Chapter 9

Protein Modification for Crystallization

Toshio Hakoshima

Abstract

Technological advances in data collection with synchrotron radiation sources and phasing methods including automated model building and validation have highlighted crystallization as the rate-limiting step in X-ray diffraction studies of macromolecular structures. Although protein crystallization remains a stochastic event, protein engineering with the advent of recombinant methods enables us to generate target proteins possessing a higher propensity to form crystals suitable for X-ray diffraction data collection. This chapter presents an overview of protein engineering methods designed to enhance crystallizability and discusses examples of their successful application.

Keywords Protein engineering, Recombinant DNA technology, Flexible loops, Non-conserved insertion, Secondary structure prediction, Natural variation, Artificial linker

1 Introduction

Advanced recombinant technology and biochemical installations significantly reduce efforts required for protein production and purification. Moreover, the use of superb crystallization screening kits coupled with high-performance crystallization robots has changed previously laborious trial-and-error crystallization experiments to routine laboratory work that can be executed by specialist and nonspecialist researchers alike. However, the preparation of single well-diffracting crystals of the target proteins remains a time-consuming challenge. Once single well-diffracting crystals have been obtained, however, X-ray data collection using synchrotron radiation and phasing of the intensity data followed by structural refinement are relatively straightforward tasks.

Two approaches have been employed to improve protein crystal quality and size. Firstly, natural variation in amino acid sequences of homologues or homologues from different species can be exploited to identify a target with suitable crystallization properties. Alternatively, artificial modification of target proteins by the use of recombinant techniques can be employed to enhance the

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_9, © Springer Japan 2016

target protein's propensity to crystallize or to improve the diffraction quality of the resulting crystals. In this review, we firstly introduce some examples of the use of homologue proteins to demonstrate the impact of natural sequence variations on crystallizability and crystal lattices and then discuss current progress in protein engineering methodologies used to improve the crystal quality of target proteins that are recalcitrant to crystallization in their wild-type form. Protein engineering methodologies can execute internal deletion of non-conserved flexible loops in addition to frequently used N- and C-terminal truncations, in addition to the use of fusion proteins between tags and target proteins and between ligands and target proteins. Although these approaches require preliminary optimization screening, the screening procedures are fairly well established and therefore can be routinely performed to obtain diffraction-quality crystals.

2 Methods

2.1 Homologous Proteins

Historically, natural variations in the amino acid sequences of homologues were exploited to identify targets with suitable crystallization properties during the purification procedure [1, 2]. Extensive application of this approach resulted in a scramble to report on the first structure determination of transcription factors in 1990s. How much variation in the sequences is needed to enhance the propensity to crystallize or to improve the diffraction quality? Human and mouse genomes share well-conserved sequences of their homologues with high amino acid identity (>90 %), and their conserved functionally important domains display high identity (>95 %). These high sequence identities significantly decrease the possibility of significant improvements in crystallization or crystal quality. In practice, homologues with less than 80 % identity are potential targets for improvement trials. Recent examples are mammalian T-lymphoma invasion and metastases 1 and 2 (Tiam1 and Tiam2), which are Rac-specific guanine exchange factors (Rac-GEFs) [3, 4], and dwarf 14 (D14) and related proteins, which are plant hormone receptor candidates [5].

Tiam1 possesses a novel functional PHCCEx domain (~30 kDa) for plasma membrane association and specific binding to a class of membrane proteins. The domain boundary of the mouse Tiam1 PHCCEx domain was delineated following extensive screening of expression constructs, since several constructs produced proteins that were easily degraded during the protein purification steps. The optimized construct produced a stable protein sample that was successfully crystallized in the form of needlelike crystals of a hexagonal lattice ($P6_422$, $a = b = 113.5$ Å, $c = 113.8$ Å, $\gamma = 120°$), although the crystals diffracted poorly up to 4.5 Å using synchrotron radiation at SPring-8. Several trials to improve

the diffraction by changing conditions or using additives were unsuccessful. Human and mouse Tiam1 PHCCEx domains share high sequence identity (>90 %). Thus, focus was then set on Tiam2, which is a functional homologue of Tiam1 with 65 % sequence identity. Our sequence alignment showed that the Tiam2 PHCCEx domain possesses no large insertion or deletion compared with Tiam1, suggesting that the sequence variation is relatively high but suitable for possible improvement. With this sequence variation, protein samples of the Tiam2 PHCCEx domain were purified in a similar manner to that of Tiam1. Crystallization screening of the Tiam2 PHCCEx domain yielded two crystal forms, chunky crystals of tetragonal ($P4_32_12$, $a = b = 105.6$ Å, $c = 287.6$ Å) and monoclinic ($P2_1$, $a = 46.7$ Å, $b = 104.8$ Å, $c = 116.0$ Å, $\beta = 80.6°$) lattices. The tetragonal crystals diffracted at 3.2 Å and the monoclinic up to 2.08 Å, which are sufficient for structure determination and detailed characterization of the molecular structure.

D14 and related D14-like (D14L) proteins belong to an α/β hydrolase family based on amino acid sequences and are candidates for strigolactone and karrikin receptors, respectively. *Arabidopsis thaliana* (*At*) and *Oryza sativa* (*Os*, rice) D14 share 74 % amino acid identity. The recombinant protein of *At*D14 was easily prepared as a soluble protein and concentrated to 20 mg/mL to yield crystals following conventional crystallization screening. However, the diffraction limit of these crystals was around 4 Å, the mosaicity was large, and the diffraction spots appeared as streaks. Compared with *At*D14, *Os*D14 possesses an additional non-conserved sequence of 54 residues at the N-terminus. This N-terminal non-conserved extension contains many Gly and Ser residues and was predicted as an intrinsically disordered random coil. In general, *Os* proteins often possess such additional sequences predicted to form random coils. N-terminal truncated *Os*D14 (Δ54) could be prepared as a soluble protein, although its solubility was poor and the maximum concentration was 3 mg/mL. Despite the limited suitability for structural work, the orthorhombic crystals ($P2_12_12_1$, $a = 48.0$ Å, $b = 88.2$ Å, $c = 121.2$ Å) of *Os*D14 (Δ54) diffracted at 1.45 Å. D14L is also referred to as KARRIKIN INSENSITIVE 2 (KAI2) and shares about 50 % amino acid identity with D14. The recombinant protein of *At*D14L was efficiently expressed, easily purified, concentrated to 20 mg/mL, and crystallized in monoclinic crystals ($P2_1$, $a = 51.0$ Å, $b = 55.6$ Å, $c = 53.1$ Å, $\beta = 115.8°$) that diffracted up to 1.15 Å.

2.2 Internally Truncated Proteins

As already mentioned in the case of *Os*D14, N- and/or C-terminal truncation(s) can be frequently implemented in an effort to improve protein properties such as stability, solubility, and crystallizability. This approach could be extended to internal loop regions that may prevent crystallization of the target proteins. One recent

example is the C-terminal cargo-recognition domain comprising myosin tail homology 4 (MyTH4) and 4.1/ezrin/radixin/moesin (FERM) subdomains, the so-called MyTH4–FERM cassette, found in nonconventional myosins [6]. A DNA fragment encoding the MyTH4–FERM cassette (residues 1486–2058) of human myosin-X cloned into the pET47b [+] vector (Novagen) produced a soluble protein, although this protein was unstable and suffered partial degradation during purification. No crystals were obtained from the purified sample. A protease labile region was found in the FERM domain. Compared with the canonical FERM domain from ezrin/radixin/moesin (ERM) proteins, the FERM domain of the myosin-X MyTH4–FERM cassette contains a non-conserved insertion of ca. 60 residues (1850–1910) located between α2B and α3B helices (Fig. 1). Using time-of-flight mass spectroscopy (TOF-MS), we identified the cleavage site at S1892-F1893, which was located within the non-conserved insertion. We designed S1892A and F1893A mutants to prevent this partial degradation. However, these mutant proteins were still degraded during purification. Next, we designed truncated proteins comprising deletion of residues forming the internal non-conserved insertion. Nucleotides encoding residues 1845–1891 (Δ47), 1872–1891 (Δ20), or 1882–1891 (Δ10) of the non-conserved insertion were deleted from the plasmid using inverse PCR. To test the cargo-binding affinity of these internally truncated MyTH4–FERM cassettes, we performed pull-down assays with a GST-fused netrin receptor, deleted in colorectal cancer (DCC), which is a myosin-X cargo protein. We found that Δ47 possessed reduced affinity, while Δ20 and Δ10 showed retained affinity. The MyTH4–FERM cassettes (Δ20) were successfully purified without degradation, crystallized in a monoclinic lattice ($P2_1$, $a = 185.2$ Å, $b = 49.6$ Å, $c = 94.0$ Å, $\beta = 116.7°$), and diffracted at 1.9 Å resolution. The complex between the MyTH4–FERM cassettes (Δ20) and DCC was also crystallized in a related lattice ($P2_1$, $a = 85.4$ Å, $b = 49.5$ Å, $c = 93.4$ Å, $\beta = 117.1°$) and diffracted at 2.2 Å resolution.

Interestingly, an independent structural work of a fusion protein between the myosin-X MyTH4–FERM cassettes and DCC showed that, after extensive trials, deletion of a 36-residue fragment (residues 1871–1906) within the non-conserved insertion was necessary for crystallization [7]. Moreover, another structural work of the MyTH4–FERM cassettes of myosin VIIA, which is a close homologue of myosin-X, showed that a 30-residue deletion (residues 1037–1066) in the MyTH4 domain but not in the FERM domain was necessary for crystallization of the cassette bound to Sans [8]. This 30-residue deletion is part of the non-conserved long insertion (residues 1030–1080) between helices α1M and α2M, compared with the myosin-X MyTH4 domain.

Internal deletion was also explored in recent structural work of the yeast Ire kinase-nuclease domain [9]. Ire1 is an ancient

```
mRadixin        MPKPINVRVTTMDAELEFAIQ-PNTTGKQLFDQVVKTVGLREVW---FFGLQYVDSKGYSTW-- 52
sfMoesin        MPKSMNVRVTTMDAELEFAIQ-QTTTGKQLFDQVVKTIGLREVW---FFGLQYTDSKGDLTW--
hsMyoX          ---MTSTVYCHGGGSCKITIN-SHTTAGEVVEKLIRGLAMEDSR--NMFALFEY--NGHVDKAIE 1756
XeMyoX          -SHMTTSVYCHGGGSCQISIN-SHTTAGEVVEKLIRGLSMDNSR--NMFALFEH--NKHTDRAVE
DmMyoX          -RSARRQIYRLPGGAERVVNTRCSTVVADVIAELCALLGVESEAEQQEFSLYCIVQGDAFTMPLA

mRadixin        LKLNKKVTQQDVKKEN----------PLQFKFRAKFFP-EDVSEELIQEITQRLFFLQVKEAI 110
sfMoesin        IKLYKKVMQQDVKKEN----------PLQFKFRAKFYP-EDVADELIQEITLKLFYLQVKNAI
hsMyoX          SRTVVADVLAKFEKLAATSEVGDL--PWKFYFKLYCF--LDTDNVPKDSVEFAFMFEQAHEAV 1815
XeMyoX          SRVIVADVLAKFERLAGTGDEEDDLGPWNLYFKLYCF--LDVQSVPKEGIEFAFMFEQAHESL
DmMyoX          ADEYILDVTTELLKSGQ---------PFYLIFCRSVWHFALKREPAPMPLYVEVLFNQVAPDYLEGLLLELPGN

mRadixin        LNDEIY-------CPPETAVLLASYAVQAKYGDYN----------
sfMoesin        LSDEIY-------CPPETSVLLASYAVQARHGDHN----------
hsMyoX          IHGHHP-------APEENLQVLAALRLQYLQGDYT----HAAIPPLEEVYSLQRLKARISQSTKTFTPCE 1875
XeMyoX          TSGHFP-------APEETLQHLAALRLQYQHGDFS----V--TWSLDTVYPVQRLKAKILQATKSSTSGH
DmMyoX          LEGLLLELPGNGVPVPEMVRDMARIAALLHRAADL----------

mRadixin        ----------------YLANDKEIHKPRLLPQRVLEQHKLTKEQWEERIQNWHEEHRGML 183
sfMoesin        ----------------FLANDPAVHGPRLLPQRVTDQHKMSREEWEQSITNWQEHRGML
hsMyoX          RLEKRRSELEGTLRSFRTGSVVRQKVEEEQMLDMWIKEEVSSARASIIDKWRKFQGMN 1935
XeMyoX          TLERRRTSFLEGTLKGFKVGSMRKQKVEEEQMMEMWVKEELSAARTSIAEKWSRLQGVS
DmMyoX          ----------------SHVPAMKEIKFLLPKPALGIREIRPAQWVGLVQSAWPQVANLS

mRadixin        REDSMMEYLKIAQDLEMYGVNYFEIKNKKG--------------TELWLGVDALGLNIYEHDDKLTPKIGF 240
sfMoesin        REDAMMEYLKIAQDLEMYGVNYFEIRNKKN--------------TELWLGVDALGLNIYEKDDKLTPKIGF
hsMyoX          QEQAMAKYMALIKEWPGYGSTLFDVECKEGGFP-----------QELWLGVSADAVSVYKRGEG-RPLEVF 1994
XeMyoX          QHQAMVKYMAIVSEWPGYGPTLFDVEYKEGGFP-----------NDLWLGVSAENVSVYKRGDA-KPLETF
DmMyoX          PGQVKAQFLNVLATWPLFGSSFFAVKRIWAEEGPHVEDNHSPMWRDLILALNRRGVLFLDPNTH-ETLQHW

mRadixin        PWSEIRNISFND--------KKFVIKPIDKKAPDFVFYAPRLRINKRILALCMGNHELYMRRRKPDTI 300
sfMoesin        PWSEIRNISFN--------DRKFIIKPIDKKAPDFVFFAPRVRVNKRILALCMGNHELYMRRRKPDTI
hsMyoX          QYEHILSFGAP---------ANTYKIVVDERELLFETSEVVDVAKLMKAYISMIVKKRYSTTRSASSQ 2054
XeMyoX          QYEHIIFFGAP--------QPNTFKITVDDRELFFETTQVGEITKIMRAYINMIVKKRCSVRSVTSQD
DmMyoX          SFMEVISTRKVRSEDGALFLDMKVGNLMQQRVIRVQTEQAHEISRLVRQYITMAQISQRDKRELN

mRadixin        EVQQMKAQARVDSSGAA
hsMyoX          GSSR     2058
XeMyoX          SQSSNWAR
```

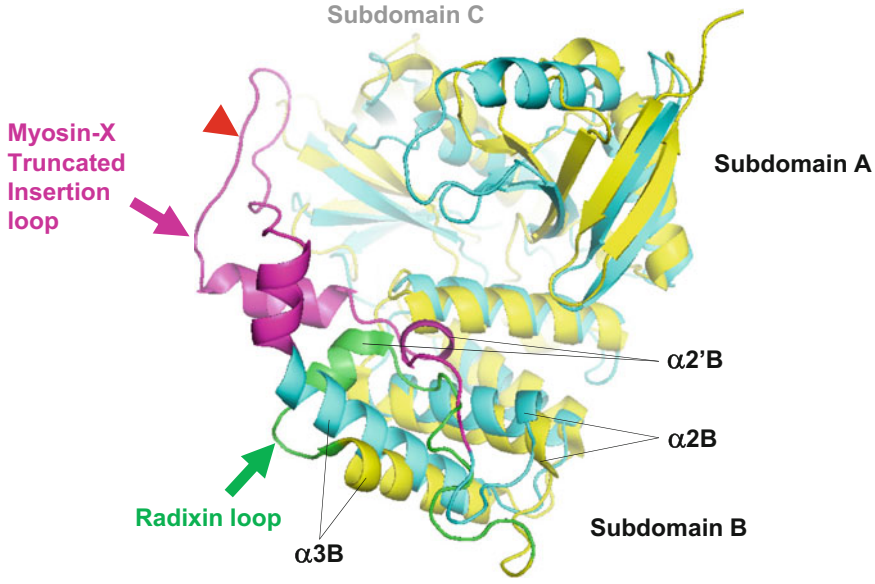

Fig. 1 Detection of non-conserved insertion of the FERM domain in the MyTH4–FERM cassette of nonconventional myosin-X in comparison with ERM proteins. *Top*: Sequence alignment of nonconventional myosin-X

transmembrane sensor of endoplasmic reticulum (ER) stress with dual protein kinase and ribonuclease activities of the cytoplasmic domain. Crystals derived from the Ire1 cytoplasmic domain were not suitable for structure determination. To improve crystallizability, the cytoplasmic domain was engineered to produce a variant containing a 24-residue (C869–F892) internal deletion within the kinase domain that removes a protease labile loop. This loop is part of 30-residue (863–898) insertion between αE and αEF and located at the C-terminal flanking region of the activation loop. It is well known that protein kinase domains frequently possess insertion at the activation loop and flanking regions, which contributes to the uniqueness of each kinase. Although the activation loop is a critical segment for kinase activity, the 24-residue deletion of Ire1 did not affect the enzymatic properties of the protein in vitro. The variant formed crystals that facilitated structural determination at 2.4 Å resolution.

In conclusion, deletion of non-conserved insertions represents one promising approach to improve protein stability and crystallizability. Careful sequence alignment of target proteins is essential for this approach. Long stretches of non-conserved insertions are primary candidates for deletion. In the case of the aforementioned examples, target proteins contained non-conserved insertions comprising more than 30 residues. The position and length of the peptide stretch to be deleted should be optimized by trial-and-error experiments followed by appropriate activity assays. The consideration of known crystal structures of homologues to the target proteins would greatly assist the design of the deletion.

2.3 Fusion and Chimera Proteins

Protein tags are routinely used in recombinant protein expression in order to facilitate the purification of target proteins [10, 11]. Other than short oligopeptides such as a hexahistidine, the use of highly soluble stable proteins, such as GST (glutathione S-transferase), MBP (maltose-binding protein), or thioredoxin, in the preparation and expression of fusion proteins can improve crystallizability and/or diffraction quality by modifying crystal

Fig. 1 (continued) (MyoX) from different sources and ERM proteins, radixin and moesin. Conserved or semi-conserved nonpolar residues are in red or orange. Non-conserved insertions are marked with *blue boxes*. The cleavage site of myosin-X during purification is in the *blue box* and highlighted with a *red circle*. The sources are mouse (m), *Spodoptera frugiperda* (sf), *Homo sapiens* (hs), *Xenopus laevis* (xl), and *Drosophila melanogaster* (dm). *Bottom*: Structural comparison between the obtained myosin-X FERM domain (cyan) with internal truncation (Δ20, see text) and the radixin FERM domain bound to the ICAM-2 peptide (*yellow*) (PDB accession code: 1J19). The structure of the radixin FERM domain represents the canonical FERM domain structures. The FERM domain contains subdomains *A*, *B*, and *C*. The non-conserved insertion found in the myosin-X FERM domain is inserted between α2B and α 3B helices of the subdomain *B*. The truncated insertion of myosin-X (*magenta*) displays a distinct conformation from that of radixin (*green*). Protease labile site is indicated with a *red arrow*

contacts. This strategy was originally applied to the DNA-binding domain of DNA replication-related element-binding factor (DREF), which was crystallized as a fusion protein with *Escherichia coli* GST [12]. The key point of the utility of this technique lies in the design of the linker between the tag and the target protein, which is a determinant factor affecting crystallizability. Long linkers containing a protease cutting site and residues from multicloning sites found in commercial MBP-fusion expression kits should be converted to a shorter linker to limit conformational flexibility. Since the C-terminal end of MBP contains an α-helix, an oligo-alanine stretch was repeatedly employed for this linker [13]. This approach has successfully been applied to a variety of target proteins [14–17]. The alanine stretch of the linker is expected to form an α-helix that reduces the flexibility between the tag and target proteins, and in some cases the alanine linker exists as a loop and produces no direct contact between the tag and target proteins [17]. Generally, linkers comprising three or five alanines have been frequently tested for optimization of crystallizability of the fusion proteins. Some mutations to reduce surface entropy have also been applied to MBP in the fusion protein approach.

Another application concerning the use of fusion proteins relates to stabilization of protein–ligand or protein–protein complexes by increasing the local concentration in an effort to overcome the relatively weak affinity of ligand binding to form a complex. In this case, the ligand protein (or peptide) and its binding protein (or receptor) are fused by a linker peptide. Unlike the linker employed for fusion proteins of tag and target proteins described above, the linker in this case should be sufficiently flexible to facilitate ligand approach and direct binding to the binding site. The choice of linker length is dependent on the distance between the N- and C-terminal ends of the ligand and the binding protein. If the structure of the binding protein is known, extensive modeling could provide sufficient information for design of the linker length and connection to the N- or C-terminal end of the binding protein. If the structure of the binding protein is unknown, fusion proteins with the ligand linked to the N- or C-terminal end of the binding protein should be produced to determine which is most suitable for complex formation. Since the linker is designed to possess flexibility, small residues are employed such as glycine or a mixture of alanine and serine. For example, the structure of the complex between α-catenin and β-catenin was successfully determined using a fusion protein comprising the α-catenin-binding segment of β-catenin (residues 118–151) linked to the N-terminus of the D1 domain of α-catenin via a linker comprising five glycine residues [18]. In this fusion protein, the N-terminal 55 residues of the α-catenin D1 domain were removed since the N-terminal residues inhibit β-catenin binding to the D1 domain. Another example is a fusion protein between the myosin-X MyTH4–FERM cassette

and the DCC peptide, as described above [7]. To improve the quality of the complex crystals, fusion proteins for crystallization were tested, and fusion of the DCC peptide to the C- but not N-terminal end of the myosin-X MyTH4–FERM cassette yielded high-quality crystals of the complex. This C-terminal fusion protein contained two linker residues (Ser and His) between the MyTH4–-FERM cassette and the DCC peptide as a result of the cloning process. Fortunately, the C-terminal very end of the cassette and the N-terminal very end of the DCC peptide were sufficiently flexible to form the complex. However, compared with the non-fused 1:1 complex [6], the conformation of the DCC peptide and its binding mode to the cassette was altered somewhat, probably due to the fusion. Thus, the application of fusion proteins to ligand–protein complexes should be accompanied with additional experimental tests to verify the binding mode and ligand conformation.

References

1. Kendrew JC, Parrish RG, Marrack JR, Orlans ES (1954) The species specificity of myoglobin. Nature (London) 174:946–949
2. Campbell JW, Due'e E, Hodgson G, Mercer WD, Stammers DK, Wendell PL, Muirhead H, Watson HC (1972) X-ray diffraction studies on enzymes in the glycolytic pathway. Cold Spring Harb Symp Quant Biol 36:165–170
3. Terawaki S, Kitano K, Hakoshima T (2008) Crystallographic characterization of the membrane-targeting domain of Rac-specific guanine nucleotide exchange factors Tiam1 and 2. Acta Crystallogr F64:1039–1042
4. Terawaki S, Kitano K, Mori T, Zhai Y, Higuchi Y, Itoh N, Watanabe T, Kaibuchi K, Hakoshima T (2010) The PHCCEx domain of Tiam1/2 is a novel protein- and membrane-targeting module. EMBO J 29:236–250
5. Kagiyama M, Hirano Y, Mori T, Kim S-Y, Kyozuka J, Seto Y, Yamaguchi S, Hakoshima T (2013) Structures of D14 and D14L in the strigolactone and karrikin signaling pathways. Genes Cells 18:147–160
6. Hirano Y, Hatano D, Takahashi A, Toriyama M, Inagaki N, Hakoshima T (2011) Structural basis of cargo recognition by the myosin-X MyTH4-FERM domain. EMBO J 30:2734–2747
7. Wei Z, Yan J, Lu Q, Pan L, Zhang M (2011) Cargo recognition mechanism of myosin X revealed by the structure of its tail MyTH4-FERM tandem in complex with the DCC P3 domain. Proc Natl Acad Sci U S A 108:3572–3577
8. Wu L, Pan L, Wei Z, Zhang M (2011) Structure of MyTH4-FERM domains in myosin VIIa tail bound to cargo. Science 331:757–760
9. Lee KP, Dey M, Neculai D, Cao C, Dever TE, Sicheri F (2008) Structure of the dual enzyme Ire1 reveals the basis for catalysis and regulation in nonconventional RNA splicing. Cell 132:89–100
10. Uhlen M, Forsberg G, Moks T, Hartmanis M, Nilsson B (1992) Fusion proteins in biotechnology. Curr Opin Biotechnol 3:363–369
11. Malhotra A (2009) Tagging for protein expression. Methods Enzymol 463:239–258
12. Kuge S, Fujii Y, Shimizu T, Hirose F, Matsukage A, Hakoshima T (1997) Use of a fusion protein to obtain crystals suitable for X-ray analysis: crystallization of a GST-fused protein containing the DNA-binding domain of DNA replication-related element-binding factor, DREF. Protein Sci 6:1783–1786
13. Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B (2003) Crystal structures of fusion proteins with large-affinity tags. Protein Sci 12:1313–1322
14. Kobe B, Center RJ, Kemp BE, Poumbourios P (1999) Crystal structure of human T cell leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins. Proc Natl Acad Sci 96:4319–4324
15. Ke A, Wolberger C (2003) Insights into binding cooperativity of MATa1/MATalpha2 from the crystal structure of a MATa1

homeodomain-maltose binding protein chimera. Protein Sci 12:306–132

16. Monné M, Han L, Schwend T, Burendahl S, Jovine L (2008) Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. Nature 456:653–657
17. Ullah H, Scappini EL, Moon AF, Williams LV, Armstrong DL, Pedersen LC (2008) Structure of a signal transduction regulator, RACK1, from Arabidopsis thaliana. Protein Sci 17:1771–1780
18. Pokutta S, Weis WI (2000) Structure of the dimerization and beta-catenin-binding region of alpha-catenin. Mol Cell 5:533–543