nature structural & molecular biology

Structural basis for RNA recognition by a dimeric PPR-protein complex

Jiyuan Ke[1,7], Run-Ze Chen[2,7], Ting Ban[2,3,7], X Edward Zhou[1], Xin Gu[1], M H Eileen Tan[1,4], Chen Chen[1,4], Yanyong Kang[1], Joseph S Brunzelle[5], Jian-Kang Zhu[3,6], Karsten Melcher[1] & H Eric Xu[1,2]

Thylakoid assembly 8 (THA8) is a pentatricopeptide repeat (PPR) RNA-binding protein required for the splicing of the transcript of *ycf3,* a gene involved in chloroplast thylakoid-membrane biogenesis. Here we report the identification of multiple THA8-binding sites in the *ycf3* intron and present crystal structures of *Brachypodium distachyon* THA8 either free of RNA or bound to two of the identified RNA sites. The apostructure reveals a THA8 monomer with five tandem PPR repeats arranged in a planar fold. The complexes of THA8 bound to the two short RNA fragments surprisingly reveal asymmetric THA8 dimers with the bound RNAs at the dimeric interface. RNA binding induces THA8 dimerization, with a conserved G nucleotide of the bound RNAs making extensive contacts with both monomers. Together, these results establish a new model of RNA recognition by RNA-induced formation of an asymmetric dimer of a PPR protein.

Pentatricopeptide repeat (PPR) proteins form a large family of RNA-binding proteins with more than 450 members in *Arabidopsis thaliana*[1–3]. They are conserved in yeast, plants and humans, and they have crucial roles in RNA metabolism. In plants, PPR proteins are involved in RNA editing, mRNA stabilization and splicing of many genes in chloroplasts and mitochondria[4,5]. PPR proteins, which are characterized by degenerate 35-aa motifs, are arranged in tandem repeats with an average of ~12 PPR repeats in *A. thaliana*[4]. The crystal structures of the PPR domain of human mitochondrial RNA polymerase and that of PRORP1 from *A. thaliana* reveal that each PPR motif adopts two antiparallel α-helices[6–8].

An intriguing question is how PPR proteins specifically recognize their target RNA sequences. Understanding this fundamental question will not only improve understanding of their biological functions but also help protein engineering of PPR proteins in biotechnology applications of targeting specific RNA sequences for transcriptional and translational regulation. Several PPR proteins and their endogenous specific RNA-recognition sequences have been characterized biochemically[9–12]. The recently reported structure of human IFIT5, a closely related tetratricopeptide repeat protein, in complex with viral 5′-triphosphate RNAs (PPP-RNAs) revealed a sequence-nonspecific interaction between the helical domain and the viral single-stranded PPP-RNA[13]. However, the structural mechanism of specific RNA recognition by PPR proteins remains unknown. Studies have demonstrated base-specific recognition of RNA by PUF domains, with amino acids at positions 12 and 16 of each PUF repeat interacting with RNA bases through hydrogen bonds, and amino acids at position 13 making base-stacking interactions[14,15]. The modular recognition of RNA by PUF proteins suggests that PPR proteins may also bind RNA in a similar fashion, with roughly one PPR repeat recognizing one RNA base[16]. Bioinformatics analysis of PPR protein sequences indicates that residues at positions 6 and 1′ of each PPR motif determine the RNA-base specificity[17].

Thylakoid assembly 8 (*tha8*) is a maize gene, the mutations of which cause defects in the biogenesis of chloroplast thylakoid membranes[3]. The *tha8* gene encodes a small PPR protein that is localized to chloroplasts, where it is required for the splicing of the *ycf3-2* and *trnA* group II introns[3]. *A. thaliana* has a THA8 ortholog with conserved functions in the biogenesis of chloroplast thylakoid membranes[3] and a THA8-like (THA8L) protein, which shares 26% sequence identity with THA8 but has unknown functions[18]. Whereas most PPR proteins have more than ten PPR motifs, THA8 belongs to a subfamily of small PPR proteins that are involved in the splicing of group II introns through specific RNA binding. In this paper, we identify the THA8 RNA-binding sites and report the crystal structures of THA8 both free of RNA and bound to two of the identified RNA sites. Our structures validate a previously proposed combinatorial amino acid code of RNA recognition[17,19,20] but also reveal an unexpected mechanism of RNA recognition by RNA-induced formation of an asymmetric dimer of a PPR protein.

[1]Laboratory of Structural Sciences, Center for Structural Biology and Drug Discovery, Van Andel Research Institute, Grand Rapids, Michigan, USA. [2]Van Andel Research Institute–Shanghai Institute of Materia Medica (VARI-SIMM) Center, Center for Structure and Function of Drug Targets, Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. [3]Shanghai Center for Plant Stress Biology, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [4]Department of Obstetrics & Gynecology, National University Hospital, Yong Loo Lin School of Medicine, National University of Singapore, Singapore. [5]Life Sciences Collaborative Access Team, Synchrotron Research Center, Northwestern University, Argonne, Illinois, USA. [6]Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, Indiana, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to H.E.X. (eric.xu@vai.org).

Received 12 August; accepted 8 October; published online 3 November 2013; doi:10.1038/nsmb.2710

RESULTS

THA8 structure and RNA binding

To unravel the structural basis of RNA recognition by PPR proteins, we focused our studies on THA8, a small PPR protein with a defined phenotype and molecular function in RNA splicing[3]. The THA8 proteins from maize, rice and *Brachypodium distachyon* are highly conserved, with >73% sequence identity among them. We expressed and purified the THA8 proteins from all three species. Only the THA8 protein from *B. distachyon* produced high-quality crystals in both RNA-free and RNA-bound states. We thus focused our studies on this protein and determined the crystal structure of apo-THA8 at 1.6-Å resolution (**Table 1**). It revealed five tandem PPR repeats in a similar arrangement as seen in the recently determined THA8L structure[18] (**Supplementary Fig. 1a**,**b**). Each PPR repeat is formed by two antiparallel α-helices connected by a short turn linker. The linker between repeats 3 and 4 is very long in THA8 as compared to that of THA8L (**Supplementary Fig. 1b**,**c**).

THA8 also specifically binds to the same short RNA targets as identified for THA8L[18] (**Supplementary Fig. 2**), thus suggesting a conserved function between THA8 and THA8L. Using an AlphaScreen binding assay (**Supplementary Fig. 3**), we found that THA8 binds specifically to the short purine-rich sequence present in the Zm1a RNA (**Supplementary Fig. 3b**,**c**). Competition with short RNA sequences from different species (**Supplementary Fig. 3d**) revealed that THA8 binds to Zm4 RNA with the highest affinity (half-maximal inhibitory concentration (IC$_{50}$) = 4 nM) (**Supplementary Fig. 3e**,**f**). Removal of up to six additional nucleotides from the 3′ end of Zm4 has little effect on their binding to THA8 (**Supplementary Fig. 3g**), thus suggesting that the G-A–rich sequence itself is sufficient for high-affinity binding to THA8. Similarly to THA8L, THA8 also preferentially binds to single-stranded RNA (**Supplementary Fig. 4a**).

Crystal structures of THA8–RNA complexes

We determined the crystal structures of THA8 in complex with Zm4 13-mer RNA and Zm1a-6 12-mer RNA at a resolution of 2.8 Å and 3.0 Å, respectively (**Table 1**). Both complex structures revealed that THA8 is assembled into an asymmetric dimer with the RNA bound at the dimeric interface formed by the C-terminal part of one monomer and the N-terminal part of the other monomer (**Fig. 1a**). Although the monomeric surface is relatively flat (**Supplementary Fig. 1a**), the dimer formation creates a concave surface at the interface with a strong positively charged potential, which is complementary to the negatively charged RNA molecule (**Fig. 1b**). The opposite side of the dimer complex is highly negatively charged (**Fig. 1b**). The dimer formation buries a surface area of 759 Å^2 for each monomer, and Zm-4 RNA binding buries additional 443-Å^2 and 195-Å^2 surface areas for two monomers, respectively (calculated with PDBePISA (http://www.ebi.ac.uk/msd-srv/prot_int/)). In each complex, only 4- or 5-nt RNA fragments of Zm4 or Zm1a-6 were resolved, even though both RNAs have 12 or 13 nt. The observed length of RNA in the structure correlates with the small size of the G-A fragment sufficient for binding to THA8 (**Supplementary Fig. 3g**). The binding of both RNAs is anchored by a G nucleotide, whose base is stacked between two tyrosine residues (Y169 and Y205) from position 3 of PPR motifs 4 and 5 (**Fig. 1c**). The G base makes extensive hydrogen-bond interactions with T172 and D203 in one monomer and with K57 and H97 from the adjacent monomer (**Fig. 1c**). The G base is nearly buried within a pocket formed by the THA8 dimer (**Fig. 1d**), and this mode of binding is well supported by the excellent electron density in both structures (**Fig. 1c**). Residues that contact the G base (K57, H97, Y169, T172, D203 and Y205) are conserved (**Supplementary Fig. 1c**), and their mutations greatly affected THA8's RNA-binding activity (**Fig. 1e**). The residues that contact the G base (T172 and D203) are from position 6 of motif 4 and position 1 of motif 5 (**Supplementary Fig. 1d**); thus, the mode of G-base contact is consistent with the proposed model of the combinatorial amino acid code[17].

Table 1 Data collection and refinement statistics

	Apo-THA8	Merged S-SAD data[b]	THA8–Zm4 RNA	THA8–Zm1a-6 RNA
Data collection				
Space group	$P6_5$	$P6_5$	$P4_12_12$	$P4_12_12$
Cell dimensions				
a, b, c (Å)	73.7, 73.7, 61.9	73.7, 73.7, 61.8	90.3, 90.3, 81.7	88.3, 88.3, 78.9
α, β, γ (°)	90, 90, 120	90, 90, 120	90, 90, 90	90, 90, 90
Resolution (Å)	50–1.6	50–2.0	50–2.8	50–3.0
R_{merge}[a]	0.042 (0.883)	0.089 (0.856)	0.05 (0.944)	0.114 (1.19)
$I / \sigma I$	30.4 (2.9)	58.9 (4.2)	34.4 (3.3)	19.2 (2.2)
Completeness (%)	100 (100)	99.8 (98.4)	100 (100)	100 (100)
Redundancy	11.3 (11.3)	67.1 (21.6)	14.1 (14.7)	10.2 (10.5)
Refinement				
Resolution (Å)	50–1.6		50–2.8	50–3.0
No. reflections	23,941		8,359	6,334
R_{work} / R_{free} (%)	21.2 / 24.4		20.5 / 25.8	20.5 / 23.2
No. atoms				
Protein	1,468		1,500	1,491
Ligand/ion			91	90
Water	179		14	15
B factors				
Protein	27.2		91.3	76.0
Ligand/ion			150.3	142.8
Water	37.1		81.1	62.8
r.m.s. deviations				
Bond lengths (Å)	0.006		0.009	0.010
Bond angles (°)	1.04		1.14	1.08

[a]Values in parentheses are for highest-resolution shell. [b]Merged data from four apo-THA8 crystals collected at 1.70 Å to measure the sulfur anomalous signal (S-SAD). All other data were collected at 0.979 Å.

Besides the base contact, the phosphate backbone of RNA is contacted by charge interactions (**Fig. 1f**) with the THA8 dimer, which forms an extensive patch of positively charged surface surrounding the THA8 dimer interface (**Fig. 1b**). We systematically mutated most of the positively charged surface residues of THA8 and found that many of these mutations disrupted THA8 RNA binding (**Supplementary Fig. 3c**), including residues R58, R94, R173, R176, K212 and R216, which directly interact with the backbone of Zm4 RNA (**Fig. 1f**). Other charge-reversal mutations (R33E, R36E, R41E, R43E, R64E, R79E, R109E, R115E, R181E and R183E) also substantially reduced THA8 and Zm1a RNA interaction (data not shown). Those residues do not directly interact with the observed short RNA sequence, and R33E, R36E, R41E and R43E residues are disordered in the

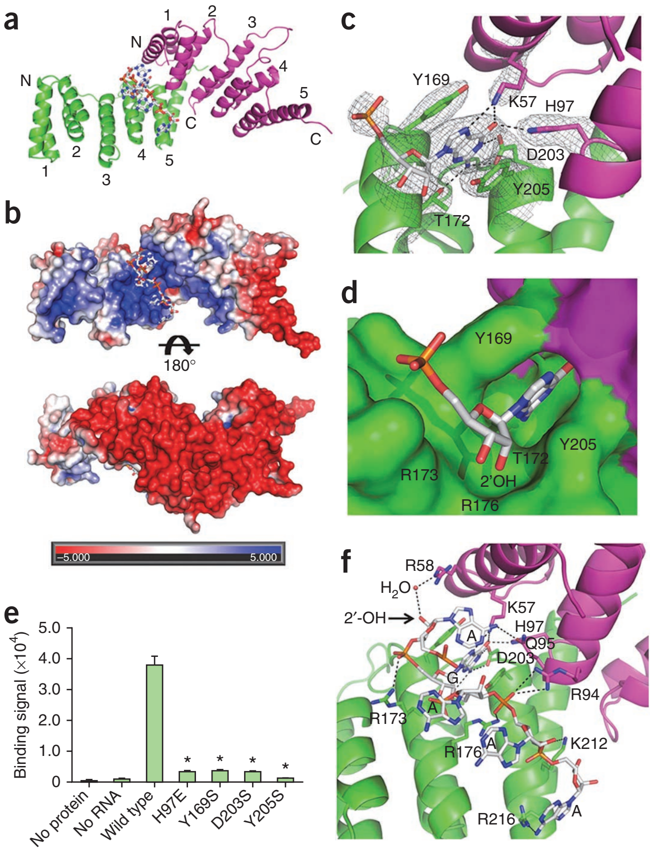


Figure 1 The structure of THA8 in complex with a 13-nucleotide Zm-4 RNA. (**a**) Two THA8 monomers, shown in cartoon representation and colored in green and magenta. The bound Zm-4 RNA fragment (AGAAA) is shown in stick model at the dimer interface. (**b**) Surface charge distribution of the two different sides of the THA8 dimer. A color-coded bar shows an electrostatic scale from –5 to +5 eV. The bound RNA fragment is shown as stick model. (**c**) Close-up view of the THA8-dimer interactions with the G nucleotide of the AGAAA motif. The carbon atoms are colored in green and magenta for two monomers that interact with Zm4 RNA. Hydrogen-bond interactions are indicated by black dashed lines. The $2F_o - F_c$ map contoured at 1σ is shown surrounding the interaction site. (**d**) Surface presentation of the conserved binding pocket for the G nucleotide at the dimer interface. (**e**) Mutational effects of the THA8 residues recognizing the G nucleotide in THA8–Zm1a RNA interaction. The binding between 10 nM biotin-Zm1a RNA and 50 nM His_6-THA8 wild-type or mutant proteins measured by AlphaScreen assay ($n = 3$; error bars, s.d.) is shown. $*P < 0.01$, compared to the wild type (Student's t test). The binding assay was repeated once with similar results. (**f**) Interaction of the Zm4 RNA fragment (AGAAA) with THA8-dimer proteins (colored in green and magenta). The carbon atoms are colored in green and magenta for two monomers. Hydrogen bonds are indicated by black, dashed lines. The RNA molecule is shown as stick models with carbon atoms colored in white.

complex structure, thus suggesting that the residues may affect the THA8-RNA interaction indirectly.

Preferential binding of single-stranded RNA by THA8

The strong preference of THA8 for single-stranded RNA (**Supplementary Fig. 4**) is readily explained by the THA8–RNA complex structures. The deep burying of the G base into the dimeric pocket would prevent the binding of double-stranded RNA or RNA–DNA hybrids (**Fig. 1d**). THA8 also greatly prefers a single-stranded RNA over DNA, as shown by the 100-fold-weaker binding of single-stranded Zm4 DNA relative to the corresponding binding of single-stranded RNA (**Supplementary Fig. 4a**). The complex structures reveal that the ribose of the G nucleotide docks nicely into the THA8 binding pocket (**Fig. 1d**), with its 2′-hydroxyl group of the G-nucleotide ribose making a direct hydrogen bond with R176. In addition, the 2′-hydroxyl of the preceding A nucleotide makes a water-mediated hydrogen bond with R58 from the neighboring THA8 (**Fig. 1f** and **Supplementary Fig. 4b**). These contacts are consistent with the strong preference of THA8 for single-stranded RNA over DNA.

RNA binding induces THA8 dimerization and oligomerization

Comparison of RNA-free and RNA-bound structures revealed major conformational changes in THA8 upon binding RNA, including a shift of α-helices from motifs 1 and 2 as well as side chain movements of four critical residues (Y169, T172, D203 and Y205) that form the G-binding pocket (**Fig. 2a**). The RNA-free THA8 structure is in a monomeric form (**Supplementary Fig. 1a**). In contrast, RNA-bound THA8 adopts an asymmetric dimer with the bound RNA at the dimer interface, thus suggesting that RNA binding may induce THA8 dimerization. To test this hypothesis, we developed an AlphaScreen assay to examine the dimer or oligomer status of THA8, using biotin- and hexahistidine (His_6)-tagged THA8 proteins (**Fig. 2b**). The biotin- and His_6-tagged THA8 proteins did not yield any binding signal in the absence of RNA but produced a strong binding signal in the presence of Zm4 RNA, thus suggesting THA8-THA8 interaction upon the RNA binding. Mutations in the key residues (H97, S99 and L102; **Fig. 2c**) at the dimer interface greatly reduced THA8 RNA binding (**Fig. 2d**), results indicating that an intact dimeric interface is required for THA8-RNA interactions.

To further validate that RNA induces THA8 dimerization, we measured the size of THA8 protein in the absence and presence of Zm4 RNA, using dynamic light scattering. The RNA-free THA8 had a radius of 3.1 nm, which increased to 5.3 nm upon binding of Zm4 RNA. This size increase is RNA specific, because the control RNA did not induce any radius change for THA8 protein (**Supplementary Fig. 5a**). Also, Zm4 RNA did not induce a radius change for the THA8 dimerization mutant (S99R) and RNA-binding-site mutant (Y169S) (**Supplementary Fig. 5a**). We also used analytical size-exclusion chromatography to examine RNA-induced THA8 dimerization. In this experiment, THA8 alone elutes as a monomer peak, and addition of Zm4 RNA induces a left shift of the peak (**Supplementary Fig. 5b**) indicating a dimer formation. In contrast, Zm4 RNA did not induce dimerization for S99R and Y169S mutants. Together, these results support that RNA is an allosteric ligand whose binding induces dimerization and conformational changes of THA8.

The asymmetric nature of the THA8 dimer indicates that it can extend to form a THA8 oligomer (**Fig. 3a**,**b**). We used a gel mobility shift assay to examine the THA8–Zm4 RNA interaction. A protein–RNA complex band was present between THA8 and Zm4 RNA but not between THA8 and HB9 RNA (a negative control), in accordance with the AlphaScreen results. Interestingly, the mobility of the band of the protein–RNA complex was further reduced with increasing concentration of THA8, thus indicating a protein concentration–dependent formation of THA8 dimers and oligomers upon RNA binding (**Fig. 3c**). We also used cross-linking to examine RNA-induced THA8 protein dimerization and oligomerization. Using a low concentration of glutaraldehyde (0.004%), we found that Zm4 RNA indeed induces THA8 protein dimerization and oligomerization (dimer, trimer and tetramer bands, **Fig. 3d**) with Zm4 RNA but not with control RNA. We also examined RNA-induced dimerization of THA8L, using the biotin- and His_6-tagged THA8L proteins. We found that Zm4 also induces THA8L

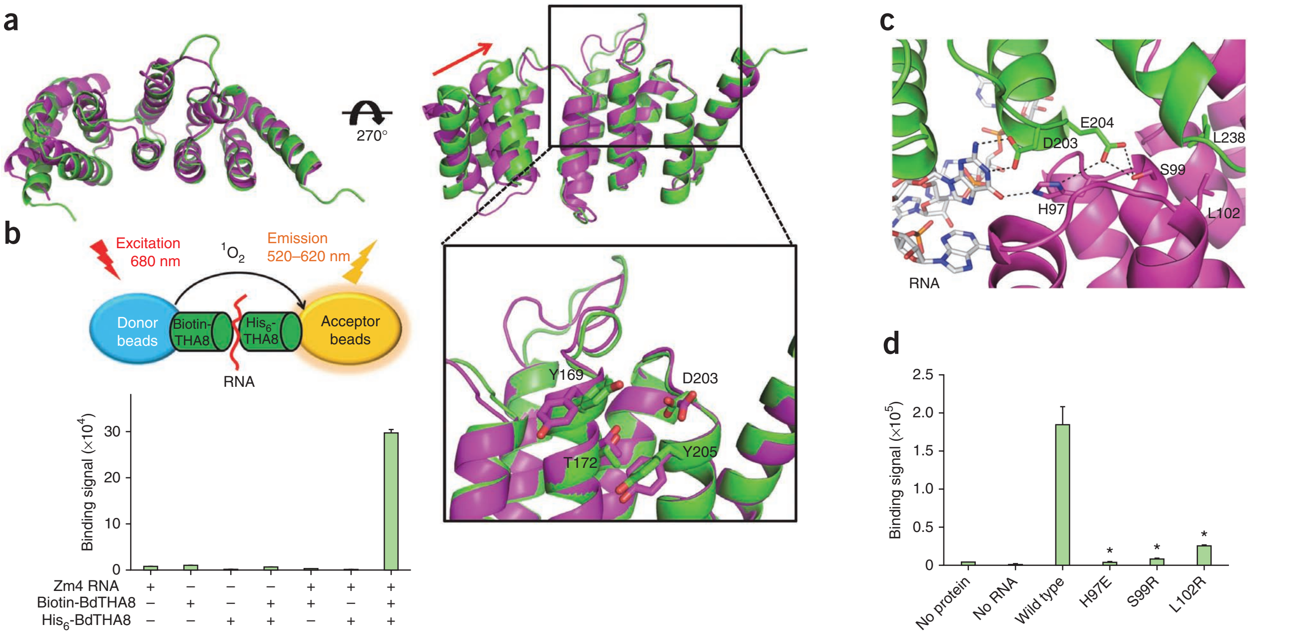
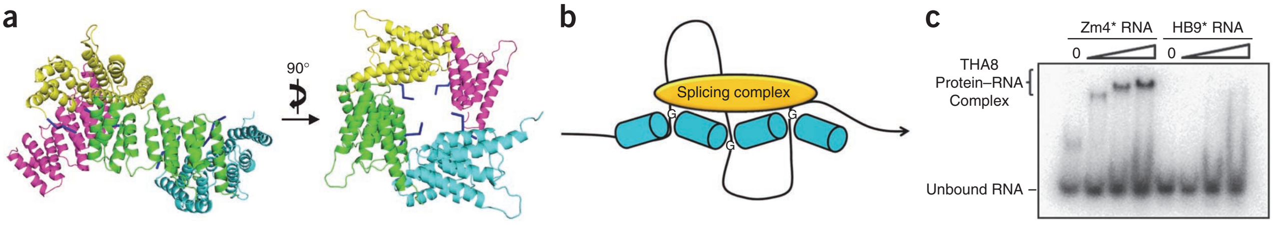


Figure 2 RNA induces conformational changes and dimerization of THA8. (**a**) Superposition of apo-THA8 (magenta) and Zm-4 RNA–bound (green) structures in cartoon representation. The arrow in red indicates the shift of the two N-terminal helices upon RNA binding. A close-up view of the THA8 RNA-binding pocket with four residues involved in binding of the G nucleotide is shown in stick models. (**b**) Induction of THA8 protein dimerization by Zm4 RNA. Binding between 100 nM biotin-THA8 and 100 nM His$_6$-THA8 in the presence or absence of 100 nM Zm-4 RNA measured by AlphaScreen binding assay ($n = 3$; error bars, s.d.) is shown. The binding assay was repeated once with similar results. (**c**) Close-up view of the THA8 dimer interface in the vicinity of the Zm-4 RNA–binding site. The carbon atoms are colored in green and magenta for the two THA8 monomers and white for RNA. Hydrogen-bond interactions are indicated by black, dashed lines. (**d**) Mutational effects of the dimer-interface residues on the THA8–Zm1a RNA interaction. The binding between 10 nM biotin-Zm1a RNA and 50 nM His$_6$-THA8 wild-type or mutant proteins measured by AlphaScreen assay ($n = 3$; error bars, s.d.) is shown. *$P < 0.01$, compared to the wild type (Student's *t* test). The binding assay was repeated once with similar results.

protein dimerization (**Fig. 3e**), thus suggesting that RNA-induced dimerization is a conserved feature of these small PPR proteins.

DISCUSSION

In this study, we have identified the THA8-binding sites present in multiple copies in the *ycf3* gene. We further characterized the preference of THA8 RNA binding for single-stranded RNA and determined the crystal structures of THA8 in both RNA-free and RNA-bound states. The structures reveal an unexpected mode of RNA binding, with the bound RNA at the asymmetric dimer interface formed by two THA8 monomers. This new mode of RNA recognition is supported by our extensive biochemical and mutagenesis data and has important



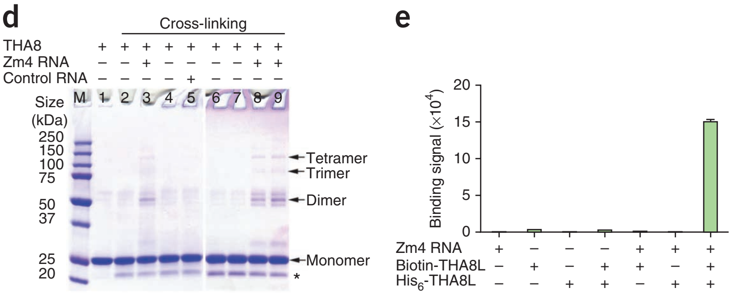


Figure 3 Zm4 RNA induces THA8 protein dimerization and oligomerization. (**a**) Side and top views of the four adjacent THA8 molecules in the THA8–Zm4 RNA complex structure. Four molecules are shown in cartoon representation and are colored in magenta, yellow, green and cyan. Four AGAAA RNA fragments are shown as blue ribbon models. (**b**) A proposed model for the interaction of THA8 with pre-mRNA. The single-stranded pre-mRNA chain wraps around the THA8 oligomer, thus bringing several discontinuous regions of RNA into proximity to facilitate the splicing of the intron by other splicing factors present in the splicing complex. Four cylinders represent four THA8 monomers. The orange oval represents other protein factors present in the splicing complex. (**c**) Gel mobility shift assay to detect Zm4 RNA–THA8 interaction. Asterisk indicates radiolabel. (**d**) SDS-PAGE analysis of THA8 protein in the presence or absence of the indicated RNA, cross-linked with 0.004% glutaraldehyde. (**e**) Zm4 RNA induction of THA8L protein dimerization. Binding was measured similarly as for THA8 by AlphaScreen binding assay ($n = 3$; error bars, s.d.). The binding assay was repeated once with similar results.

implications for the functional assembly of THA8–RNA complexes in the splicing of the *ycf3* introns.

The most important feature of RNA recognition by THA8 is that RNA is bound at the dimer interface formed by asymmetric packing of two monomers. The short PPR proteins (five PPR repeats) in the THA8 family are distinct from other longer PPR proteins, such as PPR5 (ref. 11; ten PPR repeats), PPR10 (refs. 9,10; 19 PPR repeats), CRP1 (ref. 11; 14 PPR repeats) and HCF152 (ref. 12; 12 PPR repeats). These PPR proteins have high RNA-binding capacity and have been reported to specifically bind longer RNA sequences, a result consistent with the RNA code prediction[17]. Instead, THA8 protein binds to a short purine-rich RNA sequence. For small PPR proteins, the capacity for RNA binding is inherently small. Structurally, each THA8 monomer has a roughly flat, rectangular shape that is not optimal for binding RNA. Dimerization creates a concave surface with strong positive charge potential that is optimal for binding short RNA fragments with strong negative charges. The dimer formation also increases the binding affinity of THA8 for RNA by engaging more residues from both molecules for contacting the G nucleotide. We have recently reported the crystal structure of another small PPR protein, THA8L[18], which shares 26% sequence identity with THA8 (**Supplementary Fig. 1c**). THA8 and THA8L share a similar structure and RNA binding specificity[18]. Importantly, RNA binding also promotes THA8L dimerization (**Fig. 3e**). Thus, RNA-induced dimerization and oligomerization may be a conserved feature for the THA8 family of small PPR proteins.

RNA induced THA8 dimerization and oligomerization has important implications for THA8 protein function. THA8 is involved in the splicing of two specific group II introns *in vivo* (in *ycf3* and *trnA*)[3]. Splicing of group II introns is a complex process involving protein–RNA complexes containing many splicing factors. Indeed, THA8 works with at least two other splicing factors, WTF1 and RNC1, for the splicing of the *trnA* intron[3]. That the *ycf3* intron II contains multiple binding sites for THA8 and that THA8 adopts the asymmetric-dimer packing mode suggest that THA8 binds to pre-mRNA as an oligomer. This binding mode allows assembly of THA8 onto pre-mRNA, and this assembly may induce the long linear RNA to form a condensed structure with several loop structures (**Fig. 3b**). This would bring discontinuous regions of intron RNA into proximity for interaction with other splicing factors in the splicing complex for efficient splicing.

The RNA binding mode of dimeric and oligomeric THA8 is distinct from the modular recognition of RNA by PUF[14,15] and DNA by TALE[21,22] proteins. Each repeat from a TALE protein recognizes a specific DNA base by using two hypervariable residues at positions 12 and 13. Each PUF repeat uses residues at positions 12 and 16 to interact with an RNA base through hydrogen bonds and residues at position 13 to make base-stacking interactions. We anticipate that PPR proteins may also recognize the target RNAs in a modular fashion.

We found a specific interaction of THA8 with only the G nucleotide, which was mainly contacted by PPR motifs 4 and 5 of one THA8 molecule. The key residues involved in recognition of the G nucleotide (T172 and D203) are entirely consistent with the RNA code prediction[17]. However, the N-terminal PPR motifs (motifs 1–3) of THA8 may have lost their base-specific interaction capacity during evolution and may be used mainly to assist motifs 4 and 5 of another molecule to form the RNA-binding pocket. We speculate that the RNA-recognition capacity for THA8 is inherently small because of its small size, and dimerization and oligomerization represent a key mechanism for such a small PPR protein to increase its binding affinity to RNA. We anticipate that a structure of a PPR protein with more PPR motifs in complex with its target RNA will elucidate more base-specific interactions between the PPR protein and RNA.

In summary, we have presented comprehensive biochemical and structural studies of THA8, a prototypical PPR protein with a defined phenotype and molecular function in RNA splicing[3]. The RNA-free THA8 structure reveals five tandem PPR motifs, a result consistent with our identification of short G-A–rich sequences as the THA8-binding sites in RNA. Importantly, the THA8-binding sites are conserved and are present in multiple copies in the *ycf3* intron. The structures of THA8 in complex with two different RNA fragments from the *ycf3* intron not only validate the previously proposed recognition code for a G nucleotide but also reveal an unexpected dimeric association of THA8 for RNA binding. The detailed interactions with the G base, the ribose and the phosphate backbone provide a basis for selective recognition of single-stranded RNA by THA8 over other types of nucleic acids. The THA8 dimer complex formation is mediated by the N-terminal PPR motifs 1 and 2 and the C-terminal motifs 4 and 5 (**Fig. 3a**). This asymmetric dimer interface, which would allow THA8 to assemble into oligomers, may thus help to organize and condense the *ycf3* intron for splicing because of its multiple THA8-binding sites (**Fig. 3a**,**b**). The unexpected observation of RNA binding by the dimeric THA8, together with the extensive biochemical and mutational data, provides a structural basis for rationalizing RNA recognition by short PPR proteins and for designing sequence-specific RNA-binding proteins.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Atomic coordinates and structure factors have been deposited in the Protein Data Bank, under accession codes 4ME2 (apo-THA8), 4N2Q (THA8–Zm4 complex) and 4N2S (THA8–Zm1a-6 complex).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank staff members of the Life Science Collaborative Access Team of the Advanced Photon Source (APS) for assistance in data collection at the beamlines of sector 21, which is in part funded by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (grant 085P1000817). Use of APS was supported by the Office of Science of the US Department of Energy, under contract no. DE-AC02-06CH11357. This work was supported by the Jay and Betty Van Andel Foundation, the Ministry of Science and Technology (China) (grants 2012ZX09301001-005 and 2012CB910403), Amway (China), the US National Institutes of Health (grant R01 DK071662 to H.E.X.) and by the Chinese Academy of Sciences (J.-K.Z.).

AUTHOR CONTRIBUTIONS

H.E.X. and J.-K.Z. conceived of the study; H.E.X., J.K., J.-K.Z. and K.M. supervised the study; J.K., R.-Z.C., T.B., X.G., M.H.E.T., C.C. and Y.K. performed experiments of RNA binding, protein expression, purification and crystallization; J.K. and J.S.B. carried out data collection; J.K. and X.E.Z. performed model building, refinement and data analysis; and H.E.X. and J.K. wrote the manuscript with contribution from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Delannoy, E., Stanley, W.A., Bond, C.S. & Small, I.D. Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. *Biochem. Soc. Trans.* **35**, 1643–1647 (2007).

2. Schmitz-Linneweber, C. & Small, I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends Plant Sci.* **13**, 663–670 (2008).
3. Khrouchtchova, A., Monde, R.A. & Barkan, A. A short PPR protein required for the splicing of specific group II introns in angiosperm chloroplasts. *RNA* **18**, 1197–1209 (2012).
4. Lurin, C. *et al.* Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16**, 2089–2103 (2004).
5. O'Toole, N. *et al.* On the expansion of the pentatricopeptide repeat gene family in plants. *Mol. Biol. Evol.* **25**, 1120–1128 (2008).
6. Ringel, R. *et al.* Structure of human mitochondrial RNA polymerase. *Nature* **478**, 269–273 (2011).
7. Howard, M.J., Lim, W.H., Fierke, C.A. & Koutmos, M. Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5′ processing. *Proc. Natl. Acad. Sci. USA* **109**, 16149–16154 (2012).
8. Small, I.D. & Peeters, N. The PPR motif: a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci.* **25**, 46–47 (2000).
9. Pfalz, J., Bayraktar, O.A., Prikryl, J. & Barkan, A. Site-specific binding of a PPR protein defines and stabilizes 5′ and 3′ mRNA termini in chloroplasts. *EMBO J.* **28**, 2042–2052 (2009).
10. Prikryl, J., Rojas, M., Schuster, G. & Barkan, A. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proc. Natl. Acad. Sci. USA* **108**, 415–420 (2011).
11. Williams-Carrier, R., Kroeger, T. & Barkan, A. Sequence-specific binding of a chloroplast pentatricopeptide repeat protein to its native group II intron ligand. *RNA* **14**, 1930–1941 (2008).
12. Nakamura, T., Meierhoff, K., Westhoff, P. & Schuster, G. RNA-binding properties of HCF152, an *Arabidopsis* PPR protein involved in the processing of chloroplast RNA. *Eur. J. Biochem.* **270**, 4070–4081 (2003).
13. Abbas, Y.M., Pichlmair, A., Gorna, M.W., Superti-Furga, G. & Nagar, B. Structural basis for viral 5′-PPP-RNA recognition by human IFIT proteins. *Nature* **494**, 60–64 (2013).
14. Filipovska, A., Razif, M.F., Nygard, K.K. & Rackham, O. A universal code for RNA recognition by PUF proteins. *Nat. Chem. Biol.* **7**, 425–427 (2011).
15. Wang, X., McLachlan, J., Zamore, P.D. & Hall, T.M. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**, 501–512 (2002).
16. Zhelyazkova, P. *et al.* Protein-mediated protection as the predominant mechanism for defining processed mRNA termini in land plant chloroplasts. *Nucleic Acids Res.* **40**, 3092–3105 (2012).
17. Barkan, A. *et al.* A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet.* **8**, e1002910 (2012).
18. Ban, T. *et al.* Structure of a PLS-class pentatricopeptide repeat protein provides insights into mechanism of RNA recognition. *J. Biol. Chem.* doi:10.1074/jbc.M113.496828 (18 September 2013).
19. Takenaka, M., Zehrmann, A., Brennicke, A. & Graichen, K. Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. *PLoS ONE* **8**, e65343 (2013).
20. Yagi, Y., Hayashi, S., Kobayashi, K., Hirayama, T. & Nakamura, T. Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PLoS ONE* **8**, e57286 (2013).
21. Deng, D. *et al.* Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**, 720–723 (2012).
22. Mak, A.N., Bradley, P., Cernadas, R.A., Bogdanove, A.J. & Stoddard, B.L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716–719 (2012).

ONLINE METHODS

Cloning and mutagenesis. The *tha8* cDNAs from different species, corresponding to maize THA8 residues 31–257, were synthesized by Genewiz and were cloned into pET24a vector (Novagen) with a His_6-Sumo tandem fusion tag at the N terminus and an ULP1 protease cleavage site between the His_6-Sumo tag and THA8. Site-directed mutagenesis was carried out with the QuikChange method (Stratagene) or the GeneTailor System (Invitrogen). Mutations and all plasmid constructs were confirmed by sequencing before protein expression.

Protein expression and purification. For protein expression, *E. coli* BL21 (DE3) cells transformed with pET24a His_6-Sumo–THA8 vector were used to inoculate 4 l of LB media. When cell density reached an OD_{600} of ~1.0, protein expression was induced with 0.1 mM IPTG at 16 °C overnight. Cells were harvested, resuspended in 50 ml buffer A (20 mM Tris, pH 8.0, 200 mM NaCl and 10% glycerol) per 2 l of cells and lysed by an APV2000 cell homogenizer (SPX corporation). The lysate was centrifuged at 16,000 r.p.m. for 30 min, and the supernatant was loaded on a 50 ml Ni-chelating Sepharose column (GE Healthcare). The column was washed with 90% buffer A and 10% buffer B (20 mM Tris, pH 8.0, 200 mM NaCl, 500 mM imidazole and 10% glycerol) and eluted with 50% buffer A and 50% buffer B. The eluted His_6-Sumo–THA8 protein was dialyzed against buffer A and cleaved overnight at 4 °C with ULP1 at a protease/protein ratio of 1:500. The cleaved His_6-Sumo tag was removed by passage through a 5-ml Ni-chelating Sepharose column (GE Healthcare), and the protein was further purified by a 120 ml Superdex gel-filtration column in buffer C (25 mM Tris, pH 8.0, 200 mM ammonium acetate, 1 mM dithiothreitol and 1 mM EDTA). The protein eluted from the gel-filtration column at a volume corresponding to the size of a monomer at a purity >95% as judged by SDS-PAGE. For preparation of biotinylated THA8 protein, the *tha8* cDNA (codons 31–257) was cloned into the first cloning site of pET-Duet1 vector with a His_6-Sumo tag at the N terminus and a biotin acceptor peptide sequence (AviTag) at the C terminus. The second cloning site included the coding sequence of the biotinylation enzyme BirA. His_6-Sumo–THA8 fusion protein was coexpressed with BirA in BL21(DE3) cells in the presence of 40 μM biotin to allow *in vivo* biotinylation of THA8 protein. The fusion protein was purified by Ni-NTA chromatography, proteolytic release of the His_6-Sumo tag and size-exclusion chromatography as described for the wild-type protein.

Crystallization. Purified THA8 protein was concentrated to about 10 mg/ml before crystallization trials. Initial screening with PEG Rxn HT (Hampton Research) identified that polyethylene glycol (PEG) is favorable for crystal formation. Optimization trays with PEG were set up manually with the sitting-drop method at 20 °C. Rod-shaped crystals of ~200 μm in length were obtained with 0.5 μl of the purified protein and 0.5 μl of well solution (18% PEG 1500, 0.1 M HEPES, pH 7.5, and 0.2 M L-proline). These crystals diffracted X-rays to 1.6–2.0 Å at LS-CAT of the Advanced Photon Source (APS) synchrotron. To prepare the THA8–RNA complexes, we mixed THA8 protein with different RNAs at a molar ratio of 1:1 and tested them in crystallization trials with commercial screens (Hampton Research). Among different THA8–RNA crystals, only crystals of THA8 in complex with either Zm-4 or Zm1a RNA diffracted X-rays to 2.8–3.0 Å, with a well solution of 14% PEG 4000 and 0.1 M MES, pH6.0.

Data collection and structure determination. All crystals were transferred to well solution with 22% (v/v) ethylene glycol as cryoprotectant before flash freezing in liquid nitrogen. Data collection was performed at sector 21-ID-D (LS-CAT) of the APS synchrotron. A native data set was collected to 1.6 Å. Initial structure determination by molecular replacement with the PPR domain from the crystal structure of proteinaceous RNase P 1 (PDB 4G23, which shares about 20% sequence identity with THA8) as a search model failed to yield any correct solution. To solve the phase problem, four data sets of native THA8 crystals were collected at a wavelength of 1.70 Å to measure the sulfur anomalous signal. These four data sets were processed with XDS[23], combined with Pointless, and merged with Scala of the CCP4 suite[24] (S-anomalous data), as previously described[25]. Initial phases were established by SHELX[26] with the native data and the S-anomalous data (**Table 1**). Five sulfur atoms were found by SHELXD with a CCall/CCweak score of 41.4/23.7, and subsequent phasing was performed with SHELXE. Density modification for the initial electron density map was performed with DM[27]. A crude model was built automatically with the CCP4 program buccaneer with R/R_{free} of 0.430/0.464. The model was further improved by Phenix autobuild[28] and by several cycles of manual building with Coot[29]. The model was further refined against the native data with Refmac of CCP4 (ref. 30) to an R factor of 0.21 and an R_{free} factor of 0.24 (**Table 1**). The diffraction data for two complex crystals were collected similarly as for native crystals. The THA8–Zm4 RNA and THA8–Zm1a-6 RNA complex structures were solved by molecular replacement with the apo-THA8 structure as a search model. The models were refined with Refmac, and the RNA fragments were built on the basis of the electron density with Coot. Several cycles of manual model building and refinement were performed to refine the complex structure of THA8–Zm4 RNA to an R factor of 0.21 and an R_{free} factor of 0.26 and the complex structure of THA8–Zm1a-6 RNA to an R factor of 0.21 and an R_{free} factor of 0.23 (**Table 1**).

Wavelengths of data collection are shown in the footnote of **Table 1**. All data were collected at 100 K at the LS-CAT beamlines (Sector 21). Ramachandran statistics are as follows: apo-THA8 structure, 95.4% in favored regions, 4.6% in additional allowed regions, 0% generously allowed regions; THA8–Zm4 RNA complex structure, 92.0% in favored regions, 8.0% in additional allowed regions, 0% generously allowed regions; THA8–Zm1a-6 RNA complex structure, 93.1% in favored regions, 6.9% in additional allowed regions, 0% generously allowed regions.

AlphaScreen binding assay. Interactions between His_6-tagged THA8 and biotin-RNA were assessed by luminescence-based AlphaScreen technology (Perkin Elmer) with a hexahistidine detection kit that our group has used extensively[31]. Biotin-RNAzm1a (Gene Pharma) was attached to streptavidin-coated donor beads, and His_6-tagged THA8 was attached to nickel-chelated acceptor beads. The donor and acceptor beads were brought into proximity by the interaction between His_6-tagged THA8 and biotin-RNAzm1a. The binding mixtures, containing the indicated amounts of protein and RNA and 5 μg/ml of streptavidin-coated 'donor' beads and Ni-chelate–coated 'acceptor' beads, were incubated in 50 mM MOPS, pH 7.4, 100 mM NaCl, and 0.1 mg/ml BSA for 2–3 h before data collection with an Envision plate reader (PerkinElmer). For competition assays, increasing concentrations of unlabeled nucleic acids were added in addition to the labeled protein and RNA. Each data point was an average of triplicate measurements, with standard errors indicated. For competition binding assays, the IC_{50} values were derived from curve fitting based on a competitive-inhibitor model with GraphPad Prism.

Gel mobility shift assay. A gel mobility shift assay was performed to detect RNA binding by THA8 protein. Both Zm4 and control HB9 RNA oligoribonucleotides were 5′-end-labeled with [γ-^{32}P]ATP by T4 polynucleotide kinase according to the manufacturer's protocol (Invitrogen). The labeled RNA probes were separated from unincorporated nucleotides by centrifugation with G-25 quick-spin columns (Roche). The binding reaction contained 25 mM Tris, pH 8.0, 0.125 mM EDTA, 10% glycerol, 25 mM KCl, 1 ng of labeled RNA oligonucleotides and increasing amounts of THA8 protein (0, 0.4, 1.2 and 4 μM). Binding reactions were incubated for 30 min at room temperature and resolved on 6% native polyacrylamide gel running in 0.5× TBE buffer at 100 V. Results were visualized on a PhosphorImager (FujiFilm).

Analytical size-exclusion chromatography. Analytical size-exclusion chromatography was performed with Agilent 1260 Infinity (Agilent Technologies). The column was run on a Waters HPLC system at a flow rate of 0.35 ml min^{-1} with dual-mode detection at 280 and 220 nm. The column was equilibrated with running buffer (20 mM Tris, pH 7.5, 50 mM NaCl and 2% glycerol) to obtain a stable baseline. After that, 20 μl of ~1.0 mg/ml THA8, THA8/RNA mixture, S99R, Y169S mutations and RNA mixture solutions incubated for 30 min in ice were centrifuged, and then the supernatant was loaded on the column to detect the elution time of the main peak and the association state of THA8.

Dynamic lighting scattering. THA8, THA8/RNA mixture, S99R- and Y169S-mutation solutions were diluted to a final concentration of ~1.0 mg/ml. The experiments were performed at a preadjusted temperature of 25 °C. The sizes of THA8 and PPR/RNA complexes were measured by dynamic light scattering (DLS) with a DynaPro NanoStar (Wyatt). The DLS technique measures the time-dependent fluctuation in the intensity of scattered light that occurs

doi:10.1038/nsmb.2710

because of the motion of the particles. The analysis of these fluctuations enables the determination of the translational diffusion coefficients of particles, which can be transformed to a size distribution. The radii of particles were calculated according to the instructions of the instrument.

Protein cross-linking. THA8 proteins at 0.25 mg/ml in 1× PBS were incubated with or without 0.5 nmol of Zm4 RNA or HB9 RNA (a negative control) for 30 min and then cross-linked with 0.004% glutaraldehyde for 25 min at room temperature. The reactions were stopped with Tris-glycine buffer and analyzed by SDS-PAGE.

23. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
24. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760–763 (1994).
25. Liu, Q. *et al.* Structures from anomalous diffraction of native biological macromolecules. *Science* **336**, 1033–1037 (2012).
26. Sheldrick, G.M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 479–485 (2010).
27. Cowtan, K. dm: an automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 34–38 (1994).
28. Terwilliger, T.C. *et al.* Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 582–601 (2009).
29. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
30. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240–255 (1997).
31. Melcher, K. *et al.* A gate-latch-lock mechanism for hormone signalling by abscisic acid receptors. *Nature* **462**, 602–608 (2009).

doi:10.1038/nsmb.2710