVKERYNFLEEEAEERCRLAAQACIRACERYLALCTESSREQRQH
Design2.0       FEEFIGKCEDEGLIKKVKERYNFLEEEAEERCRLAAHHCIHACERYLALCTESSREQRQH
Design3.0       FEEFIGKCEDEGLIKKVKERYNFLEEEIYERCRLAAHHCIHACERYLALCTESSREQRQH

WT 2ARH/3KAW    AGDCADLCRLAALLLERRSPWAPAACELAARYALACAERCDGDEPLERECAGACRRFVEA
Design1.0       AGDCADLCRLAALLLERRSPWAPAACELAARYALACAERCDGDEPLERECAGACRRFVAA
Design2.0       AGDCADLCRLAALLLERRSPWAPAACELAARYALACAERCDGDEPLERECAGACRRFVAA
Design3.0       AGDCADLCRLAALLLERRSPWAPAACELAARYALACAERCDGDEPLERECAGACRRFVAA

WT 2ARH/3KAW    CRPLL
Design1.0       CAPLL
Design2.0       CAPLL
Design3.0       CHPLL
```

**Fig. 2** Sequence alignment of various designs. The short linker is shaded differently. The preceding region is the trimeric domain and the following region is the dimeric domain. The mutations in each design, when compared with the wild-type, are highlighted.

are present in the asymmetric unit of this crystal form. To our surprise, the fusion protein was arranged in the crystal lattice as a tetramer instead of the intended tetrahedral 12-mer assembly (Fig. 3A). Within the tetrameric assembly, the 2-fold interface involving the native dimeric domain (3KAW; the four-helix bundle domain) formed correctly. However, the trimeric domain (2ARH) did not form the previously

**Table I.** Crystallographic data

| Design | 2ARH-3-3KAW | 2ARH-3-3KAW-2.0 | 2ARH-3-3KAW-3.0 |
|---|---|---|---|
| Space group | $P6_322$ | $R3$ | $P3_121$ |
| Unit cell dimensions | 191.61, 191.61, 114.69 | 121.43, 121.43, 207.82 | 112.93, 112.93, 150.01 |
| Resolution (Å) | 95.81–4.20 | 93.83–2.2 | 46.49–4.25 |
| Measured reflections | 183546 | 338449 | 43354 |
| Unique reflections | 9491 | 57982 | 8061 |
| Completeness | 99.9% (99.1%) | 99.9% (98.7%) | 98.9% (94.6%) |
| Rsym | 6.3% (51.0%) | 7.4% (65.1%) | 11.5% (80.8%) |
| $I/\sigma(I)$ | 35.75 (7.55) | 13.81 (3.00) | 12.66 (2.03) |
| Refinement | | | |
| Asymmetric unit | 2 molecules | 2 molecules | 2 molecules |
| Matthews coefficient | 4.31 | 4.32 | 4.03 |
| Solvent content | 71.5% | 71.5% | 69.5% |
| Data used for refinement | 8541 | 55024 | 7654 |
| Data used for $R_{\mathrm{free}}$ | 949 | 2958 | 402 |
| Final $R_{\mathrm{work}}$ | 0.333 | 0.228 | 0.266 |
| Final $R_{\mathrm{free}}$ | 0.376 | 0.249 | 0.288 |
| RMSD | | | |
| Bonds (Å) | 0.008 | 0.013 | 0.004 |
| Angles (°) | 1.033 | 1.443 | 0.832 |
| PDB ID | 4ZSV | 4ZSX | 4ZSZ |

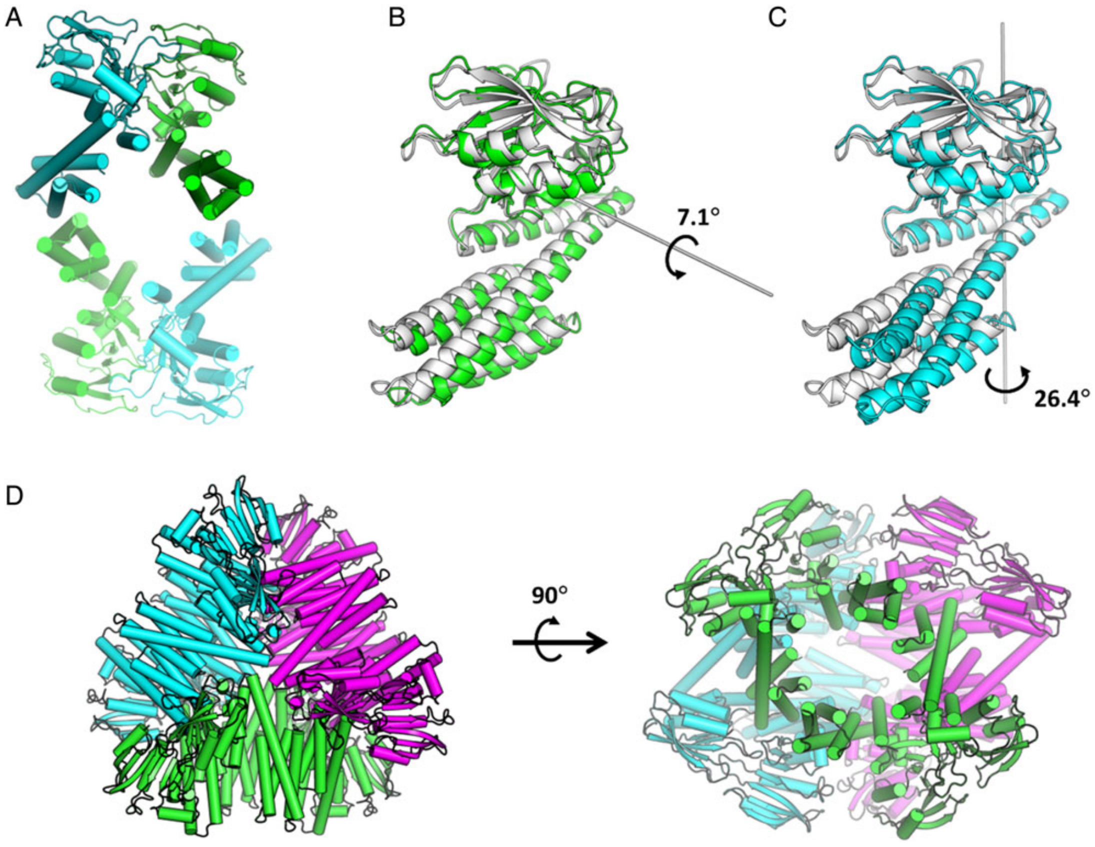

**Fig. 3** Crystal structures of 2ARH-3-3KAW and its deviation from the design. (**A**) The two protein chains in the asymmetric unit of the $P6_322$ crystal are shown in darker shades; the two independent chains are shaded differently. The symmetry-related copies forming the tetramer are shown in lighter shades. Comparisons of the two chains in the asymmetric unit with the idealized designed monomer are shown in (**B**) and (**C**). The designed model is shown as a white ribbon with the observed structures of the two independent chains in darker ribbons. The crystal structures are superimposed by the trimeric domain to highlight the relative rotation of the dimeric domain, which originates from bending in the helix linker region. The rotation axis and the rotation angle describing the helix bending in each case are shown. (**D**) A higher order 12-mer assembly created by three tetramers in the crystal, which the PISA program predicts could be stable in solution, viewed from the top (left) and the side (right).

reported trimeric interface, but formed a symmetric dimeric interface instead. In addition to this departure from the design, we found that the two independent monomers in the asymmetric unit displayed different amounts of deformation from the designed monomer structure (Fig. 3B and C and Table II). By comparing the two conformers in the crystal structure, the deformations could be attributed to the helix linker region. In comparing conformer A to the intended fusion model, the helix bending was only 7.1°, but in monomer B, the bending was more dramatic, reaching 26.4°. Despite the large local deformation, the overall structures of the fusion molecules were quite similar to the designed monomer structure. When superimposing the crystal structures on the ideal model by least-square fitting of the 294 $C_\alpha$ atoms, the RMSDs were only 1.6 and 3.4 Å for conformers A and B, respectively (Table II and Supplementary Fig. S2).

We further analyzed the oligomeric state in the crystal lattice computationally using the PISA program (Krissinel and Henrick, 2007). This analysis suggested that a higher order grouping, a 12-mer assembly, was present in the crystal (Supplementary Fig. S3). This 12-mer assembly is entirely dissimilar from the intended tetrahedral shape; three D2 tetramers are arranged in a plane to form a disk (Fig. 3D). This configuration in the crystal state suggested an alternative explanation for the size exclusion chromatogram. Our initial interpretation was that the peak ~470 kDa corresponded to the designed 12-mer tetrahedron. However, this peak could instead reflect the distinct 12-mer conformation seen in the crystal state. In the crystal arrangement (essentially a trimer of tetramers), the association energy between tetramers, as predicted by PISA, is much weaker than the association energy for subunits within a tetramer (Supplementary Fig. S3). This is consistent with the major peak at the tetramer molecular weight in the SEC chromatogram in addition to the 12-mer. We also analyzed the unexpected dimeric interface for the 2ARH domain observed in the crystal in comparison with its anticipated (native) trimeric interface. The area buried in the observed dimeric interface was smaller (1156 ± 9 $Å^2$ per monomer vs. 1848 ± 24 $Å^2$ per monomer in the native trimeric interface) and the estimated solvation free energy was weaker (−6.7 ± 2.1 kcal/mol per dimeric interface vs. −32.7 ± 3.5 kcal/mol per trimeric interface). It appears that the smaller assembly state obtained is built using weaker interfaces than those possible for a larger species built with native interfaces.

In view of the failure to form the intended 12-mer tetrahedron, we investigated a wide range of solution conditions in an attempt to identify conditions where a 12-mer was formed exclusively. Those experiments failed to identify any such conditions. We reasoned that the formation of smaller species might indicate a kinetically driven assembly outcome, because routes to large oligomers necessarily proceed through intermediates with fewer subunits. We therefore sought to favor (kinetically and thermodynamically) the formation of larger assemblies by increasing the protein concentration. This led invariably to aggregation—i.e. the formation of even larger assembly states than intended—possibly in the form of network like gels.

**Table II.** Summary of structural deviation from the ideal subunit

| | Rotation/shift | All $C_\alpha$ least-squares RMSD (Å) |
|---|---|---|
| 2ARH-3-3KAW | | |
| Conformer A | 7.1°/1.9 Å | 1.6 |
| Conformer B | 26.4°/0.6 Å | 3.4 |
| 2ARH-3-3KAW-2.0 | | |
| Conformer A | 26.1°/0.3 Å | 2.8 |
| Conformer B | 34.8°/2.6 Å | 2.9 |
| 2ARH-3-3KAW-3.0 | | |
| Conformer A | 15.1°/0.0 Å | 2.0 |
| Conformer B | 12.9°/0.4 Å | 1.9 |

The deviations reported describe differences between dimeric domains after superimposing trimeric domains.

## Crystal structure of 2ARH-3-3KAW-2.0

It was surprising that the reported trimeric domain did not form a trimer in the first crystal structure, but formed a dimer instead. We investigated what might cause this discrepancy. Because the 6xHis-tag at the N-terminus of 2ARH-3-3KAW came very close to the trimeric interface, we suspected that the 6xHis-tag might potentially be responsible for the unintended dimer formation. Because there was no room to accommodate the 6xHis-tag at the C-terminus of the 2ARH-3-3KAW fusion molecule, we used an alternative strategy for protein purification. It has been shown that histidine pairs in the (i, i + 3) or (i, i + 4) positions of an α-helix can bind to a nickel ion (Salgado *et al.*, 2009). We exploited this phenomenon and installed multiple bi-histidine motifs onto the first helix of the four-helix bundle of the dimeric domain. One of these variants could be obtained with high purity by IMAC chromatography without the presence of a terminal His-tag, and we dubbed this mutant 2ARH-3-3KAW-2.0 (Fig. 2).

We crystallized this mutant and solved its structure at 2.2 Å resolution in space group *R*3 (Table I). In this crystal structure, the trimeric domain indeed formed the intended trimeric interface as reported previously. However, the 2ARH-3-3KAW-2.0 fusion protein nonetheless formed a D3 hexameric assembly (Fig. 4A), and not the intended tetrahedral, 12-mer assembly. PISA analysis showed that no other higher order assemblies were present in the crystal lattice. This was consistent with a nearly complete disappearance of the 12-mer peak in the SEC chromatogram (Supplementary Fig. S4), which was instead characterized by a very broad peak centered around the hexamer molecular weight.

We inspected the crystal structure of 2ARH-3-3KAW-2.0 and found that, similar to the original 2ARH-3-3KAW design, deviations from the designed model originate in the helix linker (Fig. 4B and C; Table II). There were only minor changes within the trimeric and dimeric domains. The bending of the helix linkers in the two independent conformers is, however, much larger (26 and 35°) compared with those observed in the 2ARH-3-3KAW structure. Since the trimeric and dimeric domains maintained their native interfaces in the 2.0 variant, the flexibility of the helix linker seems to be the main culprit in allowing the formation of smaller assembly states. The native subunit interfaces are satisfied, but at the expense of bent helix geometry under a different symmetry than was designed.

## Crystal structure of 2ARH-3-3KAW-3.0

To test if the helix linker could be strengthened in a way that might benefit the assembly of the intended tetrahedral cage, we used the Rosetta suite of programs (Leaver-Fay *et al.*, 2011) to suggest amino acid changes to be made in the helix linker region. We recognized that a lone helix linker might not be able to provide the rigidity needed to maintain the desired orientation, so in addition to the optimization of the helix linker we also optimized a region on the four-helix bundle that comes into contact with the trimeric domain. In different designs, between two and six amino acid changes were considered. Several of the lowest-energy candidates were produced and one of them could be purified readily based on the pairs of histidines that were introduced; we named this Rosetta-optimized mutant 2ARH-3-3KAW-3.0 (see Fig. 2 for the introduced mutations).

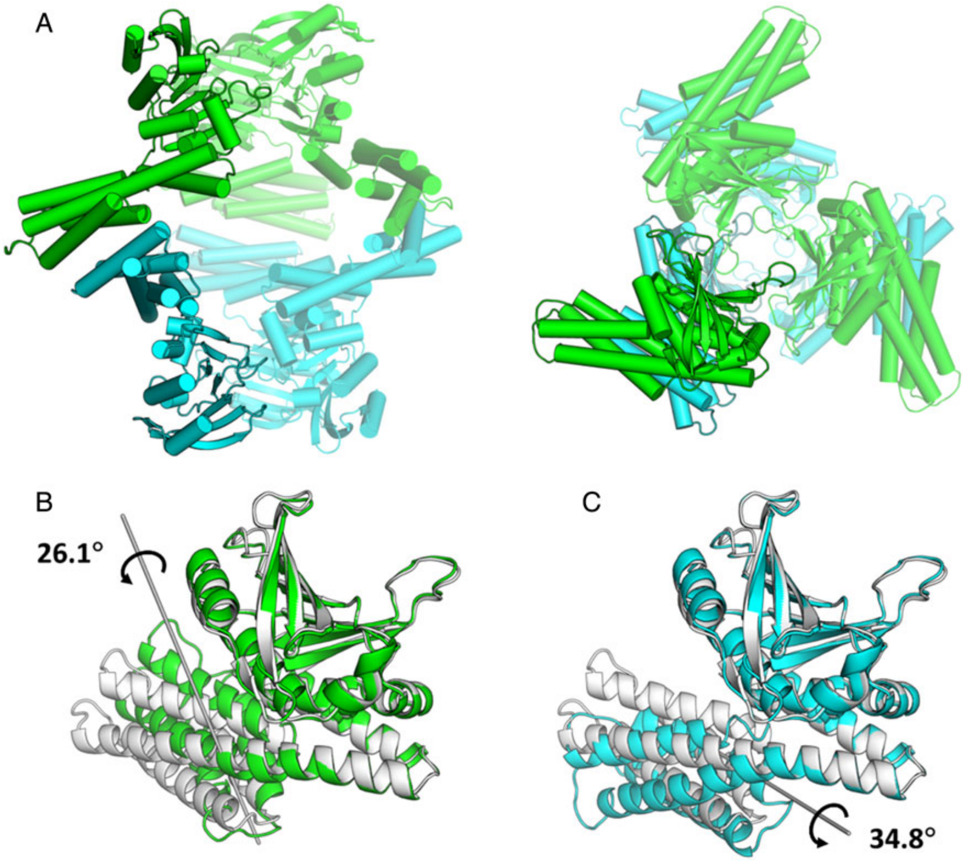

**Fig. 4** Crystal structure of 2ARH-3-3KAW-2.0 and its deviation from the design. (**A**) The content of the asymmetric unit of the crystal is shown in darker shades; the two independent chains are shaded differently. A view along one approximate (non-crystallographic) 2-fold axis is shown on the left and a view along the crystallographic 3-fold axis is shown on the right. The symmetry-related copies forming the hexamer are shown in lighter shades. Comparisons of the two chains in the asymmetric unit with the idealized model are shown in (**B**) and (**C**). The ideal model is shown as a white ribbon and the two independent chains in the crystal structure are shown as darker ribbons. The rotation axis and the rotation angles describing the helix bending are shown.

2ARH-3-3KAW-3.0 crystallized in the $P3_121$ space group and a structure was determined at 4.2 Å resolution (Table I). To our surprise, the reported trimeric domain again formed a dimeric interface, although no 6xHis-tag was present in this design. The fusion protein formed a compact tetramer in the crystal, similar to what was observed in the original 2ARH-3-3KAW design. Crystal packing analysis using the PISA program showed, however, that the disk-shaped higher order assembly observed earlier was not present in this crystal lattice. SEC for this variant indeed showed a dominant peak for the tetramer and only a very minor peak near the 12-mer molecular weight (Supplementary Fig. S4).

One concern during our design was about the high local negative charge of the helix linker (five glutamates within a stretch of six residues), but this did not prevent helix formation. However, helix flexibility again permitted the formation of an alternate assembly. The helix bending in the crystal was 15.1 and 12.9° for conformers A and B (Fig. 5B and C; Table II). The changes in bulky hydrophobic amino acids indicated by the Rosetta program were not sufficient to prevent bending of the helix linkers.

Motivated by the recurring tendency of the naturally trimeric interface of the 3KAW protein to form a spurious dimeric interaction, we attempted in limited experiments to block the dimeric interaction by mutation, without disrupting the desired trimeric interaction. Those experiments were unsuccessful, producing only monomeric protein.

## Discussion

We showed here that connecting two protein domains (which themselves have alpha-helical termini) using a short alpha-helical linker, produces a fusion protein in which the two domains are indeed connected by a spanning alpha helix. That feature offers an important ability to predict and control the relative orientation of two joined proteins. This result is in line with recent studies from our group, where the helix fusion strategy (Padilla *et al.,* 2001) has been used to control the relative orientation of two oligomeric proteins, thereby creating large self-assembling protein cages (Lai *et al.*, 2012a, 2013, 2014).

The primary objective of the present study was likewise to create a large symmetric cage by the fusion of two oligomers. It was an unexpected observation that, in several different experiments, our designed protein crystallized to reveal various smaller assemblies instead of the intended 12-subunit tetrahedral structure. Analysis of the crystal structures revealed that two factors contributed to the formation of unintended oligomers. First, the trimeric 2ARH domain proved capable of forming unanticipated dimeric associations in addition to the trimeric state seen in the deposited PDB structure. The dimeric interaction likely represents a minor form that is populated in the crystalline state under certain conditions, although a retrospective SEC analysis on the isolated 2ARH domain showed multiple overlapping peaks that provide evidence of polymorphic assembly behavior (see Supplementary Fig. S5). Secondly, the helix linker was able to bend

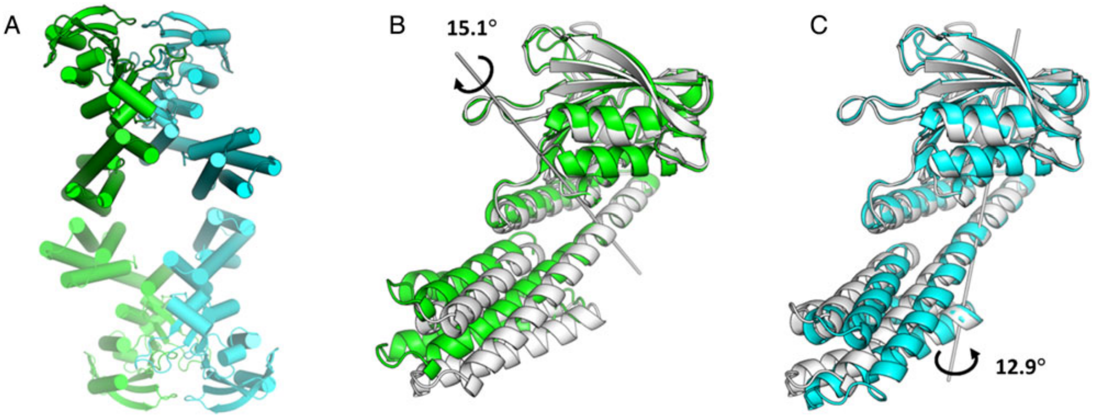

**Fig. 5** Crystal structure of 2ARH-3-3KAW-3.0 and its deviation from the design model. (**A**) The two protein chains comprising the asymmetric unit are shown in darker shades. The symmetry-related copies forming the tetramer are shown in lighter shades. Comparisons of the two chains in the asymmetric unit with the idealized model are shown in (**B**) and (**C**). The ideal model is shown in a white ribbon and the two independent chains in the crystal structure shown as darker ribbons. The rotation axis and the rotation angles are shown.

and twist, causing a range of angular perturbations (from 7 to 35°) that led to alternate assembly outcomes: tetramers and hexamers. In all cases, however, the fusion protein did in fact adopt a configuration in fairly close agreement with the computationally designed configuration (i.e. within ~3 Å deviation over ~300 $C_\alpha$ atoms). It is a particularly puzzling finding that, despite the ability of the fusion protein to very nearly adopt the correctly designed configuration (e.g. with a helix bent by as little as 7°), we did not in any case observe 12 copies of a subunit in that configuration assembling together to form the intended tetrahedral cage. That observation raises the possibility that kinetic phenomena may have an important role in dictating the outcome, with smaller assembly species finding ways to satisfy all the necessary protein subunit interfaces, even at the expense of adopting configurations of somewhat higher energy. Attempts to test this hypothesis and to shift the outcome to larger species by increasing the protein concentration were prevented by aggregation. Whether other approaches might be fruitful—e.g. by attempting complete unfolding and refolding—will require further study. The observed behavior in this system highlights key challenges in current efforts to achieve more highly reliable design of large, geometrically specific protein assemblies.

Currently the success rates for creating large, geometrically specific protein assemblies can be estimated at ~10% for both the helix fusion method and the interface design method. A common issue between the two strategies is that a large fraction of the designs are typically insoluble. It was observed in the interface design strategy that designs with closely related sequences can exhibit significantly different solubilities, emphasizing that even modest mutations at the surface of proteins can have detrimental consequences. For the helix fusion strategy, we suspect that one cause of insolubility may be the exchange of structural elements between the fused domains, which could lead to misfolding events. Another possible cause for the high incidence of insoluble designs is undesirable flexibility between the fused domains, which would allow the formation of extended networks when the designed proteins are over-expressed in bacterial hosts.

The use and properties of various linkers for fusing protein domains together have been widely studied (Argos, 1990; George and Heringa, 2002; Yu *et al.*, 2015). In the present study, we observed in multiple crystal structures that alpha-helical linkers can adopt a wide range of curved conformations, while the overall helical features (e.g. backbone hydrogen bonding) are maintained. The observed curvature patterns were typically smooth, with bending spread across the length of the helical segment rather that occurring as a sharp kink as occurs often in alpha helices containing proline (Deville *et al.*, 2008). Curved alpha helices similar to those we observed are not uncommon in natural proteins (Barlow and Thornton, 1988; Kumar and Bansal, 2012). The average radius of curvature for the helical fusion segments observed in our crystal structures is 41 ± 12 Å, compared with a typical range of 65 ± 34 Å for natural helices (Kumar and Bansal, 1998). The consequences of the allowed helix curvature in the present work are that the fused domains can rotate away from their idealized orientations, leading to either the formation of unintended interfaces or alternate assembly forms in which the intended interfaces are all satisfied. Helix bending, therefore, presents a challenge for applications that, like the present one, require precise geometric control, and further work to improve the rigidity of alpha-helical connections may be important. For other synthetic biology-related applications besides the construction of large, perfectly symmetric protein assemblies, the demand for orientational control between protein components will generally be more permissive. In those cases, the helix fusion strategy may offer geometric control within the desired range. Helix linkers have been used in synthetic fusion proteins of various functions, such as in a FRET biosensor (Sivaramakrishnan and Spudich, 2011), in a synthetic bifunctional enzyme (Arai *et al.*, 2001) and in a therapeutic fusion protein with improved oral efficacy (Bai and Shen, 2006). In one case involving a synthetic allosteric DNA-binding fusion protein, a helix linker designed to be continuous between the helical termini of the component proteins has already been explored (Strickland *et al.*, 2008).

## Supplementary data

Supplementary data are available at *PEDS* online.

## Acknowledgements

The authors thank Dan McNamara for help with cloning and protein expression, Michael Collazo for help with protein crystallization and Drs Duilio Cascio and Michael Sawaya for help with home X-ray data collection.

We thank the staff at the 24-ID-C beamline of the Advanced Photon Source for assistance with synchrotron X-ray data collection.

## Funding

This work was supported by NSF grant CHE-1332907 and by the BER program of the Department of Energy Office of Science.

## References

Adams,P.D., Afonine,P.V., Bunkoczi,G., *et al.* (2010) *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 213–221.

Arai,R., Ueda,H., Kitayama,A., Kamiya,N. and Nagamune,T. (2001) *Protein Eng.*, **14**, 529–532.

Argos,P. (1990) *J. Mol. Biol.*, **211**, 943–958.

Bai,Y. and Shen,W.C. (2006) *Pharm Res.*, **23**, 2116–2121.

Barlow,D.J. and Thornton,J.M. (1988) *J. Mol. Biol.*, **201**, 601–619.

Bricogne,G., Blanc,E., Brandl,M., *et al.* (2011) *BUSTER Version 2.10.0*. Cambridge, UK: Global Phasing Ltd.

Brodin,J.D., Ambroggio,X.I., Tang,C., Parent,K.N., Baker,T.S. and Tezcan,F.A. (2012) *Nat. Chem.*, **4**, 375–382.

Chen,A.H. and Silver,P.A. (2012) *Trends Cell Biol.*, **22**, 662–670.

Der,B.S., Machius,M., Miley,M.J., Mills,J.L., Szyperski,T. and Kuhlman,B. (2012) *J. Am. Chem. Soc.*, **134**, 375–385.

Der,B.S. and Kuhlman,B. (2013) *Curr. Opin. Struct. Biol.*, **23**, 639–646.

Delebecque,C.J., Lindner,A.B., Silver,P.A. and Aldaye,F.A. (2011) *Science*, **333**, 470–474.

Deville,J., Rey,J. and Chabbert,M. (2008) *Proteins*, **72**, 115–135.

Dueber,J.E., Wu,G.C., Malmirchegini,G.R., Moon,T.S., Petzold,C.J., Ullal, A.V., Prather,K.L. and Keasling,J.D. (2009) *Nat. Biotechnol.*, **27**, 753–759.

Fletcher,J.M., Harniman,R.L., Barnes,F.R., *et al.* (2013) *Science*, **340**, 595–599.

George,R.A. and Heringa,J. (2002) *Protein Eng.*, **15**, 871–879.

Good,M.C., Zalatan,J.G. and Lim,W.A. (2011) *Science*, **332**, 680–686.

Grueninger,D., Treiber,N., Ziegler,M.O., Koetter,J.W., Schulze,M.S. and Schulz,G.E. (2008) *Science*, **319**, 206–209.

Grunberg,R. and Serrano,L. (2010) *Nucleic Acids Res.*, **38**, 2663–2675.

Ha,J.H., Karchin,J.M., Walker-Kopp,N., Huang,L.S., Berry,E.A. and Loh,S.N. (2012) *J. Mol. Biol.*, **416**, 495–502.

Ha,J.H., Shinsky,S.A. and Loh,S.N. (2013) *Biochemistry*, **52**, 600–612.

Heinig,M. and Frishman,D. (2004) *Nucleic Acids Res.*, **32**, W500–W502.

Hoover,D.M. and Lubkowski,J. (2002) *Nucleic Acids Res.*, **30**, e43.

Kabsch,W. (2010) *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 125–132.

King,N.P. and Lai,Y.T. (2013) *Curr. Opin. Struct. Biol.*, **23**, 632–638.

King,N.P., Bale,J.B., Sheffler,W., McNamara,D.E., Gonen,S., Gonen,T., Yeates, T.O. and Baker,D. (2014) *Nature*, **510**, 103–108.

Krissinel,E. and Henrick,K. (2007) *J. Mol. Biol.*, **372**, 774–797.

Kumar,S. and Bansal,M. (1998) *Biophys. J.*, **75**, 1935–1944.

Kumar,P. and Bansal,M. (2012) *J. Biomol. Struct. Dyn.*, **30**, 773–783.

Lai,Y.T., Cascio,D. and Yeates,T.O. (2012a) *Science*, **336**, 11–29.

Lai,Y.T., King,N.P. and Yeates,T.O. (2012b) *Trends Cell Biol.*, **22**, 653–661.

Lai,Y.T., Tsai,K.L., Sawaya,M.R., Asturias,F.J. and Yeates,T.O. (2013) *J. Am. Chem. Soc.*, **135**, 7738–7743.

Lai,Y.T., Reading,E., Hura,G.L., Tsai,K.L., Laganowsky,A., Asturias,F.J., Tainer, J.A., Robinson,C.V. and Yeates,T.O. (2014) *Nat. Chem.*, **6**, 1065–1071.

Leaver-Fay,A., Tyka,M., Lewis,S.M., *et al.* (2011) *Methods Enzymol.*, **487**, 545–574.

Lee,H., DeLoache,W.C. and Dueber,J.E. (2012) *Metab. Eng.*, **14**, 242–251.

McCoy,A.J., Grosse-Kunstleve,R.W., Adams,P.D., Winn,M.D., Storoni,L.C. and Read,R.J. (2007) *J. Appl. Crystallogr.*, **40**, 658–674.

Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) *Acta Crystallogr. D Biol. Crystallogr.*, **53**, 240–255.

Ni,T.W. and Tezcan,F.A. (2010) *Angew. Chem. Int. Ed. Engl.*, **49**, 7014–7018.

Padilla,J.E., Colovos,C. and Yeates,T.O. (2001) *Proc. Natl Acad. Sci. U.S.A.*, **98**, 2217–2221.

Salgado,E.N., Lewis,R.A., Mossin,S., Rheingold,A.L. and Tezcan,F.A. (2009) *Inorg. Chem.*, **48**, 2726–2728.

Salgado,E.N., Ambroggio,X.I., Brodin,J.D., Lewis,R.A., Kuhlman,B. and Tezcan,F.A. (2010) *Proc. Natl Acad. Sci. U.S.A.*, **107**, 1827–1832.

Schulz,G.E. (2010) *J. Mol. Biol.*, **395**, 834–843.

Sinclair,J.C., Davies,K.M., Venien-Bryan,C. and Noble,M.E. (2011) *Nat. Nanotechnol.*, **6**, 558–562.

Sivaramakrishnan,S. and Spudich,J.A. (2011) *Proc. Natl Acad. Sci. U.S.A.*, **108**, 20467–20472.

Strickland,D., Moffat,K. and Sosnick,T.R. (2008) *Proc. Natl Acad. Sci. U.S.A.*, **105**, 10709–10714.

Worsdorfer,B., Woycechowsky,K.J. and Hilvert,D. (2011) *Science*, **331**, 589–592.

Yu,K., Liu,C., Kim,B.G. and Lee,D.Y. (2015) *Biotechnol. Adv.*, **33**, 155–164.