Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

The design and structural characterization of a synthetic pentatricopeptide repeat protein

Benjamin S. Gully,[a] Kunal R. Shah,[a,b] Mihwa Lee,[a] Kate Shearston,[a,b] Nicole M. Smith,[a] Agata Sadowska,[a] Amanda J. Blythe,[a] Kalia Bernath-Levin,[a,b] Will A. Stanley,[a,b]‡ Ian D. Small[b] and Charles S. Bond[a]*

[a]School of Chemistry and Biochemistry, The University of Western Australia, Crawley, Western Australia, Australia, and [b]Australian Research Council Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, Western Australia, Australia

‡ Current address, Northern Institute for Cancer Research, Newcastle University, Newcastle-upon-Tyne, England.

Correspondence e-mail: charles.bond@uwa.edu.au

Received 5 June 2014
Accepted 12 November 2014

PDB reference: *synth*PPR3.5, 4ozs

Proteins of the pentatricopeptide repeat (PPR) superfamily are characterized by tandem arrays of a degenerate 35-amino-acid α-hairpin motif. PPR proteins are typically single-stranded RNA-binding proteins with essential roles in organelle biogenesis, RNA editing and mRNA maturation. A modular, predictable code for sequence-specific binding of RNA by PPR proteins has recently been revealed, which opens the door to the *de novo* design of bespoke proteins with specific RNA targets, with widespread biotechnological potential. Here, the design and production of a synthetic PPR protein based on a consensus sequence and the determination of its crystal structure to 2.2 Å resolution are described. The crystal structure displays helical disorder, resulting in electron density representing an infinite superhelical PPR protein. A structural comparison with related tetratricopeptide repeat (TPR) proteins, and with native PPR proteins, reveals key roles for conserved residues in directing the structure and function of PPR proteins. The designed proteins have high solubility and thermal stability, and can form long tracts of PPR repeats. Thus, consensus-sequence synthetic PPR proteins could provide a suitable backbone for the design of bespoke RNA-binding proteins with the potential for high specificity.

1. Introduction

Controlling and manipulating macromolecular interactions has generated great interest over the past decade. Much of this biotechnological research initially studied short nucleic acids; however, the functional variance of proteins and subcellular targeting capabilities provide greater control and precision. While transcription-activator-like effector (TALE) proteins provided a route towards sequence-specific DNA binding (Boch *et al.*, 2009; Deng *et al.*, 2012; Yang *et al.*, 2000), with many downstream applications (Morbitzer *et al.*, 2010; Zhang *et al.*, 2011; Miller *et al.*, 2011; Hockemeyer *et al.*, 2011), an RNA-binding equivalent has proven elusive. PUF (Pumilio and FBF homology) proteins, characterized by imperfect ~36-amino-acid helical repeats (PUF domain) bind RNA in a modular mode (Wang *et al.*, 2002) which can be engineered to recognize specific RNA (Cheong & Hall, 2006; Filipovska *et al.*, 2011). Applications for such PUF proteins (Ozawa *et al.*, 2007; Tilsner *et al.*, 2009; Wang *et al.*, 2009) exist, although their RNA-binding capability (a maximum of ~16 nucleotides) is limited, as the structure is proposed to ultimately form a closed circle when additional PUF domains are added, thereby limiting specificity. As such, the search for a modular RNA-binding protein with greater specificity or targeting potential is a requirement for the biotechnological control of complex transcriptomes of higher organisms.

A potential solution to this need comes from the pentatricopeptide protein (PPR) superfamily. First identified a decade ago (Small & Peeters, 2000), PPR proteins are characterized by tandem degenerate 35-amino-acid repeat motifs that display some limited homology to tetratricopeptide repeat (TPR) motifs. Although PPR proteins are found throughout the eukarya (for example, there are seven human PPR-containing proteins), the potential for their use as a backbone for bespoke proteins is raised by their massive expansion in the plant kingdom (Barkan & Small, 2014). Over 450 divergent PPR proteins have been identified in *Arabidopsis thaliana* alone (Lurin *et al.*, 2004). PPR proteins are essential for organelle biogenesis in plants (Lurin *et al.*, 2004), RNA stabilization (Beick *et al.*, 2008; Choquet, 2009; Prikryl *et al.*, 2011), editing (Okuda *et al.*, 2006), post-transcriptional maturation (Williams-Carrier *et al.*, 2008; Delannoy *et al.*, 2007) and organellar translational control (Davies *et al.*, 2012; Zoschke *et al.*, 2012, 2013; Pfalz *et al.*, 2009; Barkan *et al.*, 1994), with genetic mutations resulting in cytoplasmic male sterility (Bentolila *et al.*, 2002; Akagi *et al.*, 2004), seed development (Gutiérrez-Marcos *et al.*, 2007) and other phenotypic impairments (Cushing *et al.*, 2005).

More than one subclass of PPR motifs exist: the originally identified, canonical 'P' motifs are found in tandem arrays of up to 30 'P-class' repeats. Longer and short ('L' and 'S') motifs have also been described, and are typically found in arrays of repeating PLS triplets (Lurin *et al.*, 2004). Despite this divergence, a set of very similar amino-acid codes has been

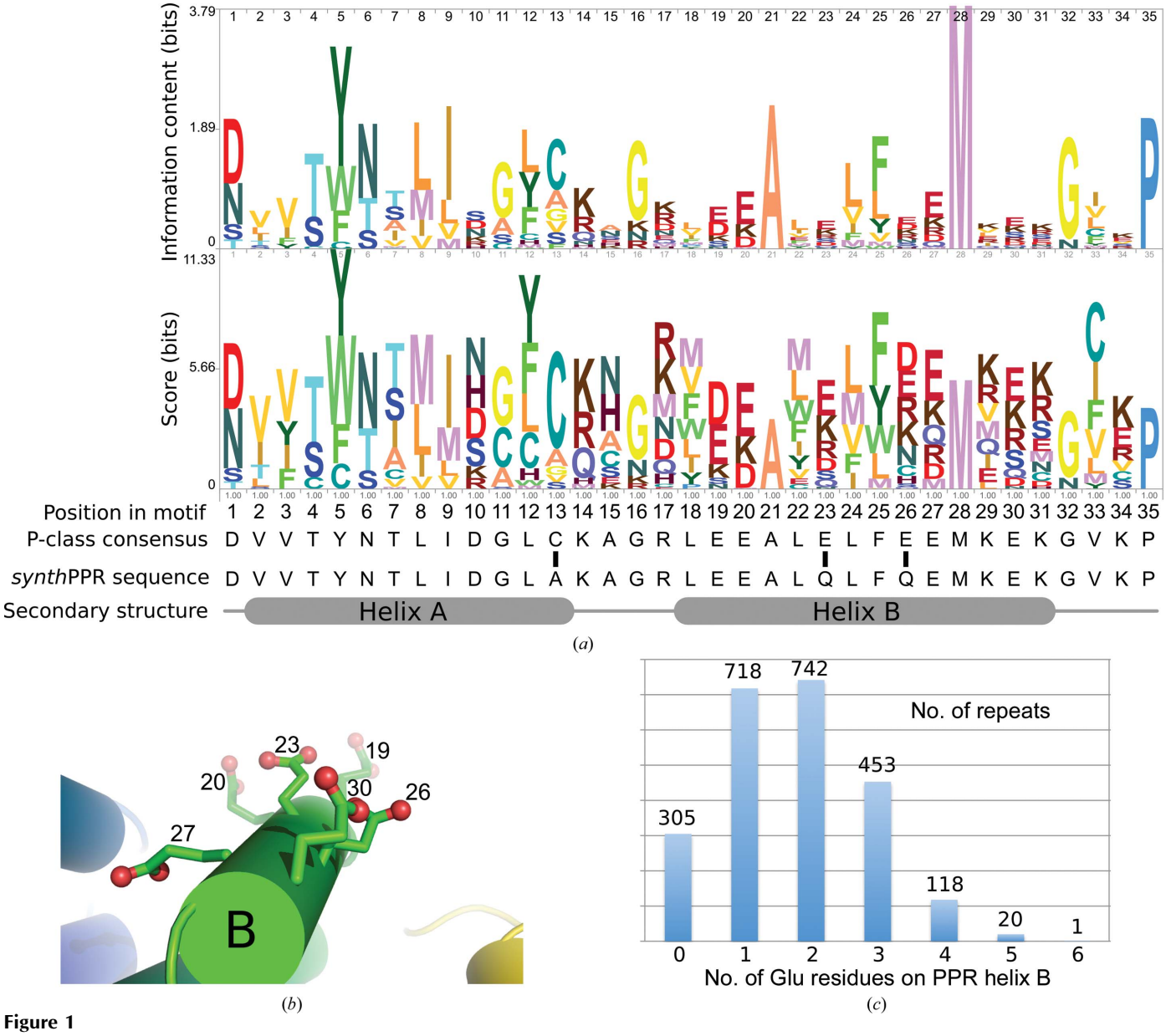


Figure 1
Sequence analysis of PPR domains. *SKYLIGN* representations of an alignment of 2357 PPR motifs displaying information content above background and score. The consensus sequence is provided, along with the modified consensus used for structural work. (*b*) A model of the predicted structure revealed a cluster of six glutamate residues on predicted helix B. (*c*) Only one real sequence from all 2357 has six glutamate residues on helix B.

elucidated that determines the sequence specificity of binding of single-stranded RNA by P-class and PLS-class PPR proteins (Barkan *et al.*, 2012; Takenaka *et al.*, 2013; Yagi *et al.*, 2013).

In this code, the nature of residue 6 of one motif and residue 1 of the subsequent motif (indicated by 1′) are most closely correlated with the identity of the base to be coordinated. In this way, tracts of tandem repeats can bind specific oligonucleotides.

PPR proteins are distantly related to TPR proteins, although the latter are generally associated with protein–protein interactions rather than protein–RNA interactions. The first example of a TPR structure was the three-TPR domain of protein phosphatase-5 (Das *et al.*, 1998), which showed the novel antiparallel α-helical architecture of the family. Subsequently, this led to the design and structural characterization of an idealized TPR motif (Main *et al.*, 2003; Kajander *et al.*, 2007). Biophysical measurements confirmed the α-helical content of PPR proteins (Beick *et al.*, 2008), and recently determined crystal structures of proteins that contain P-type PPR motifs: human mtRNAP (Ringel *et al.*, 2011), *A. thaliana* PRORP1 (Howard *et al.*, 2012) and most recently *Bracypodium distachyon* THA8 (Ke *et al.*, 2013) illustrate an antiparallel α-helical structure similar to that of TPR proteins. The first RNA-bound structure of a PPR protein has recently been described (Yin *et al.*, 2013): *Zea mays* PPR10 contains an array of PPR motifs bound to an 18 nt RNA sequence from the *psaJ* transcript.

Nevertheless, the structural context and limited sequence identity of the PPR repeats in these proteins in comparison with a canonical PPR repeat have shown that an engineered scaffold based on sound design principles is essential to support future PPR protein-engineering efforts to design bespoke RNA-binding proteins. Partly inspired by published work on designed TPR proteins, we defined a P-class PPR consensus sequence, modelled its structure, altered the sequence based on observations from the model, expressed the protein and determined its crystal structure.

2. Materials and methods

2.1. Source of sequence data

A set of 2357 previously identified *A. thaliana* PPR motifs consisting of exactly 35 amino acids (Small & Peeters, 2000) was used to obtain a consensus model using the *HMMER* v.2.26 package (Johnson *et al.*, 2010) as described in Lurin *et al.* (2004). Based on this analysis, a canonical PPR sequence of DVVTYNTLIDGLCKAGRLEEALELFEEMKEKGVKP was selected for this work. Fig. 1(*a*) presents the sequence alignment as analysed using *SKYLIGN* (Wheeler *et al.*, 2014) represented by information content and score. Minor discrepancies between the *HMMER* and *SKYLIGN* consensus sequences are observed at two low information content positions. A three-dimensional model of a four-motif consensus PPR array was constructed using the methodology described in Fujii *et al.* (2011).

2.2. Molecular biology, cloning and protein purification

The design principles for generating a synthetic protein are described in §3. The designed gene sequence, terminated by a double stop (TAATAA) codon and flanked by *Not*I and *Nco*I restriction-enzyme sites, was optimized and synthesized by GenScript and provided in a pUC57 vector. The *synth*PPR3.5 and *synth*PPR5.5 genes (containing three or five tandem PPR motifs, respectively) were subcloned into a pETM-11 plasmid (EMBL Protein Core Facility) using standard restriction (*Not*I and *Nco*I, New England Biolabs) and ligation (T4 DNA ligase, New England Biolabs) methods.

The resulting expression plasmids were used to transform *Escherichia coli* Rosetta 2 (DE3) cells (Novagen). 250 ml 2×YT medium supplemented with 1%(*w*/*v*) D-glucose, 50 m*M* Tris–HCl pH 7.5, 50 µg ml^{-1} kanamycin and 50 µg ml^{-1} chloramphenicol was inoculated with a transformed bacterial pre-culture and shaken at 37°C until the optical density at 600 nm reached 0.6. The culture was then cooled on ice for 5 min and expression was induced with 1 m*M* isopropyl β-D-1-thiogalactopyranoside with shaking overnight at 16°C. Selenomethionine-derivatized protein was generated *via* inhibition of the methionine-synthesis pathway in M9 medium supplemented with 1 m*M* $CaCl_2$, 2 m*M* $MgSO_4$, 0.5%(*w*/*v*) D-glucose and 0.0001%(*w*/*v*) thiamine. When an optical density at 600 nm of 0.6 was reached, the culture was supplemented with lysine, phenylalanine and threonine at 100 mg ml^{-1}, isoleucine, leucine and valine at 50 mg ml^{-1} and L-selenomethionine at 60 mg ml^{-1}. The culture was then induced with 1 m*M* isopropyl β-D-1-thiogalactopyranoside and shaken overnight at 16°C for protein expression.

The bacterial pellet was resuspended in buffer *A* [50 m*M* Tris–HCl pH 8.0, 500 m*M* KCl, 10%(*v*/*v*) glycerol, 1 m*M* DTT] supplemented with 0.13 m*M* PMSF (Roche), one Mini cOmplete protease-inhibitor tablet (Roche), 0.5 µl Benzonase (Sigma–Aldrich) and 1 mg ml^{-1} lysozyme and lysed under high pressure using a Emulsiflex C5 homogeniser (Avestin). The supernatant was loaded onto a pre-equilibrated 5 ml HisTrap column (GE Healthcare), washed with buffer *A* and eluted with a gradient of 0–500 m*M* imidazole in buffer *A*. The peak fraction was then subjected to gel-filtration chromatography (BioLogic DuoFlow, Bio-Rad) on a HiLoad 16/60 Superdex 200 prep-grade column (GE Healthcare) in buffer *A*. Peak fractions were confirmed *via* SDS–PAGE (NuPage Novex 4–12%, bis-tris gel, Invitrogen) stained with Coomassie Brilliant Blue (Supplementary Fig. S1[1]). Protein concentration was determined from the absorbance at 280 nm. The histidine tag was not cleaved for any subsequent experiments.

2.3. Circular dichroism

Circular dichroism was carried out with a sample of *synth*PPR3.5 at 0.2 mg ml^{-1} in 10 m*M* sodium borate buffer pH 8.0 using a Jasco J-810 spectropolarimeter equipped with a Peltier heating environment. Data were recorded with a 1 mm quartz cuvette using a 260–200 nm measurement range, 1 nm

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: KW5103).

bandwidth, 5 s response time, three accumulations and a 1 nm data pitch. All measurements were made in triplicate. Measurements were taken at 20°C following by heating to 95°C, remeasurement, cooling to 20°C and remeasurement. Baseline correction using buffer measurements at 20°C was employed.

2.4. Crystallization

Purified *synth*PPR3.5 in buffer *A* was concentrated to 15 mg ml^{-1} by centrifugation (Centricon) and screened for crystallization in 96-well sitting-drop format with 96-3 LVR Intelli-Plates (Art Robbins Scientific), a Phoenix liquid-handling robot (Art Robbins Scientific) and the sparse-matrix screens Crystal Screen, Crystal Screen 2, Index, PEG/Ion and Natrix (Hampton Research; 250 nl protein:250 nl crystallant; 50 µl reservoir). Crystal hits were optimized, resulting in diffraction-quality crystals that grew from 3 µl drops in 24-well sitting-drop CrysChem plates (Hampton Research) with a 2:1 ratio of crystallant [100 m*M* sodium citrate pH 3.35, 8%(*w*/*v*) PEG 3350] and protein equilibrated against 1 ml crystallant at 20°C. Crystals required cryoprotection with 10%(*w*/*v*) PEG 3350 and 15%(*v*/*v*) glycerol to maintain optimum diffraction. Crystals of selenomethionine-derivatized *synth*PPR3.5 and *synth*PPR5.5 were also grown in these optimized conditions.

2.5. X-ray data collection and structure solution

All diffraction data sets were collected using an ADSC Quantum 315r detector on the MX2 beamline of the Australian Synchrotron at a temperature of 100 K. Native data were collected at a wavelength of 0.9537 Å. For *synth*PPR3.5, selenomethionine-derivative (Se-*synth*PPR3.5) crystal diffraction data were collected at a peak wavelength of 0.9792 Å (f' −8.02, f'' 3.84) determined empirically from a Se fluorescence scan. All diffraction data were processed with *XDS* (Kabsch, 2010). The anomalous substructure determined by *phenix.autosol* (Adams *et al.*, 2010) included six Se atoms (x, y, z and occupancies: Se 1, −12.492, −59.940, −75.901, 2.05; Se 2, −3.836, −63.464, −84.503, 1.92; Se 3, −1.477, −33.605, −35.488, 1.58; Se 4, −5.349, −12.466, −24.587, 1.67; Se 5, −1.204, −43.042, −26.829, 1.31; Se 6, −7.979, −9.308, −67.705, 0.46). Phasing resulted in a figure of merit of 0.358 and SAD phases to 2.6 Å resolution. *phenix.autobuild* was used for density modification and model building, producing a model with five PPR motifs (and five Se atoms) in the asymmetric unit (R_{work} = 0.29; R_{free} = 0.32). The sixth potential Se position was spurious and did not correspond to any atomic position. At this point a switch to the isomorphous higher resolution native data was made. Iterative cycles of model building and refinement used *Coot* (Emsley *et al.*, 2010), *phenix.refine* and *BUSTER* (Bricogne *et al.*, 2011). *REFMAC* (Murshudov *et al.*, 2011) was used for the final stages of refinement, as a LINK statement was required to form a peptide bond between residue 175 of one asymmetric unit and residue 1 of the next. Coordinates were manipulated with *PDB-MODE* (Bond, 2003), sequence alignments were performed with *ALINE* (Bond & Schüttelkopf, 2009) and molecular graphics, interhelical angles and helix vectors were generated in *PyMOL* (Schrödinger). Superhelical parameters were derived by adapting *DynDom* (Hayward & Berendsen, 1998): individual PPR motifs were treated as domains, allowing the comparison of a protein consisting of motifs 1 and 2 with one consisting of motifs 1 and 3. The result is a translation axis and a rotation about the axis that describes the superhelix.

2.6. Modelling of protein–RNA complex

Coordinates of a nine-repeat *synth*PPR protein derived from the crystal structure and an oligouridine RNA octamer were prepared using *PDB-MODE* (Bond, 2003). Distance geometry restrained simulated annealing and Powell minimization was performed using *XPLOR-NIH*, with the backbone of the protein fixed and with atoms from the uracil base restrained to be within 5 ± 1 Å of the side-chain atoms of Asn6 and Asp1 of adjacent PPR motifs. All 20 parallel runs produced effectively identical structures, indicating a global minimum.

2.7. Protein mass determination from crystals

Approximately ten crystals (∼0.05 × 0.1 × 0.3 mm each) were harvested from a drop using a nylon loop (Hampton Research). They were washed by transfer into a protein-free drop of reservoir solution [100 m*M* sodium citrate pH 3.35, 8%(*w*/*v*) PEG 3350] and then dissolved in ∼4 µl water. For mass spectrometry, 10% trifluoroacetic acid (TFA) was added to a final concentration of 0.1%. The protein solution was mixed in a 1:1 ratio with matrix solution II [sinapinic acid (Sigma) saturated in 30% acetonitrile, 0.1% TFA]. Spotting was performed using the double-layer method. Initially, 1 µl of matrix solution I (sinapinic acid saturated in ethanol) was spotted. After it had dried, 1 or 2 µl of the protein–matrix mixture was spotted on top. After the spots had dried, they were analyzed with an UltraFlex III matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometer (Bruker Daltonics) at 50% laser intensity with up to 3000 shots. The instrument was calibrated using Protein Calibration Standard I (Bruker) with a mass range of 5700–16 952 Da. The spectra (*e.g.* Supplementary Fig. S2) from two protein dilutions each gave two main peaks corresponding to the single-charged (16 728.069, 16 762.987 Da) and double-charged (8361.770, 8370.968 Da) protein, respectively, which correlate well with the theoretical protein sequence (average mass 16 681 Da). The 47 Da shift between the observed and expected mass could be a combination of a 2–3 Da error and various salt adducts from the complex crystallization buffer or the oxidation of methionine.

3. Results and discussion

3.1. Sequence analysis

A P-class PPR consensus sequence was derived from a profile Hidden Markov Model (HMM) generated from 2357 PPR motifs found in *A. thaliana*. In the first instance, the

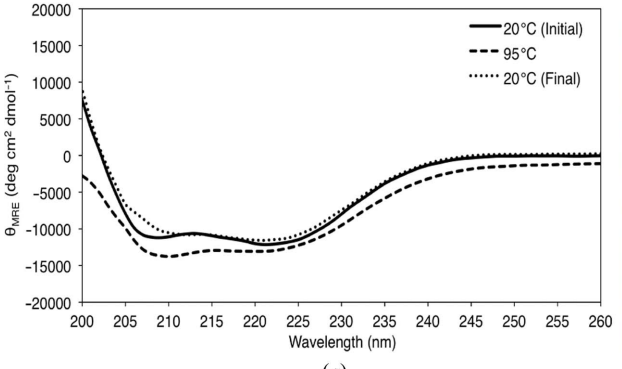
residue with the highest propensity at each of the 35 positions was used as the basis for the following work (Fig. 1*a*). In the absence of existing structural information on PPR proteins at the time of this work, we explored potential *ab initio* structural models of a conceptual PPR protein made from tandem consensus repeats using our previously described approach (Fujii *et al.*, 2010). In the resulting structure, the PPR motifs are composed of two antiparallel α-helices forming a hairpin. PPR motifs then stack together to form a superhelical solenoid structure. Helix A bears residues implicated in interaction with RNA, while helix B faces away from the RNA interface (Fujii *et al.*, 2011).

Table 1
Data-collection and refinement statistics.

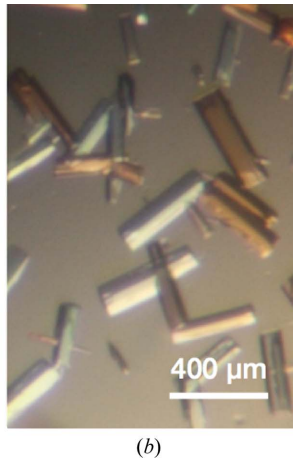
Values in parentheses are for the highest resolution shell.

	*synth*PPR3.5, PDB entry 4ozs	*synth*PPR3.5, long	Se-*synth*PPR3.5
Data collection			
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters			
a (Å)	54.0	54.2	54.1
b (Å)	75.0	75.2	75.0
c (Å)	85.1	255.9	85.7
Resolution (Å)	56.28–2.17 (2.25–2.17)	48.74–2.17 (2.25–2.17)	45.72–2.61 (2.73–2.61)
R_{merge}†	0.046 (0.709)	0.133 (2.616)	0.097 (1.271)
$CC_{1/2}$	0.999 (0.822)	0.999 (0.525)	—
$\langle I/\sigma(I)\rangle$	24.4 (2.9)	7.7 (0.8)	11.9 (1.5)
Completeness (%)	99.8 (98.9)	99.8 (99.8)	99.6 (97.3)
Average multiplicity	7.2 (7.3)	7.0 (7.1)	7.0 (6.7)
Anomalous completeness (%)	—	—	98.8 (94.1)
Anomalous multiplicity	—	—	3.7 (3.4)
Wilson *B* factor (Å^2)	48	49	—
Refinement			
Resolution (Å)	56.28–2.17		
Reflections (total/free)	17825/966		
R_{work}/R_{free}‡	0.216/0.265		
No. of atoms			
Total	1416		
Protein	1371		
Water	46		
Ramachandran outliers§ (%)	1.2		
Rotamer outliers§ (%)	5.5		
Clashscore§	13		
Average *B* factors (Å^2)			
Protein	52		
Ligand/ion	56		
Water	55		
R.m.s. deviations from ideal values¶			
Bond lengths (Å)	0.004		
Bond angles (°)	0.69		

† $R_{merge} = \sum_{hkl}\sum_j |I_j(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_j |I_j(hkl)|$. ‡ R_{work} and R_{free} are calculated from the working set of reflections and the test set, respectively, and expressed as $\sum_{hkl} ||F_{obs}| - |F_{calc}|| / \sum_{hkl} |F_{obs}|$. § Chen *et al.* (2010). ¶ Engh & Huber (1991).



(*a*)



(*b*)

Figure 2
The *synth*PPR3.5 protein is thermally stable and crystallizes readily. (*a*) Circular-dichroism spectropolarimetry of a sample at 20°C, at 95°C and returned to 20°C. Measurements were made in triplicate. Each spectrum was buffer baseline-corrected and smoothed using a 15-point Savitzky–Golay filter. Minimal structural differences are observed before and after heating. (*b*) Crystals of *synth*PPR3.5 were readily grown and have a clean orthorhombic habit.

3.2. Protein design

Armed with our consensus sequence, a predicted structure and a design approach previously applied for a consensus TPR protein (Kajander *et al.*, 2007), we designed a synthetic protein. Our first considerations were the suitability of the consensus sequence. We noted a conserved cysteine residue at position 13 and chose to substitute it with alanine, which is the next most favoured amino acid at this position, in order to reduce the potential for unwanted disulfide-bond formation. We also noted a cluster of negatively charged residues (Glu19, Glu20, Glu23, Glu26, Glu27 and Glu30) on the solvent-exposed face of helix B of our model (Fig. 1*b*). Despite the glutamic acid residues being most highly representative at each of those five positions, analysis showed that few of the 2357 individual native PPR sequences involved in building the consensus sequence have more than three negatively charged residues in this region and that only one has six (Fig. 1*c*). We therefore selected Glu23 and Glu26 to make conservative subsitutions to glutamine.

Our second consideration was the termini of the protein. At the N-terminus, a *Tobacco etch virus* protease-cleavable hexahistidine tag was included to assist purification, followed by a short α-helix-stabilizing sequence (AMGN; Dasgupta & Bell, 1993). At the C-terminus, an additional single helix was included that resembles the first half of a PPR motif, although four residues (Tyr5, Ile9, Leu12 and Ala13) were substituted by asparagine, lysine, alanine and serine, respectively,

to produce an amphipathic 'solubilizing' helix. In principle, any number of PPR motifs can be inserted between these termini: for these studies, we made *synth*PPR3.5 with three complete PPR motifs and *synth*PPR5.5 with five. The final protein sequence can be described as MKHHHHHHP-MSDYDIPTTENLYFQGAMGN-(DVVTYNTLIDGLAKA-GRLEEALQLFQEMKEKGVKP)$_n$-DVVTNNTLKDGASK-AG.

3.3. Protein expression and structure solution

The synthetic protein *synth*PPR3.5 was readily over-expressed and purified using standard methods. In contrast to wild-type PPR proteins in general, *synth*PPR3.5 is remarkably soluble and resilient to changes in buffer and temperature. In order to demonstrate the thermal stability of *synth*PPR3.5, we used circular-dichroism experiments (Fig. 2*a*). The protein can be heat-cycled from 20 to 95°C and back to 20°C with minimal change in the CD spectrum and no evidence of precipitation in the cuvette. This observation is in stark contrast to CD studies of the wild-type PPR5 protein, which unfolds irreversibly at 39°C, resulting in aggregation (Williams-Carrier *et al.*, 2008). The thermal stability of *synth*PPR3.5 bodes well for future biophysical studies of synthetic PPR proteins and potentially for their longevity in biological systems.

*Synth*PPR3.5 also crystallized readily using standard methods (Fig. 2*b*). Good-quality data could be collected to 2.17 Å resolution. Initial molecular-replacement attempts using a variety of models derived from predicted and other crystal structures (the structures of mtRNAP and AtPRORP1 containing PPR motifs had become available at this time) were unsuccessful. We chose to use single-wavelength anomalous dispersion methods on a selemethionine derivative of the protein to phase the structure. Details are presented in Table 1.

The structure-solution process revealed a number of surprises. The *synth*PPR3.5 protein contains 3.5 PPR motifs and includes five methionine residues. Two of these are part of the N-terminal tag which might be expected to be disordered. Matthews coefficient analysis of the crystals ($P2_12_12_1$, unit-cell parameters a = 54.0, b = 75.0, c = 85.1 Å) suggested either two (26% probability) or three (74%) molecules in the asymmetric unit. The observation of six Se atoms in the selenium substructure (with occupancies of 2.05, 1.92, 1.58, 1.67, 1.31 and 0.46) suggested one or two protomers in the asymmetric unit. The final structure (R_{work} = 0.22, R_{free} = 0.27) is in fact rather different, containing exactly five repeats per asymmetric unit in a superhelical arrangement (Fig. 3). Application of the 2_1 screw c axis to the coordinates generates a complete superhelical turn of ten PPR motifs, resulting in a continuous superhelix running throughout the crystal. There is no evidence of the N-terminal sequence and C-terminal helix in the electron density, nor are any breaks in electron density observed between subsequent repeats. The structure has the appearance of a seamless, infinitely long protein (Fig. 4).

3.4. Diagnosis of helical disorder

In order to investigate the discrepancy between the protein we had crystallized and the apparent structure, we undertook close examination of the data-processing results from *SCALA* and *TRUNCATE*, which yielded no anomalies. All measured parameters were perfectly normal: R_{merge} per image is in the range 3–6%, both R_{meas} and $\langle I/\sigma(I)\rangle$ vary smoothly across resolution ranges to reasonable values, the Wilson plot is exemplarily linear from 4.5 Å to the resolution limit, systematic absences are clearly observed for all three screw axes, no twinning or abnormal cumulative intensity distribution features are observed and the data show minimal anisotropy.

At this point, we closely inspected the diffraction images. For *synth*PPR3.5 we noted weak, streaky diffraction, with two evenly spaced minor peaks in between major peaks along the c axis (Fig. 5*a*). One potential explanation for this phenomenon is a superlattice in which the c axis is tripled. We considered this possibility because the convolution of our three-PPR-

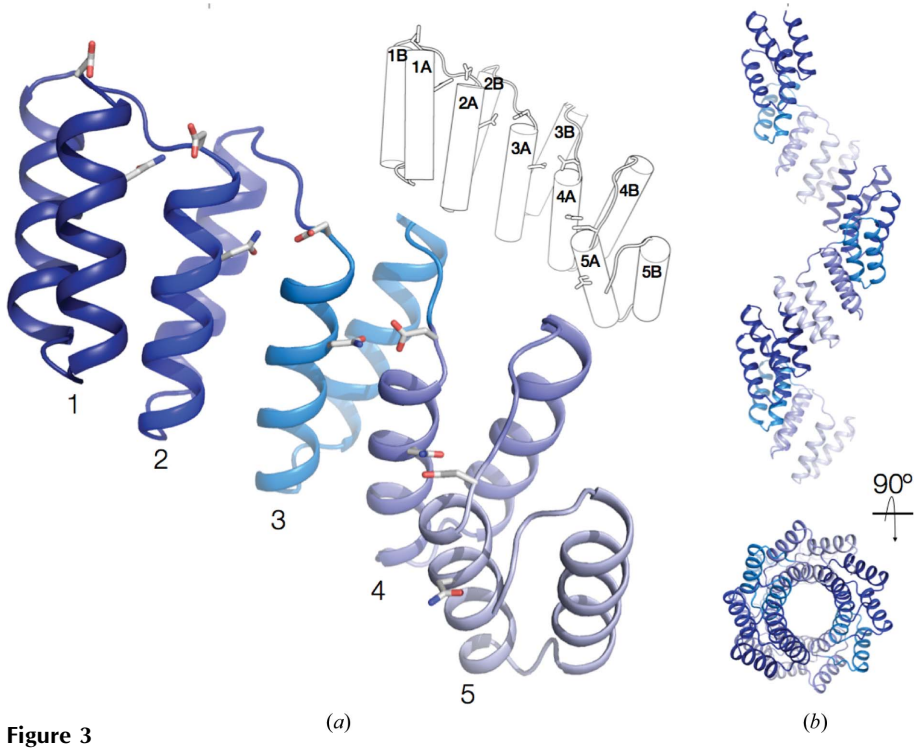


Figure 3
The crystal structure of *synth*PPR3.5 (*a*) The asymmetric unit includes five linked PPR motifs coloured in shades of blue. The $N_6D_{1'}$ RNA-binding residues are shown as sticks on the inner face of the superhelix. (*b*) The *synth*PPR superhelix represented by the content of three adjacent asymmetric units viewed perpendicular and parallel to the c axis.

motif molecule onto a unit cell containing ten protomers could result in longer range order corresponding to 30 PPR motifs. However, reprocessing of the data, still in $P2_12_12_1$, but with unit-cell parameters $a = 54.20$, $b = 75.24$, $c = 255.94$ Å (Table 1) revealed simply an overall reduction in signal and extreme systematic weakness for planes perpendicular to Miller index l. Analysis of the structure-factor amplitudes along the l axis reveals that the streaks are extremely weak, with each streak layer ($l = 3n + 1$, $l = 3n + 2$) having a mean $F/\sigma(F)$ of less than 6 at low values of l, in comparison to ∼40 for $l = 3n$ layers (Fig. 5*b*). It is clear from the images that the major diffraction spots contain no streaking and are exemplarily shaped, whereas the streaks are wide, often linking the neighbouring positions in h, k or $h + k$. Importantly, refinement of a 15-PPR-motif protein against the long unit-cell data still revealed an apparently infinite protein, with no breaks in electron density or evidence of the missing pieces of protein (Supplementary Fig. S3).

In seeking an explanation for this phenomenon, we noted that an analogous result had been obtained for the structure of the consensus TPR (cTPR) protein (Kajander *et al.*, 2007). In that case, however, an asymmetric unit cell that was too small for the crystallized proteins was observed: for example, a combination of a tetragonal fourfold screw axis ($P4_12_12_1$) and an asymmetric unit content of two cTPR motifs was observed for both an 8.5-motif and a 20.5-motif protein. The result is a hypersymmetric crystal structure of an apparently infinite TPR superhelix. In our case we observe a similar electron-density feature, yet as the asymmetric unit is large enough to contain a 3.5-PPR-motif protein, the term hypersymmetry does not seem appropriate. The best common explanation of both the *synth*PPR and cTPR structures is helical disorder. Analysis of the cTPR and *synth*PPR structures reveals infinite superhelices running along one or more crystallographic axes. Depending on the number of motifs present in the crystallized protein, one would expect breaks in electron density at regular intervals. In the case of three PPR motifs, this is incommensurate with five motifs per asymmetric unit (ten motifs per superhelical turn), and thus weak diffuse features appear in the diffraction pattern that correspond to some long-range order over three unit cells.

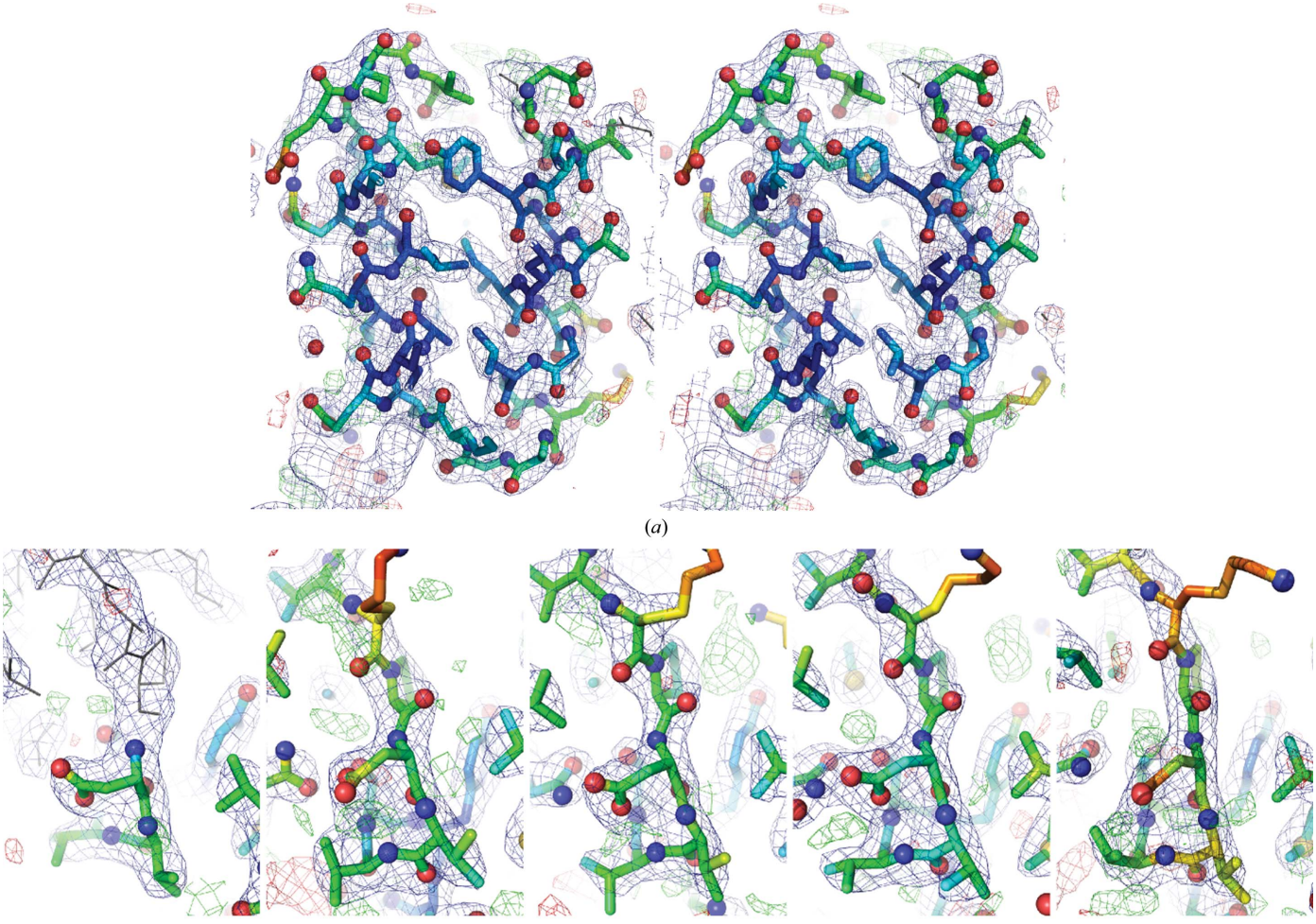


Figure 4
Electron density for the *synth*PPR3.5 structure. (*a*) A stereoview of motif 1 (residues 1–35) shown in ball-and-stick representation with sticks coloured according to *B* factor (blue, low; red, high), superimposed on $2m|F_o| - D|F_c|$, α_c (1.2σ, 0.18 e Å^{-3}, blue) and $m|F_o| - D|F_c|$, α_c (-3.0σ, -0.08 e Å^{-3}, red; $+3.0\sigma$, 0.08 e Å^{-3}, green) electron-density maps represented as mesh. (*b*) Snapshots of structure and electron density [as in (*a*)] at the five junctions between adjacent PPR motifs. Note the continuous density running between adjacent asymmetric units in the left panel.

However, there is the capacity for electron-density breaks along a superhelix in one part of the crystal to be out of register with those in another part. For *synth*PPR3.5, there are three possibilities (Fig. 6*a*). If the streaks are owing to the symmetry mismatch between the three PPR motifs of *synth*PPR3.5 and the five motifs in the asymmetric unit, then one would expect either no evidence of streaks or streaks coincident with diffraction spots for crystals of *synth*PPR5.5 (Fig. 6*b*). Indeed, *synth*PPR5.5 crystallizes under identical conditions, producing effectively indistinguishable diffraction (same space group and unit-cell parameters), except for the absence of interstitial streaks along *l*, but with evidence of very weak streaks between diffraction spots with the same value of *l* (Fig. 5*c*). Refinement against this data again reveals no evidence of chain breaks in the electron density (not shown). Owing to the rotational component of this disorder, it is not the same as the lattice translocation described by Wang *et al.* (2005). Additionally, the potential similarity to a hypersymmetric RNA duplex structure described by Shah & Brunger (1999) is not apparent in our data for similar reasons to those described by Kajander *et al.* (2007): the repeated units are perfectly identical in sequence, so no disruption of the intensity distributions is apparent and the helically disordered structure is well described by the application of crystal symmetry employed here.

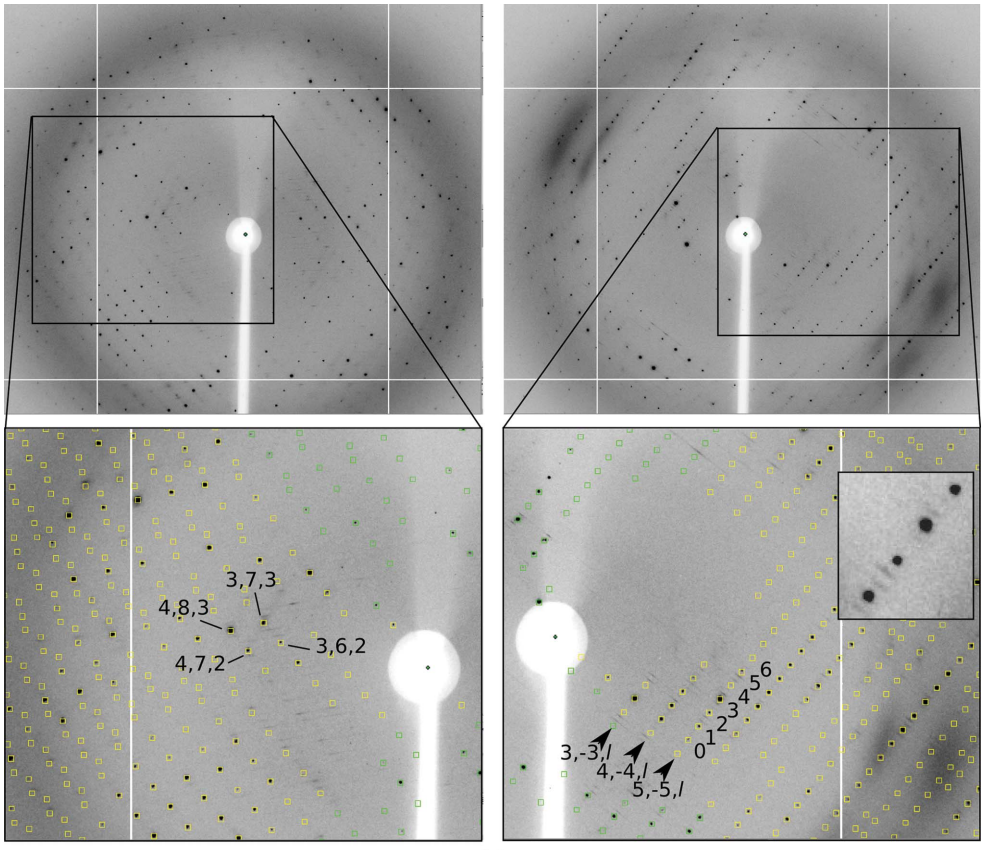


Figure 5
Helical disorder is evident in the diffraction from *synth*PPR crystals. (*a*) Two diffraction images (90° apart) from *synth*PPR3.5 crystals. Enlarged views give Miller indices for key spots. Streaking is observed between lattice points along *l*. (*b*) When the data are reprocessed with a tripled *c* axis, the mean signal-to-noise ratio for the *l* layers that correspond to the streaks is considerably lower than that for the major layers. (*c*) A diffraction image from *synth*PPR5.5 crystals reveals weak streaks perpendicular to the *l* diffraction maxima but not between them.

3.5. Verification of sample integrity

The observed phenomenon of a helically disordered superhelix raises questions about the physical status of the protein sample. The cTPR proteins described by Kajander and coworkers showed no evidence of aggregation in solution. Conversely, *synth*PPR3.5, when incubated over the course of days, displays an increased amount of sodium dodecyl sulfate/dithiothreitol-resistant laddering monitored by polyacrylamide gel electrophoresis (SDS–PAGE; Supplementary Fig. S1). Nevertheless, an explanation involving the formation of infinite assemblies of three-PPR-motif units raises the question as to what has happened to the additional parts of the protein: the N-terminal tag and the displaced C-terminal 'solvating helix'. SDS–PAGE analysis and MALDI-TOF mass spectrometry performed on a sample of approximately six-month-old crystals reveals clearly that the vast majority of protein is intact and full length, containing the histidine tag, TEV cleavage

site, helix-capping motif, three PPR motifs and a full C-terminal solvating helix (Supplementary Fig. S2; observed mass 16 728 Da; theoretical mass 16 681 Da; δ = +49 Da), although a minute amount of degraded protein in the range 14.9–15.7 kDa is observed.

4. Comparative structure analysis

Based on the excellent geometry and good-quality refinement of the structure, we consider the result of this work as an appropriate model for a consensus PPR protein of arbitrary length that is suitable for comparison with other PPR and TPR proteins. Notably, the structure is remarkably similar to the computational model on which we based our design. A four-PPR tract of predicted structure superimposes with the crystal structure with a low root-mean-square deviation (r.m.s.d.) of ∼2.5 Å for all 140 C^{α} atoms, indicating that the structure prediction had been highly successful (Supplementary Fig. S4).

4.1. Description of a canonical consensus PPR protein

Each PPR motif consists of two antiparallel four-turn α-helices (referred to as A and B, respectively) linked by a two-amino-acid intramotif turn and linked to the subsequent motif with a five-amino-acid loop (Fig. 3). Inspection of the temperature factors shows that the loops are more mobile than the helices (Supplementary Fig. S5). When the helical disorder of crystal context is considered one might expect that the occupancy of the atoms at the junction between motifs would be reduced by one third (the protomer making up the crystal consists of three PPRs). Although there is no obvious distortion of the electron density here, the temperature factors of these residues are amongst the highest within each motif (Supplementary Fig. S5).

4.2. Superhelical structure

Analysis of the superhelical parameters of the *synth*PPR structure provides an interesting comparison with cTPR. The *synth*PPR superhelix is continuous with a helical pitch of 85 Å, defining the *c* axis of the unit cell. The periodic repeat of the *synth*PPR superhelix is ten motifs, with each motif being replicated along the superhelix with a translational distance along the axis of 5.8 Å and a rotational angle of 38°. The TPR superhelix has a tighter 70 Å pitch, with a periodic repeat of eight motifs, corresponding to an axial translation of 6.4 Å and a rotational angle of 45° per motif. Helix A contributes side chains to the internal cavity in a similar arrangement to TPR proteins but with a higher conservation of hydrophilic residues and a more positively charged groove, similar to that observed in armadillo-repeat proteins. This groove corresponds to the RNA-binding surface of PPR proteins. The consensus sequence at positions 6 and 1′ of *synth*PPR corresponds to the combination expected for binding a uridine base: the $N_6D_{1'}$ RNA-binding variant has previously been shown to target U > C >>> A, G in RNA-binding studies of PPR10 (Barkan *et al.*,

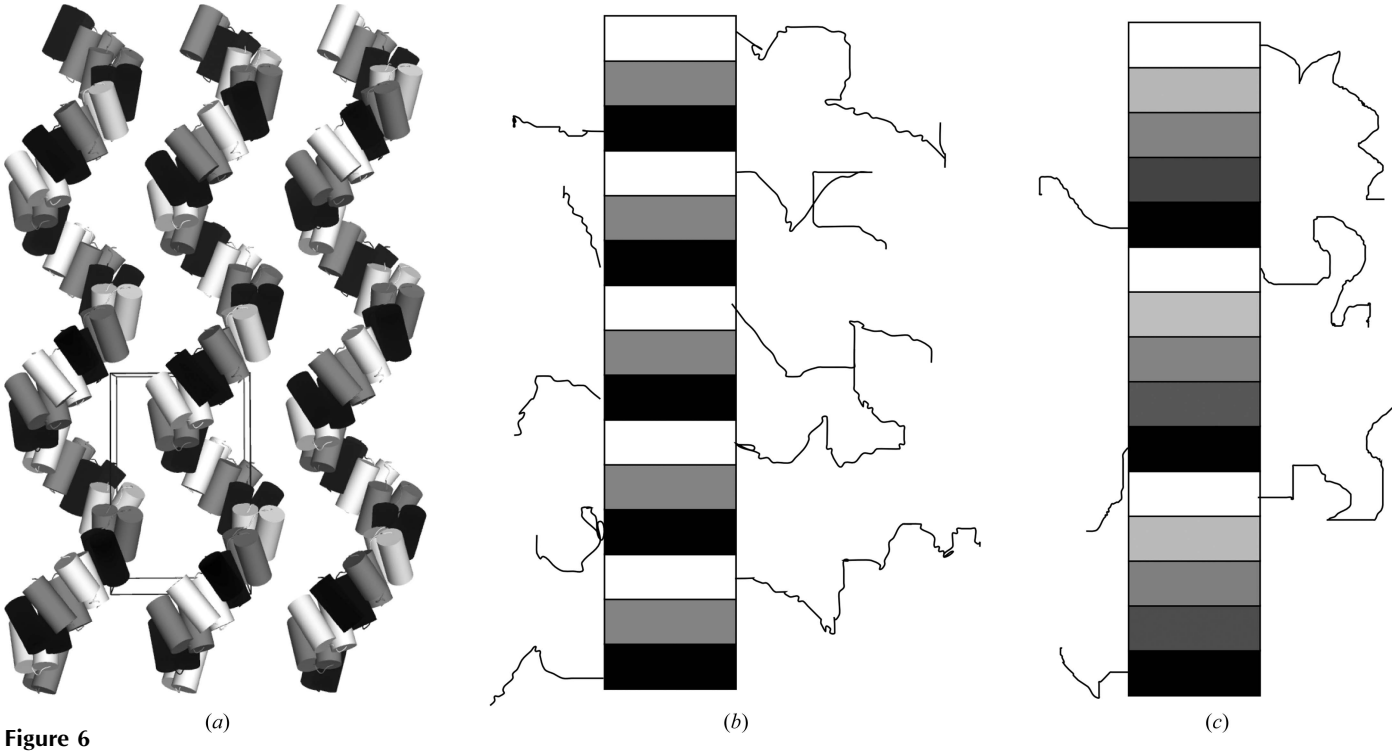


Figure 6
A schematic representation of helical disorder. (*a*) PPR motifs in three adjacent superhelices in the *synth*PPR3.5 structure are coloured as white–grey–black triplets. Arbitrary rotation/translation along the superhelical axis can place a white motif at any of these three positions. Thus, a superhelically averaged structure is observed. (*b*, *c*) Displacement of N-terminal and C-terminal parts of the polypeptide is possible for *synth*PPR3.5 and *synth*PPR5.5. Greater evidence of diffraction streaking for *synth*PPR3.5 may be owing to the higher molar concentration of disordered polypeptide.

2012). The *synth*PPR structure is consistent with this mode of binding RNA (Fig. 7).

In order to understand how the sequences of PPR and TPR proteins result in different superhelical structures, comparison can be broken down into intramotif interactions (between the residues on helices A and B of one motif) and intermotif interactions (between the helices of neighbouring motifs). As might be expected, the five PPR repeats in the *synth*PPR asymmetric unit are effectively structurally identical: they align with an r.m.s.d. of 0.22 ± 0.05 Å for all 35 C^{α} atoms (Supplementary Fig. S6). Mapping of the distribution of residue conservation, as described by the sequence logo (Fig. 1*a*), reveals that the most highly conserved residues are those with side-chain interactions between residues on helices A and B within a single motif, rather than those between motifs (Supplementary Fig. S6*g*).

4.3. Comparison between PPR and TPR motifs

A structural comparison between the *synth*PPR and cTPR (PDB entry 2fo7; Kajander *et al.*, 2007) structures reveals sequence differences that cause local intermotif differences in geometry (Fig. 8*a*), which in turn propagate to large long-range differences in superhelical structure. It is notable that all of the most highly conserved residues in both the PPR and TPR motifs are predominantly hydrophobic (Figs. 8*a* and 8*b*; with the exception of the PPR RNA-binding positions 1 and 6) and thus are buried in the various intramotif or intermotif interfaces. A comprehensive analysis of interhelical angles for a number of PPR and TPR proteins is provided in Supplementary Fig. S7. Notably, at first glance the intramotif interhelical angles of *synth*PPR and cTPR are very similar (∼167°). However, when this angle is decomposed into projections perpendicular to or parallel to the helical array (Figs. 8*b* and 8*c*), differences emerge. It is clear that helices A and B are more splayed out for PPR than TPR motifs. These differences can be accounted for by the observation that the two critical highly conserved residues buried at the interface between the helices in TPR repeats are tiny (Gly and Ala) and these are replaced by much bulkier conserved residues (Ile and Met) in PPR repeats. This observation may also explain the longer repeat length in PPRs (35 *versus* 34), allowing the helices to diverge further.

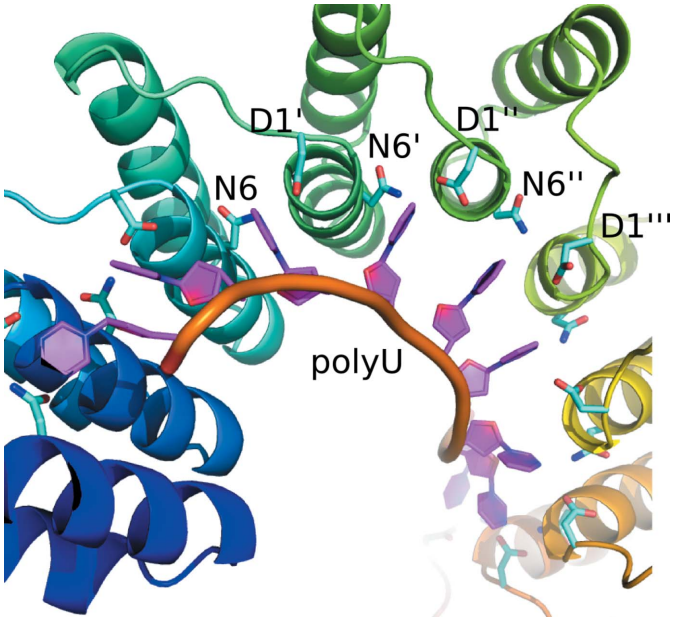


Figure 7
A model of polyU RNA (orange and magenta cartoon) bound to a *synth*PPR array (rainbow cartoon).

The most significant differences between PPR and TPR motifs, however, are apparent at the interface between adjacent motifs. The angle between helices A of adjacent motifs is ∼9° greater for TPR (22°) than PPR motifs (13°). This enhanced twist is responsible for the more highly overwound TPR superhelix (Fig. 8*c*). Analysis of the residues that make up the interface (Fig. 8*d*) shows that while residues from both helices A and B of both adjacent PPR motifs contribute to the interface, for TPR motifs helix B of the second motif barely participates, thus allowing the additional rotation. A further important structural distinction is the role of the conserved proline at the end of the motifs (Pro35 in PPR; Pro32 in TPR). Despite a similar position in the motif and a high level of conservation, the TPR proline exposes its side chain outwards, away from the interface between motifs, whereas the PPR proline is deeply buried. Its aliphatic atoms are buried in a pocket formed by Phe25, Met28, Lys29, Val33 and Thr4 and Tyr5 from the next subunit, thus playing a potentially key role in the superhelical structure. It is likely that the evolutionary constraint that causes this difference between PPRs and TPRs is the requirement that residue 6 on helix A of one motif and residue 1 of the next cooperate together to effect RNA sequence specificity.

4.4. Comparison between *synth*PPR and native PPR structures

Structural alignments of the previously crystallized PPR motifs from mtRNAP (Ringel *et al.*, 2011), PRORP1 (Howard *et al.*, 2012), THA8 (Ke *et al.*, 2013) and PPR10 (Yin *et al.*, 2013) illustrate the differences between the PPR motifs (Supplementary Fig. S6) and highlight the utility of the consensus structure described here. The intramotif angles in the previously crystallized PPR structures vary considerably relative to the well defined intramotif angle of *synth*PPR. In an attempt to define a typical P-type PPR motif structurally, we aligned the P-type motifs of mtRNAP, PRORP, THA8 and PPR10 with *synth*PPR, with resulting r.m.s.d.s of 3.1, 1.6 ± 0.3, 5.2 ± 1.2 and 1.7 ± 0.9 Å, respectively, *i.e.* a fivefold greater variance than observed within the *synth*PPR structure. The intermotif angles are similarly variable in the natural PPR structures, varying from 28 to 53° in the PPR10 structure.

The diversity of sequence represented in the existing PPR protein crystal structures makes rationalization of the specific determinants of superhelical structure difficult. The only clearly observable rule is that concerning the highly conserved methionine residue in helix B (Met28 in *synth*PPR). This residue projects towards helix A from helix B, making contacts

with residues Val2, Tyr5 and Ile9. In the apo PPR10 structure (PDB entry 4m57; Yin *et al.*, 2013; Supplementary Fig. S7) the intramotif angle is quite clearly different between those PPR motifs that conserve the methionine and those that do not. Repeats 3, 5–11, 16 and 17 have a methionine at this position and a typical interhelical angle of 168 ± 1°, whereas the other repeats have a wider angle of 158 ± 1°. Perhaps counter-intuitively, the residues at this position are smaller than methionine (valine and isoleucine), so the cause of the widened repeats is not simple steric repulsion (Fig. 9). Substitution of Met28 for less bulky side chains could alter its ability to pack against the conserved Pro35, with knock-on effects at the interface with the next PPR motif.

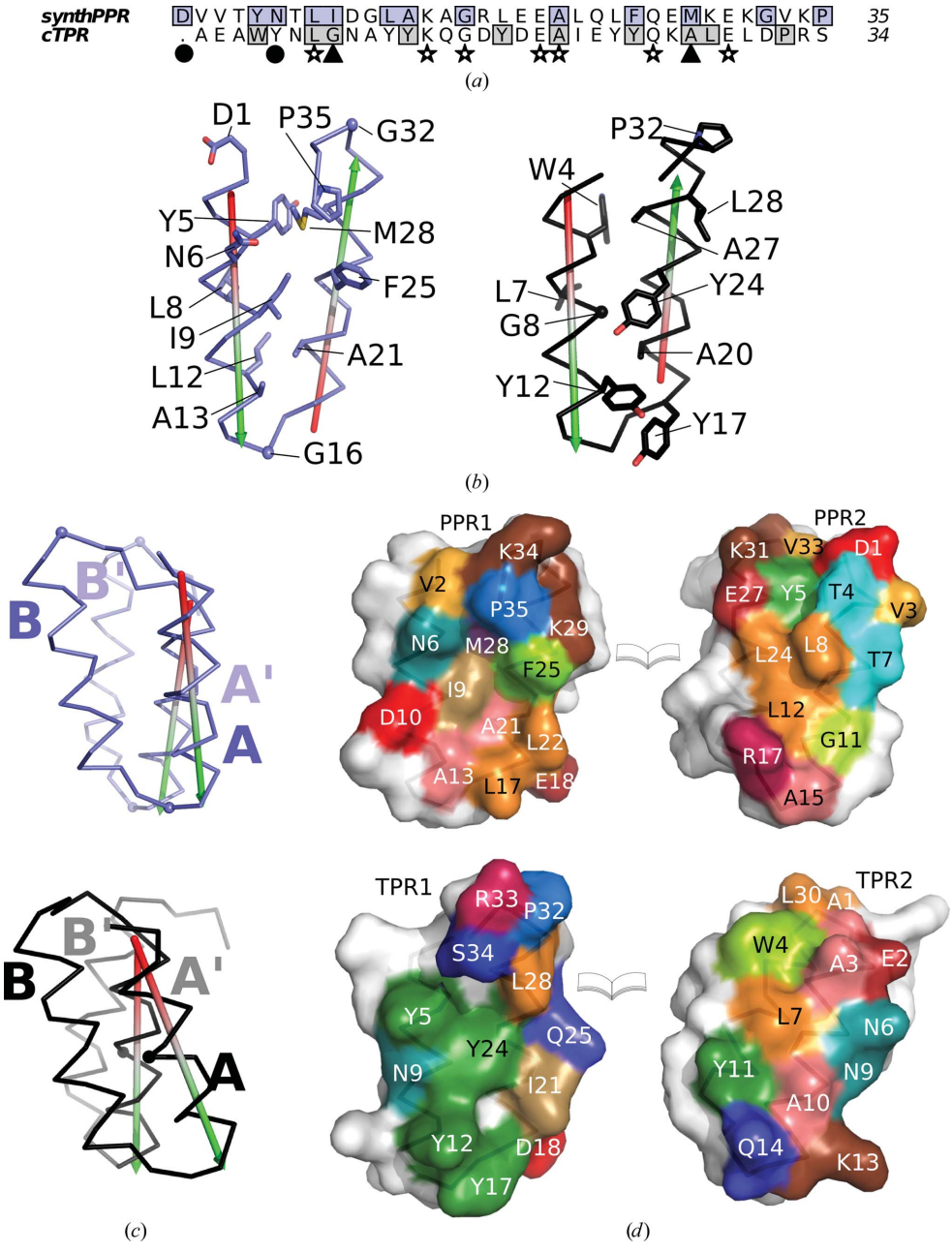


Figure 8
Comparison of the *synth*PPR and cTPR structures. (*a*) Sequence alignment of consensus PPR and TPR motifs. Most highly conserved positions are shaded (PPR, blue; TPR, grey). Circles mark RNA-binding positions in PPR motifs. Stars mark residues that are conserved between the PPR and TPR sequences. Triangles mark positions with significantly different-sized residue types. (*b*) Side views of individual PPR (blue) and TPR (black) motifs shown as ribbons, with sticks for conserved residues. Red-to-green vectors indicate the orientation of α-helices A and B. (*c*) Side views of pairs of adjacent PPR (blue) and TPR (black) motifs reveal the increased twist along a TPR superhelix compared with a PPR superhelix. Compare the angles between vectors representing helix A of adjacent repeats. (*d*) The identity of the amino-acid residues involved in the interface between adjacent motifs (PPR, top; TPR, bottom). A surface representation is used as if the motifs had been opened apart like pages of a book. Residues are coloured according to the sequence logo in Fig. 1.

4.5. The effect of 'designer' substitutions

We chose to substitute conserved glutamate residues with glutamines to minimize the size of a negatively charged cluster on the protein. The inclusion of glutamine residues in helix B, facing the outer side of the superhelix, may have played a role in improving the solubility by decreasing the charge of the outer side of the protomer, but are most likely to have influenced crystallization. Crystal contacts between superhelices involve the outer side of one protomer (protomer *A*) interacting with two protomers (α and β) in a second antiparallel superhelix (Fig. 10). These contacts are dominated by the engineered glutamines, with Gln58 of protomer *A* interacting with Glu159 of protomer α in the second superhelix, which is stabilized by Arg17 of protomer β in the antiparallel superhelix. $Gln61^{A}$ also interacts with the $Glu20^{\beta}$ stabilized by $Glu65^{A}$ and $Arg17^{\beta}$. It is thus possible that maintenance of Glu at positions 23 and 26 would have inhibited crystallization, although we have not tested this.

5. Conclusions

The results presented here provide novel insights into the design and structure of synthetic PPR proteins. The crystal structure presented here highlights which residues dictate the intramotif and intermotif relationships

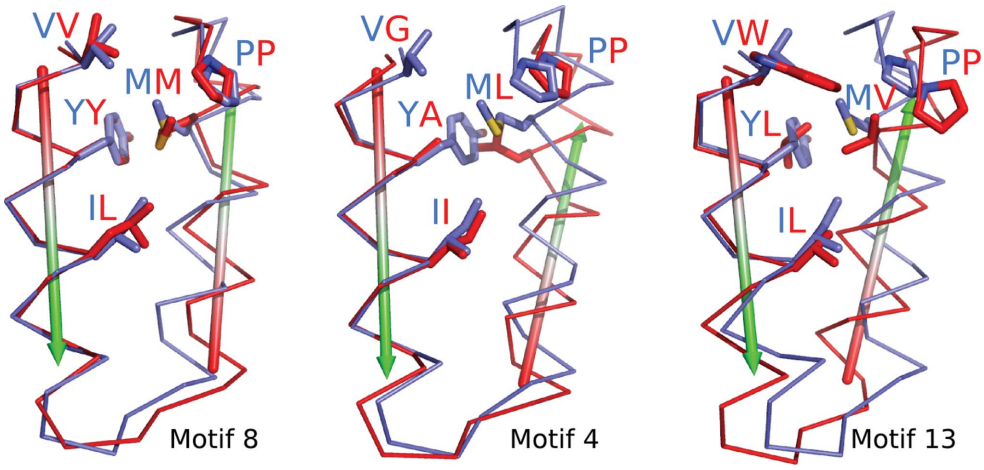


Figure 9
Comparison of *synth*PPR motifs with native PPR motifs. *synth*PPR (blue) is superimposed on motifs 8, 4 and 13 of *Z. mays* PPR10 (red; PDB entry 4m57). Motif 8 is a highly canonical motif and key residues are identical or similar to *synth*PPR, resulting in a similar interhelical angle. Motifs 4 and 13 are more divergent and are typical of the motifs that do not contain methionine at position 28. The result is a larger interhelical angle.

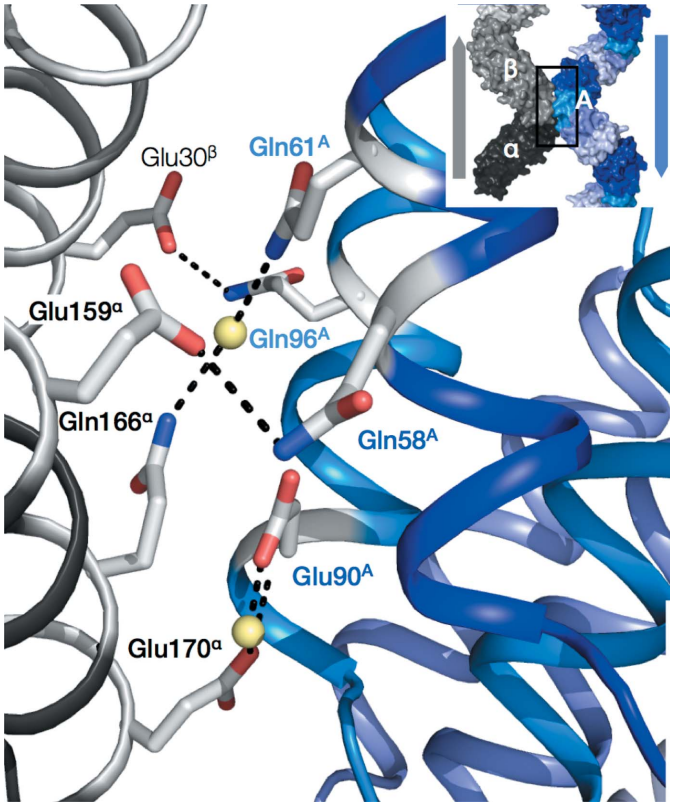


Figure 10
The crystal contacts between neighbouring superhelices (grey *versus* blue) in the *synth*PPR3.5 structure involve only a small number of hydrogen-bond contacts, either directly or *via* solvent atoms (yellow spheres). A number of glutamine residues which were substituted from glutamate in this work are involved.

of P-class PPR-motif arrays, in turn controlling the architecture of the RNA-binding surface. Additionally, the *synth*PPR structure defines the archetypal P-class motif scaffold. Importantly, the *synth*PPR proteins investigated here were highly stable in solution, unlike any of the increasing numbers of studied native PPR proteins, thus providing a basis for future studies into the detailed role of individual amino-acid positions in RNA binding and structure.

The assembly of protomers into an effectively infinite superhelix has both positive and negative implications for the utility of synthetic PPR proteins in biotechnological applications. On the one hand, long arrays replicate the unusually long RNA-binding surfaces of natural PPR proteins and support the notion that they are reproducible *in vitro*. On the other hand, however, the head-to-tail misassociation of *synth*PPR molecules could adversely modify sequence specificity, with unpredictable outcomes.

The *synth*PPR structure is an important step towards applications in biotechnology, providing a rational basis for engineering a suite of base-specific PPR motifs and RNA-specificity factors. Such customisable PPR proteins could have many uses as tools to intervene in post-transcriptional processes in living cells (Yagi *et al.*, 2014).

We are grateful for the financial support of this work by the Australian Research Council (Discovery Grant 120102870 to IDS and CSB), Pearl Technologies Ltd and The University of Western Australia (scholarship for BSG). X-ray data were collected on beamline MX2 of the Australian Synchrotron. KRS, WAS, IDS and CSB carried out bioinformatics, BSG, KS and AS carried out molecular biology and protein expression and crystallography, BSG, NMS, AJB and KBL carried out biophysical characterization, and BSG, ML and CSB analyzed the data. All authors contributed to the manuscript. The raw X-ray diffraction data used in this experiment have been deposited at https://store.synchrotron.org.au/ (Meyer *et al.*, 2014). These experimental data are openly available and licensed under Creative Commons Attribution 3.0 Australia (CC BY 3.0). The authors declare no competing financial interests.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Akagi, H., Nakamura, A., Yokozeki-Misono, Y., Inagaki, A., Takahashi, H., Mori, K. & Fujimura, T. (2004). *Theor. Appl. Genet.* **108**, 1449–1457.
Barkan, A., Rojas, M., Fujii, S., Yap, A., Chong, Y. S., Bond, C. S. & Small, I. (2012). *PLoS Genet.* **8**, e1002910.
Barkan, A. & Small, I. (2014). *Annu. Rev. Plant Biol.* **65**, 415–442.
Barkan, A., Walker, M., Nolasco, M. & Johnson, D. (1994). *EMBO J.* **13**, 3170–3181.

Beick, S., Schmitz-Linneweber, C., Williams-Carrier, R., Jensen, B. & Barkan, A. (2008). *Mol. Cell. Biol.* **28**, 5337–5347.

Bentolila, S., Alfonso, A. A. & Hanson, M. R. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 10887–10892.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A. & Bonas, U. (2009). *Science*, **326**, 1509–1512.

Bond, C. S. (2003). *J. Appl. Cryst.* **36**, 350–351.

Bond, C. S. & Schüttelkopf, A. W. (2009). *Acta Cryst.* D**65**, 510–512.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O. S., Vonrhein, C. & Womack, T. O. (2011). *BUSTER*. Cambridge: Global Phasing Ltd.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.

Cheong, C.-G. & Hall, T. M. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 13635–13639.

Choquet, Y. (2009). *EMBO J.* **28**, 1989–1990.

Cushing, D. A., Forsthoefel, N. R., Gestaut, D. R. & Vernon, D. M. (2005). *Planta*, **221**, 424–436.

Das, A. K., Cohen, P. W. & Barford, D. (1998). *EMBO J.* **17**, 1192–1199.

Dasgupta, S. & Bell, J. (1993). *Int. J. Pept. Protein Res.* **41**, 499–511.

Davies, S. M., Lopez Sanchez, M. I., Narsai, R., Shearwood, A. M., Razif, M. F., Small, I. D., Whelan, J., Rackham, O. & Filipovska, A. (2012). *FEBS Lett.* **586**, 3555–3561.

Delannoy, E., Stanley, W. A., Bond, C. S. & Small, I. D. (2007). *Biochem. Soc. Trans.* **35**, 1643–1647.

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y. & Yan, N. (2012). *Science*, **335**, 720–723.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Filipovska, A., Razif, M. F., Nygård, K. K. & Rackham, O. (2011). *Nature Chem. Biol.* **7**, 425–427.

Fujii, S., Bond, C. S. & Small, I. D. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 1723–1728.

Gutiérrez-Marcos, J. F., Dal Prà, M., Giulini, A., Costa, L. M., Gavazzi, G., Cordelier, S., Sellam, O., Tatout, C., Paul, W., Perez, P., Dickinson, H. G. & Consonni, G. (2007). *Plant Cell*, **19**, 196–210.

Hayward, S. & Berendsen, H. J. C. (1998). *Proteins*, **30**, 144–154.

Hockemeyer, D. *et al.* (2011). *Nature Biotechnol.* **29**, 731–734.

Howard, M. J., Lim, W. H., Fierke, C. A. & Koutmos, M. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 16149–16154.

Johnson, L. S., Eddy, S. R. & Portugaly, E. (2010). *BMC Bioinformatics*, **11**, 431.

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

Kajander, T., Cortajarena, A. L., Mochrie, S. & Regan, L. (2007). *Acta Cryst.* D**63**, 800–811.

Ke, J., Chen, R.-Z., Ban, T., Zhou, X. E., Gu, X., Tan, M. H. E., Chen, C., Kang, Y., Brunzelle, J. S., Zhu, J.-K., Melcher, K. & Xu, H. E. (2013). *Nature Struct. Mol. Biol.* **20**, 1377–1382.

Lurin, C. *et al.* (2004). *Plant Cell*, **16**, 2089–2103.

Main, E. R., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). *Structure*, **11**, 497–508.

Meyer, G. R., Aragao, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M. & Androulakis, S. (2014). *Acta Cryst.* D**70**, 2510–2519.

Miller, J. C. *et al.* (2011). *Nature Biotechnol.* **29**, 143–148.

Morbitzer, R., Römer, P., Boch, J. & Lahaye, T. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 21617–21622.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Okuda, K., Nakamura, T., Sugita, M., Shimizu, T. & Shikanai, T. (2006). *J. Biol. Chem.* **281**, 37661–37667.

Ozawa, T., Natori, Y., Sato, M. & Umezawa, Y. (2007). *Nature Methods*, **4**, 413–419.

Pfalz, J., Bayraktar, O. A., Prikryl, J. & Barkan, A. (2009). *EMBO J.* **28**, 2042–2052.

Prikryl, J., Rojas, M., Schuster, G. & Barkan, A. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 415–420.

Ringel, R., Sologub, M., Morozov, Y. I., Litonin, D., Cramer, P. & Temiakov, D. (2011). *Nature (London)*, **478**, 269–273.

Shah, S. A. & Brunger, A. T. (1999). *J. Mol. Biol.* **285**, 1577–1588.

Small, I. D. & Peeters, N. (2000). *Trends Biochem. Sci.* **25**, 46–47.

Takenaka, M., Zehrmann, A., Brennicke, A. & Graichen, K. (2013). *PLoS One*, **8**, e65343.

Tilsner, J., Linnik, O., Christensen, N. M., Bell, K., Roberts, I. M., Lacomme, C. & Oparka, K. J. (2009). *Plant J.* **57**, 758–770.

Wang, J., Kamtekar, S., Berman, A. J. & Steitz, T. A. (2005). *Acta Cryst.* D**61**, 67–74.

Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. (2002). *Cell*, **110**, 501–512.

Wang, Y., Cheong, C.-G., Tanaka Hall, T. M. & Wang, Z. (2009). *Nature Methods*, **6**, 825–830.

Wheeler, T. J., Clements, J. & Finn, R. D. (2014). *BMC Bioinformatics*, **13**, 15–17.

Williams-Carrier, R., Kroeger, T. & Barkan, A. (2008). *RNA*, **14**, 1930–1941.

Yagi, Y., Hayashi, S., Kobayashi, K., Hirayama, T. & Nakamura, T. (2013). *PLoS One*, **8**, e57286.

Yagi, Y., Nakamura, T. & Small, I. (2014). *Plant J.* **78**, 772–782.

Yang, B., Zhu, W., Johnson, L. B. & White, F. F. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 9807–9812.

Yin, P., Li, Q., Yan, C., Liu, Y., Liu, J., Yu, F., Wang, Z., Long, J., He, J., Wang, H.-W., Wang, J., Zhu, J.-K., Shi, Y. & Yan, N. (2013). *Nature (London)*, **504**, 168–171.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G. M. & Arlotta, P. (2011). *Nature Biotechnol.* **29**, 149–153.

Zoschke, R., Kroeger, T., Belcher, S., Schöttler, M. A., Barkan, A. & Schmitz-Linneweber, C. (2012). *Plant J.* **72**, 547–558.

Zoschke, R., Qu, Y., Zubo, Y. O., Börner, T. & Schmitz-Linneweber, C. (2013). *J. Plant Res.* **126**, 403–414.