Thioredoxin as a fusion tag for carrier-driven crystallization

LORENZO CORSINI,¹ MICHAEL HOTHORN,^{1,4} KLAUS SCHEFFZEK,¹
MICHAEL SATTTLER,¹⁻³ AND GUNTER STIER^{1,5}

¹Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

²Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

³Munich Center for Integrated Protein Science and Chair Biomolecular NMR, Department Chemie, Technische Universität München, 85747 Garching, Germany

(RECEIVED July 16, 2008; FINAL REVISION September 3, 2008; ACCEPTED September 3, 2008)

Abstract

Structural investigations are frequently hindered by difficulties in obtaining diffracting crystals of the target protein. Here, we report the crystallization and structure solution of the U2AF homology motif (UHM) domain of splicing factor Puf60 fused to *Escherichia coli* thioredoxin A. Both modules make extensive crystallographic contacts, contributing to a well-defined crystal lattice with clear electron density for both the thioredoxin and the Puf60-UHM module. We compare two short linker sequences between the two fusion domains, GSAM and GSPPM, for which only the GSAM-linked fusion protein yielded diffracting crystals. While specific interdomain contacts are not observed for both fusion proteins, NMR relaxation data in solution indicate reduced interdomain mobility between the Trx and Puf60-UHM modules. The GSPPM-linked fusion protein is significantly more flexible, albeit both linker sequences have the same number of degrees of torsional freedom. Our analysis provides a rationale for the crystallization of the GSAM-linked fusion protein and indicates that in this case, a four-residue linker between thioredoxin A and the fused target may represent the maximal length for crystallization purposes. Our data provide an experimental basis for the rational design of linker sequences in carrier-driven crystallization and identify thioredoxin A as a powerful fusion partner that can aid crystallization of difficult targets.

Keywords: carrier-driven crystallization; fusion tag; Puf60; thioredoxin; linker sequence; NMR relaxation

Supplemental material: see www.proteinscience.org

Several strategies have been developed to facilitate the crystallization of proteins, most of which target the protein sequence itself. In a recent study, the β_2 -adrenergic

G-protein coupled receptor (β_2 A-GPCR) has been crystallized by replacing an intrinsically nonstructured loop with a T4-lysozyme insertion (Cherezov et al. 2007; Rosenbaum et al. 2007). Whereas it was not possible to crystallize the wild-type GPCR (as is the case for most membrane proteins), the lysozyme insertion greatly increased the polar surface available for lattice formation and allowed the growth of three-dimensional (3D) crystals.

Carrier-driven crystallization has been introduced previously (Donahue et al. 1994; Lim et al. 1994). Short protein segments have been N- or C-terminally fused to lysozyme or glutathione-S-transferase (GST), and the fusion proteins have been crystallized under similar conditions

Present addresses: ⁴Plant Biology Laboratory, The Salk Institute, La Jolla, CA 92037, USA; ⁵Umeå Center for Molecular Pathogenesis, Umeå University, S-901 87, Sweden.

Reprint requests to: Michael Sattler, Technische Universität München, Lichtenbergstrasse 4, Garching 85747, Germany, e-mail: sattler@helmholtz-muenchen.de; fax: 49-89-28913869; or Gunter Stier, Umeå Center for Molecular Pathogenesis, Umeå University, Building 6L, S-901 87, Sweden; e-mail: gunter.stier@ucmp.umu.se (after Dec.1, Gunter.Stier@mpimf-Heidelberg.mpg.de).

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.037564.108>.

as the free lysozyme or GST (Zhan et al. 2001). Carrier-driven crystallization has also been applied to larger protein domains (for review, see Smyth et al. 2003). A 99-residue DNA-binding domain and a 318-residue hormone-binding domain have been crystallized as GST fusions, but no structures have been reported to date (Kuge et al. 1997; Lally et al. 1998). The 88-residue ectodomain of the human T-cell leukemia virus type 1 protein gp21 was crystallized as a fusion with maltose-binding protein (MBP), and the structure was solved at 2.5 Å resolution (Kobe et al. 1999). Other examples of structures of MBP fusion constructs are the *Staphylococcus* accessory regulator R (SarR, 115 residues) (Liu et al. 2001), and the MATa1 protein from *Saccharomyces cerevisiae* (50 residues) (Ke and Wolberger 2003). Also, the solution of the structure of α -actinin and of the GTPase domain of dynamin involved fusion to the myosin head domain (Kliche et al. 2001; Niemann et al. 2001). A related strategy of fusion protein crystallization, i.e., crystallization driven by the polymerization of the fused sterile alpha motif (SAM) from the protein translocation Ets leukemia, also yielded diffracting crystals in some cases (Nauli et al. 2007).

Fusion tags are routinely used to overexpress proteins in bacteria. In addition to allowing the crystallization of difficult targets, fusion tag carriers can also facilitate the solution of the phase problem by molecular replacement, as usually a high-resolution structure of the fused protein is available (Kobe et al. 1999; Kliche et al. 2001; Liu et al. 2001; Niemann et al. 2001; Ke and Wolberger 2003). In spite of these potential advantages for structure determination of proteins that are considered to be “higher hanging fruit,” carrier-driven crystallization has been applied only in a handful of cases to date. A major obstacle for the crystallization of larger proteins fused to lysozyme, GST, or MBP seems to be the choice of a suitable linker sequence between the domains. For each of the MBP chimera (Kobe et al. 1999; Liu et al. 2001; Ke and Wolberger 2003), the protein of interest was fused to an extensively mutated C terminus of the MBP domain. In the case of the lysozyme insertion into β_2 A-GPCR, the lengths of the junctions on the N- and C-terminal sides of lysozyme were carefully optimized. Residues in the C terminus of lysozyme were removed, and the positions of the truncations of the loop in β_2 A-GPCR were adjusted to minimize flexibility while not disturbing the conformational arrangement of the flanking GPCR helices (Rosenbaum et al. 2007).

Puf60 is a transcription factor (Liu et al. 2000) and a splicing factor, homologous and functionally related to U2AF⁶⁵ (Page-McCaw et al. 1999; Van Buskirk and Schupbach 2002; Hastings et al. 2007). Like U2AF⁶⁵, it comprises a long, intrinsically unstructured N-terminal domain, two central RNA recognition motifs (RRM), and a C-terminal U2AF homology motif (UHM). The UHM

mediates stable homodimerization of Puf60 in SDS-PAGE (Page-McCaw et al. 1999; Poleev et al. 2000; Rual et al. 2005; L. Corsini, M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T.G. Gibson, and M. Sattler, in prep.). To gain insight into molecular details of this unusual dimerization in reducing SDS-PAGE conditions, we wished to characterize the three-dimensional structure of the UHM domain. However, we could not obtain crystals of the UHM domain alone, and thus attempted carrier-driven crystallization with *Escherichia coli* thioredoxin 1 (Trx). We fused the N terminus of the Puf60-UHM domain to the solvent accessible C terminus of Trx via two linker sequences, GSAM and GSPPM, respectively. Only the GSAM-linked fusion protein crystallized and allowed structure determination by molecular replacement. Biochemical and functional characterization of the Puf60 structure, including a detailed analysis of the SDS-induced dimerization, are described elsewhere (L. Corsini, M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T.G. Gibson, and M. Sattler, in prep.). The methodology of construct and linker design, crystallization, and structure determination reported here suggest the use of Trx in a screen of fusion proteins for the crystallization of “difficult” proteins.

Results

Expression and crystallization of a thioredoxin-1-Puf60 fusion

The Puf60-UHM domain is well structured, as indicated by its ¹H, ¹⁵N NMR correlation spectrum (L. Corsini, M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T.G. Gibson, and M. Sattler, in prep.), and is highly soluble (>150 mg/mL in Tris-HCl or PBS buffers; data not shown). As our attempts to crystallize the UHM domain failed, although ~1300 conditions in sparse matrix screens were tested at three different protein concentrations (20 mg/mL, 50 mg/mL, 100 mg/mL; see Materials and Methods), we attempted to crystallize it as a fusion protein with an N-terminally linked thioredoxin A domain. We used a thioredoxin fusion domain, because (1) it is easy to crystallize, (2) it expresses to high levels in bacteria, and (3) its C terminus is solvent accessible, so that the fused protein does not interfere with the folding. For crystallization purposes, the linker sequence between the two fused domains has to be short enough to restrict their relative conformational flexibility. However, at the same time, the linker must be long enough to allow for an independent folding of both domains. Based on these considerations, we decided to fuse the N terminus of Puf60(UHM) to the C-terminal helix of Trx via the two alternative linker sequences GSAM and GSPPM (Fig. 1A). The GSPPM linker has one additional

Figure 1. (A) Nucleotide and amino acid sequence of the Trx and Puf60(UHM) domains and of the two alternative linkers GSAM and GSPPM. The secondary structure of the fusion protein (as seen in the crystal structure) is indicated. The nucleotide numbering refers to the GSAM construct. (B) Models of the GSAM and the GSPPM linkers in stick representation. The rigid peptide planes are shown as gray squares; rotatable backbone dihedral angles are indicated by arrows and the ϕ and ψ labels.

residue compared to the GSAM sequence, but the two proline residues have restricted ϕ dihedral angles. Thus, even though the two linker sequences differ in their length, both contain eight torsional degrees of freedom in their backbone ϕ and ψ dihedral angles (Fig. 1B).

We could express both the GSAM- and the GSPPM-linked Trx-Puf60(UHM) constructs to high levels in *E. coli*. Both proteins are soluble and elute as a single peak and at the same elution volume from a gel filtration column, corresponding to monomeric proteins (~25 kDa; data not shown). Whereas Puf60(UHM) alone did not precipitate from solutions concentrated to >150 mg/mL, both Trx fusions are soluble only up to ~100 mg/mL. The GSPPM-linked fusion did not crystallize in any of the ~1500 conditions tested at three different protein concentrations each (20 mg/mL, 50 mg/mL, 100 mg/mL; see Materials and Methods). In contrast, the GSAM fusion yielded 3D crystals in six different conditions (see Materials and Methods). Optimization of one of these conditions yielded large, diffracting single crystals.

Structure of the Trx-GSAM-Puf60(UHM) fusion

Crystals of Trx-GSAM-Puf60 have the symmetry of space group $P2_12_12_1$ ($a = 75$, $b = 89$, $c = 299$ Å), and a complete data set was collected to 2.2 Å resolution. The

structural model of Trx (PDB code 2TRX) (Holmgren et al. 1975; Katti et al. 1990), could be used to solve the phase problem by molecular replacement as implemented in PHASER 1.3 (McCoy et al. 2007). Eight Trx monomers were located in the asymmetric unit with high confidence (see Z-scores and log-likelihood gain in the Supplemental material). Most importantly, a refined version of this partial solution ($R_{\text{work}}/R_{\text{free}}$ values of 0.463/0.477) shows clear difference density for the missing Puf60(UHM) domains. Thus, a complete model could have been built even without a priori knowledge of the fold of Puf60(UHM). Nevertheless, we created a homology model (Sali and Blundell 1993) of Puf60-UHM based on SPF45-UHM (PDB code 2PE8 [Corsini et al. 2007], sequence homology/identity 63%/39%). The final molecular replacement solution comprised eight Trx and eight Puf60-UHM domains, and we could refine the model to $R_{\text{work}}/R_{\text{free}}$ 0.21/0.27 (for structural statistics see Supplemental material).

As Puf60-UHM has a weak dimerization propensity, the high protein concentration in the crystallization solution shifts the equilibrium to a population of mainly dimeric Puf60-UHM (L. Corsini, M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T.G. Gibson, and M. Sattler, in prep.). The eight molecules in the asymmetric unit are arranged in four groups of two, each connected by the Puf60-UHM homodimerization interface and additional

contacts between Puf60-UHM and Trx (Fig. 2A). The Trx and Puf60-UHM domains on one protein chain do not interact, but the extensive contacts (1150–1470 Å² buried surface area) are mediated by domains on different protein chains. The four fusion protein homodimers in the asymmetric unit are organized in a doughnut-like shape (Fig. 2B). The eight Trx domains are stacked in two

layers of four in the center of the doughnut, surrounded by a ring of eight PUF60-UHM domains. Remarkably, the four Trx domains in the top layer of the doughnut do not interact with the Trx domains in the bottom layer, thus forming a large ($\sim 50 \times 50 \times 20$ Å) 3D space in the core of the doughnut, filled with solvent only (Fig. 2C).

Trx and Puf60-UHM domains are both involved in lattice formation

To gain some insight into why the Trx-Puf60(UHM) fusion protein crystallized, while it was not possible to crystallize Puf60(UHM) alone, we analyzed the crystallographic packing in detail with the EBI-PISA server (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html) (Krissinel and Henrick 2007).

The packing interactions that connect different asymmetric units involve Trx:Trx contacts, Puf60:Puf60 and Trx:Puf60 interactions (blue, brown, and blue/brown arrowheads in Fig. 2D, respectively). One of these contacts includes a hydrogen-bonding interaction with the Ala-carbonyl in the linker sequence (GSAM; see below). In contrast, the Gly, Ser, and Met residues of the linker sequence are not involved in crystal contacts in any chain. Thus, as the linker sequence is involved in a minor crystal contact, it is possible that, for the non-crystallizing fusion protein with the GSPPM linker, this contact might be less favored.

The lattice of the Trx-Puf60-UHM fusion crystals is formed by both types of domains to a similar extent. There is no continuous, 3D lattice composed only of Trx, as the two layers of Trx domains in the doughnut do not interact with each other (Fig. 2C). The Puf60(UHM) domains also do not form a continuous lattice (Fig. 2D), indicating that the fusion of the two domains was necessary to form 3D crystals in this case. This is in contrast to the crystals of fusion constructs of GST with shorter peptides, where the crystal lattice was formed entirely by the GST domains (Lim et al. 1994; Zhang et al. 1998; Tang et al. 1999; Ware et al. 1999; Zhan et al. 2001).

Structure of the linker peptides

All of the eight linker sequences are well defined in the electron density, but adopt very different conformations (see examples in Fig. 3A). The distance spanned by the linkers, reflected by the distance between the C α atoms of the last Trx residue Ala¹⁰⁹ and the first Puf60-UHM residue, Glu⁴⁶⁰ (numbering as in Fig. 1A), varies between 8.0 Å and 12.7 Å. The conformational variability of the linker sequence can be illustrated by superimposing either the eight Trx domains or the eight Puf60-UHM domains (Fig. 3B, left and right panels, respectively). Whereas the single Trx or Puf60(UHM) domains superimpose with

Figure 2. Structure of the Trx-GSAM-Puf60(UHM) and its packing in the crystal lattice. (A) Structure of a dimer of Trx-GSAM-Puf60(UHM). The Trx domains are shown and labeled in brown; the Puf60(UHM) domains are shown and labeled in blue. The GSAM-linker is shown in orange stick representation and residues are annotated in black. (B) The asymmetric unit of the crystal has the shape of a doughnut. Same color code as in A. Trx domains in the *top* layer in the *center* are labeled in boldface; Trx domains in the *bottom* layer in *italics*. (C) Stereo drawing of one half of the doughnut shown in B. The doughnut is rotated by 90° around the *x*-axis with respect to the orientation shown in B. Note that the Trx domains in the top layer do not interact with the Trx domains in the bottom layer. Same color coding as in A and B. (D) Lattice packing from two views. In the *left* panel, symmetric Trx:Trx crystal contacts are highlighted by brown arrowheads. In the *right* panel, the Puf60(UHM):Puf60(UHM) and Trx:Puf60(UHM) contacts are highlighted by blue and blue/brown arrowheads, respectively. Same color code as in A–C. Symmetry mates were generated with PyMOL (version 0.99, DeLano Scientific LLC).

Figure 3. The linker residues have variable conformations. (A) Structures of three of the eight linkers in the asymmetric unit are shown in stick representation. The structural models are surrounded by omit electron density difference maps ($2F_o - F_c$), contoured at 1.5σ . (B, left) The eight fusion protein chains in the asymmetric unit are superimposed based on the Trx domain. The linker residues and the Puf60(UHM) domains cannot be overlaid. (Right) When the Puf60(UHM) domains of the same chains as shown in the left panel are superimposed, the linker residues and the Trx domains do not superimpose.

coordinate RMSD values of 0.2–0.4 Å over all atoms, the linker sequences cannot be superimposed. When either the Trx or the Puf60-UHM domains are superimposed, the second, nonaligned domains all have a similar rotational orientation, but different translational positions (Fig. 3B). The Gly and Ser residues of the GSAM linker undergo more conformational variability than the Ala and Met residues (compare left and right panels in Fig. 3B). This is probably due to the hydrogen-bonding crystal contact that the Ala residues are involved in (see above).

Furthermore, the Trx and Puf60(UHM) domains within a single chain do not interact in the structure (Fig. 3B). Consistently, in solution, the NMR ^1H - ^{15}N correlation spectra of the free Trx and Puf60(UHM) domains are very similar to the corresponding (sub)spectra of the fusion construct (Fig. 4A). This indicates that there are no strong, specific interdomain interactions between Trx and Puf60(UHM) in our fusion construct.

The GSPPM linker allows higher flexibility than GSAM

The fact that the backbone carbonyl of the linker Ala residue (GSAM) is involved in a crystal contact does not fully explain why the GSPPM-linked fusion did not crystallize. In fact, an analogous interaction is possible with the Pro residue in the equivalent position in the

GSPPM linker. To further characterize the two fusion proteins and the differences of the linkers we compared the conformational flexibility of the two chimeric proteins by NMR relaxation.

We prepared uniformly ^{15}N -labeled samples of both constructs. The ^1H and ^{15}N chemical shift assignments of the Trx domain (Jeng et al. 1994) and of the Puf60-UHM (Corsini et al. 2007; L. Corsini, M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T.G. Gibson, and M. Sattler, in prep.) could be transferred to obtain the ^1H , ^{15}N assignment of the NMR signals in the fusion proteins (81 of 104 Trx and 92 of 98 Puf60-UHM ^1H - ^{15}N correlation signals). A few signals could not be assigned due to signal overlap and/or chemical shift changes in the spectra of the fusion protein compared to the spectra of the isolated domains (Fig. 4A). The GSAM-linked fusion protein is structurally similar to the GSPPM-linked fusion, as the ^1H - ^{15}N HSQC spectra of the two proteins can be superimposed with only minor differences (Fig. 4B).

We analyzed to what extent the linkers allow partially independent tumbling of the Trx and Puf60(UHM) domains by recording ^{15}N NMR T_1 and T_2 relaxation times, which depend mainly on the rotational diffusion of the whole molecule and on internal motions of the amide-proton bond vector. To separate the two effects, we measured the $\{^1\text{H}\}$ - ^{15}N NOE, which is lower than ~ 0.65 for residues undergoing substantial subnanosecond

Figure 4. (A) Overlay of ^1H - ^{15}N HSQC spectra of the Trx-GSAM-Puf60(UHM) fusion protein (black), isolated Trx (red), and Puf60(UHM) (blue). The NMR signals of the free domains are very similar compared to the corresponding signals in the fusion protein, indicating that the two domains do not strongly interact in solution. (B) Overlay of ^1H - ^{15}N HSQC spectra of the Trx-GSAM-Puf60(UHM) and Trx-GSPPM-Puf60(UHM) fusion proteins in black and red, respectively. The peak indicated by the arrow corresponds to the alanine in the GSAM linker, as it is the only peak that is not present in the GSPPM-linked construct. (Note that prolines lack an amide proton and are thus not visible in the ^1H - ^{15}N HSQC spectra.)

internal motions (Kay et al. 1989; Tjandra et al. 1996). As indicated in Figure 5A,B, T_1/T_2 values of 807 ms/52 ms would be expected for a rigid, spherical, single-domain protein corresponding to the size of the Trx-Puf60(UHM) fusion proteins (Fushman et al. 1994; Daragan and Mayo 1997). In contrast, T_1/T_2 values of 470 ms/91 ms would be expected for a rigid, single-domain protein of ~ 100 residues (the size of single Trx or Puf60(UHM) domains). At 300 K and 50 MHz ^{15}N Larmor frequency, the Trx-GSAM-Puf60(UHM) construct has average T_1/T_2 values of 732 ms/57 ms for residues with a $\{^1\text{H}\}$ - ^{15}N NOE > 0.65 , whereas the Trx-GSPPM-Puf60(UHM) construct has average T_1/T_2 values of 685 ms/63 ms for the same residues (Fig. 5A–C). Thus, the GSPPM-linked fusion is significantly more flexible than the GSAM-linked construct. The Ala residue in the GSAM linker has T_1/T_2 /heteronuclear NOE values of 464 ms/189 ms/ -0.12 , indicating high structural flexibility of this linker sequence in solution.

The T_1/T_2 values of both the GSAM- and GSPPM-linked fusions indicate that the rotational diffusion of the domains is strongly reduced compared to what would be expected for isolated, independently tumbling 100-residue domains, connected by a long and completely flexible linker. Nevertheless, both linkers allow a higher mobility than what would be expected for a rigid, spherical, single-domain protein of the size of the fusion protein. In conclusion, both linkers, GSAM and GSPPM, largely restrict the rotational tumbling of the domains relative to

each other, and the domain mobility is more restricted for the GSAM-linked fusion than for the GSPPM-linked fusion protein.

Discussion

We demonstrate the use of thioredoxin-1 as a fusion tag for carrier-driven crystallization and report the first structure of a protein fused to *E. coli* thioredoxin-1 (Trx). We chose Trx mainly for three reasons. First, in our experience, it is particularly easy to express proteins as Trx fusions in *E. coli* to high expression levels. Second, the natural C terminus of Trx is the end of helix αE , which is solvent accessible and also rigidly bound to the domain. Finally, Trx itself readily crystallizes in several crystal forms, and extensive parts of its rather polar surface area can form crystal contacts (compare wild-type structures: PDB code 2TRX [space group C2], 2H6X [$P6_1$]; and mutants: 2FCH [$P2_12_12_1$], 2FD3 [$P2_1$], 1TXX [$P4_32_12_1$]). Nevertheless, thioredoxin-driven crystallization was reported only once, and the crystal structure of the fusion protein was not described (Stoll et al. 1998).

The design of the linker sequence is crucial to the success or failure of the crystallization of the fusion protein construct. We compared the two linker sequences GSAM and GSPPM, of which only the GSAM-linked fusion protein yielded crystals and allowed structure determination. Each of the eight linkers in the asymmetric unit has a different conformation, and the two Trx and Puf60-UHM domains on the same protein chain do not interact with each other in the crystal. The similarity of the ^1H and ^{15}N chemical shifts of the isolated Trx and Puf60-UHM domains, when compared to the corresponding chemical shifts in the fusion protein, indicates that the two domains do not specifically interact in solution as well (Fig. 4A). This may correlate with the fact that both domains have rather acidic pI values (4.67 for Trx, 4.34 for Puf60-UHM) (Gasteiger et al. 2005) and are affected by substantial electrostatic repulsion in neutral solution.

We also measured the relaxation properties of ^{15}N nuclei in the peptide backbone by NMR. These data indicate that both the GSAM and the GSPPM linkers significantly restrict the domain motion of the Trx and Puf60-UHM domains in solution. The effect is more pronounced for the GSAM- than for the GSPPM-linked fusion protein, suggesting that this might be a reason for the failure of the GSPPM-linked protein to crystallize. In fact, the GSAM-linker is only involved in a minor crystal contact and has different conformations in each of the proteins in the asymmetric unit of the crystal (Fig. 3A,B). Thus, it is unlikely that the residues in the linker substantially influence the crystallization. Rather, the higher flexibility of the GSPPM linker and enhanced relative domain mobility might hinder the formation of a

Figure 5. NMR relaxation analysis of the GSAM- and GSPPM-linked fusion constructs. (A) ¹⁵N T₁ and (B) ¹⁵N T₂ relaxation times at 300 K and 50.68 MHz Larmor frequency. Relaxation times of the GSAM- and GSPPM-linked fusion constructs are shown as filled triangles and open circles, respectively. The gray lines indicate relaxation times expected for a rigid 224- or 112-residue protein (corresponding to a compact fusion protein and the free Trx or Puf60(UHM) domains) as indicated at the y-axis. (C) {¹H}-¹⁵N heteronuclear NOE data.

stable building block to induce crystallization (for example corresponding to the conformation of the fusion protein dimer shown in Fig. 2A). This in turn is likely to slow down and even to inhibit the nucleation and the growth of the crystals. It is interesting to note that both the GSAM and the GSPPM linker have eight mobile ϕ and ψ dihedral angles, as the Pro residues have rigid ϕ torsional angles. Thus, the flexibility seems to be caused by the distance that can be spanned by the linker, rather than the torsional flexibility of the amino acid sequence itself. As the GSAM linker is still sufficiently long to avoid interactions of the two fused domains in the crystal (see Fig. 3B), we speculate that shorter linkers might be

even better suited to enhance crystallization. In the case of the successfully crystallized MBP fusions, three to five residues were used as linker sequences (Smyth et al. 2003), indicating that this range of linker lengths might be suitable in other cases as well.

The Ala-Met or Pro-Met residues in the linker sequence correspond to a cloning site (CCATG/G, NcoI). If additional fusion domains should be screened, it will be convenient to use standardized cloning sites, so that the same insert can be cloned into the different fusion constructs. For this reason both linker sequences were designed to end with a Met residue (the ATG codon in the NcoI restriction site). Based on the Trx-Puf60(UHM)

crystal structure, the linker residues preceding the Met could be replaced by a single Ala, or removed completely, thus shortening the linker sequence to one or two residues. Shortening the linker would further reduce the conformational flexibility of the two domains, thus potentially increasing the probability of crystallization.

The results and analyses presented here, together with previous research on fusion protein crystallization (Kliche et al. 2001; Niemann et al. 2001; Smyth et al. 2003; Cherezov et al. 2007), suggest including the thioredoxin domain and the linker sequence GSAM in a screen of fusion proteins and linkers for carrier-driven crystallization. We propose screening of combinations of fusion tags (thioredoxin, MBP, GST, NusA, lysozyme, and others) with variations of linkers, such as GSAM or shorter ones, for proteins that are difficult to crystallize with standard approaches.

Materials and Methods

Cloning and purification of the fusion constructs

A TrxA cassette with BspHI and BamHI sticky ends was co-ligated with various BglII and Acc65I linker GFP cassettes into a modified version of the commercially available pET-24d vector with NcoI and Acc65I cloning sites. A variation of linker fragments introduced a new NcoI site in front of the GFP gene. These source vectors were tested for GFP expression and purification and then used as cloning vectors to introduce target genes via NcoI and Acc65I C-terminal of His tagged carrier protein. (Fig. 1A). The vectors are available upon request.

Puf60-UHM (residues 460–559) was amplified from cDNA (BIOCAT MHS-1011-62314) with the oligos 5'-GCTCCATGG AGTCTACAGTGATGGTTCTGCGC-3' and 5'-GCTGGTACC TTACGCAGAGAGGTCACGTGTATCAAA-3' and ligated into the vector described above after restriction with NcoI/Acc65I. The Trx-GSAM-Puf60(UHM) and Trx-GSPPM-Puf60(UHM) constructs were expressed for 12 h in *E. coli* BL21(DE3)pLysS at 21°C in LB medium after induction with IPTG at OD₆₀₀ = 0.75. Both proteins were purified with Ni-NTA sepharose (Biorad) followed by size exclusion chromatography in 20 mM Tris-HCl pH 7.0, 150 mM NaCl, 5 mM β-mercaptoethanol.

Crystallization and data collection

For crystallization, Puf60(UHM) or the chimeric Trx-Puf60(UHM) fusion proteins were concentrated to 20 mg/mL, 50 mg/mL, or 100 mg/mL. We mixed 0.1 μL of protein solution with 0.1 μL of commercial crystallization solutions and dispensed it with a Mosquito Robot (TTPlabtech) in 96-well sitting-drop plates (MRC) at room temperature. Fourteen screens were tested: Qiagen Classics, Classics Lite, pH clear, Anions, Cations, AmSO₄, MPD, PEGs, PEGs II, Hampton Research Index, SaltRX, three in-house-made sparse matrix, and more systematic pH vs. anions/cations screens. No crystals developed for Puf60(UHM) or the GSPPM-linked Trx fusion. The GSAM-linked fusion crystallized in six conditions: (1) 1.8 M (NH₄)₂SO₄, 0.1 M MES pH 6.5, 10 mM CoCl₂; (2) 30% PEG 8000, 0.2 M (NH₄)₂SO₄, 0.1 M Na-cacodylate pH 6.5, 10 mM

MgCl₂; (3) 3 M NaCl, 0.1 M citrate pH 3.5; (4) 2 M (NH₄)₂SO₄, 0.1 M Bis-Tris pH 6.5; (5) 25% PEG 3350, 0.1 M Tris pH 8.5, 0.2 M MgCl₂; (6) 25% PEG 3350, 0.2 M (NH₄)₂SO₄, 0.1 M Bis-Tris pH 6.5.

As (NH₄)₂SO₄-containing conditions are prevalent, we optimized crystal growth using (NH₄)₂SO₄ and the Hampton Research Additive screen. In the final condition, Trx-GSAM-Puf60(UHM) was concentrated to about 75 mg/mL in 20 mM Tris (pH 7.0), 150 mM NaCl, 5 mM β-mercaptoethanol. Crystals were grown by vapor diffusion from hanging drops composed of 1 μL protein solution and 1 μL crystallization buffer (1.4 M (NH₄)₂SO₄/50 mM K-formate) suspended over 1 mL of the latter as reservoir solution. Crystals grew to a size of about 100 μm × 100 μm × 500 μm and were cryoprotected by serial transfer into a solution containing 20% (v/v) ethylene glycol, 1.5 M (NH₄)₂SO₄, 50 mM K-formate. Diffraction data were recorded at beamline PX01 of the Swiss Light Source, Villigen, Switzerland. Data processing and scaling were carried out with XDS (Kabsch 1993).

Structure determination and refinement

The structure of the thioredoxin-Puf60 fusion protein was solved by molecular replacement as implemented in PHASER (McCoy et al. 2007). The structure of *E. coli* thioredoxin (PDB code 2TRX) (Katti et al. 1990) and a homology model of Puf60-UHM generated with MODELLER (Sali and Blundell 1993) based on the structure of free SPF45-UHM (PDB code 2PE8) (Corsini et al. 2007) as a template were used as search models. All loops were deleted from the Puf60(UHM) homology model and all residues not identical in Puf60(UHM) and SPF45(UHM) were converted to alanine. A first solution comprised eight Trx and four Puf60(UHM) monomers. These were refined in alternating cycles of model correction in COOT (Emsley and Cowtan 2004) and restrained refinement as implemented in REFMAC (Murshudov et al. 1997) and PHENIX.REFINE (Adams et al. 2002). The models of the refined Trx and Puf60(UHM) domains were used as search models in an additional round of rotational and translational Patterson searches in PHASER, which identified a total of eight Trx monomers and eight Puf60(UHM) monomers (see Z-scores and log-likelihood gain in the Supplemental material). The final model could be refined to $R_{\text{work}}/R_{\text{free}}$ of 0.211/0.271 at 2.2 Å. The final Trx domains have RMSD to the initial search model (PDB code 2TRX) of 0.287 Å to 0.408 Å over 97–103 of 109 C_α atoms. The final Puf60(UHM) domains have RMSD to the initial homology model of 0.492 Å to 0.635 Å over 65–68 of 100 C_α atoms. See Supplemental material for structural statistics. Structural visualization was done with PyMOL (DeLano Scientific LLC). The eight UHM domains in our crystal structure can be superimposed onto the reported solution structure of Puf60-UHM (PDB code 2DNY) with RMSD of 0.9 to 1.1 Å over 90 of 100 C_α atoms. The solution structure, however, does not indicate dimerization of the Puf60-UHM.

NMR

For relaxation experiments, ¹⁵N-labeled GSAM- and GSPPM-linked fusion proteins were concentrated to 13.9 mg/mL and 14.3 mg/mL, respectively (OD₂₈₀). All NMR spectra were recorded at 300 K on a Bruker DRX500 spectrometer, processed with NMRPipe (Delaglio et al. 1995) and analyzed with NMRView (Johnson and Blevins 1994). Backbone ¹H and ¹⁵N

resonances of the Trx and the Puf60(UHM) domains were transferred from the assigned spectra of free, reduced Trx (Jeng et al. 1994) and Puf60(UHM) (L. Corsini, M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T.G. Gibson, and M. Sattler, in prep.). We assigned the Ala in the linker sequence (GSAM) by comparing the HSQC spectra of the GSAM and GSPPM-linked fusion proteins (labeled in Fig. 4B). ^{15}N relaxation data were recorded as described (Farrow et al. 1994).

Protein Data Bank accession code

Structure coordinates and reflection data have been deposited in the PDB under accession number 3DXB.

Acknowledgments

We thank the staff at beamline PX01 of the Swiss Light Source, Villigen, Switzerland, for technical assistance during data collection, Christoph Müller and Stephen Cusack for sharing beam time, and Jane H. Dyson for providing us with the NMR assignment of reduced and oxidized thioredoxin. This work was supported by the Deutsche Forschungsgemeinschaft (Sa 823/5) and EU grant 3D Repertoire (LSHG-CT-2005-512028). M.H. acknowledges support by the Peter and Traudl Engelhorn Stiftung, Penzberg, Germany.

References

- Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. 2002. PHENIX: Building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **58**: 1948–1954.
- Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.G., Thian, F.S., Kobilka, T.S., Choi, H.J., Kuhn, P., Weis, W.I., Kobilka, B.K., et al. 2007. High-resolution crystal structure of an engineered human β_2 -adrenergic G protein-coupled receptor. *Science* **318**: 1258–1265.
- Corsini, L., Bonnal, S., Basquin, J., Hothorn, M., Scheffzek, K., Valcarcel, J., and Sattler, M. 2007. U2AF-homology motif interactions are required for alternative splicing regulation by SPF45. *Nat. Struct. Mol. Biol.* **14**: 620–629.
- Daragan, V. and Mayo, K.H. 1997. Motional model analyses of protein and peptide dynamics using ^{13}C and ^{15}N NMR relaxation. *Prog. Nucl. Magn. Reson. Spectrosc.* **31**: 63–105.
- Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. 1995. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biol. NMR* **6**: 277–293.
- Donahue, J.P., Patel, H., Anderson, W.F., and Hawiger, J. 1994. Three-dimensional structure of the platelet integrin recognition segment of the fibrinogen γ chain obtained by carrier protein-driven crystallization. *Proc. Natl. Acad. Sci.* **91**: 12178–12182.
- Emsley, P. and Cowtan, K. 2004. Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**: 2126–2132.
- Farrow, N.A., Muhandiram, R., Singer, A.U., Pascal, S.M., Kay, C.M., Gish, G., Shoelson, S.E., Pawson, T., Forman-Kay, J.D., and Kay, L.E. 1994. Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry* **33**: 5984–6003.
- Fushman, D., Weisemann, R., Thüring, H., and Rüterjans, H. 1994. Backbone dynamics of ribonuclease T1 and its complex with 2'GMP studied by two-dimensional heteronuclear NMR spectroscopy. *J. Biol. NMR* **4**: 61–78.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. 2005. Protein identification and analysis tools on the ExPASy server. In *The proteomics protocols handbook*. (ed. J.M. Walker), pp. 571–607. Humana Press, Totowa, NJ.
- Hastings, M.L., Allemand, E., Duelli, D.M., Myers, M.P., and Krainer, A.R. 2007. Control of pre-mRNA splicing by the general splicing factors PUF60 and U2AF65. *PLoS One* **2**: e538. doi: 10.1371/journal.pone.0000538.
- Holmgren, A., Soderberg, B.O., Eklund, H., and Branden, C.I. 1975. Three-dimensional structure of *Escherichia coli* thioredoxin-S2 to 2.8 Å resolution. *Proc. Natl. Acad. Sci.* **72**: 2305–2309.
- Jeng, M.F., Campbell, A.P., Begley, T., Holmgren, A., Case, D.A., Wright, P.E., and Dyson, H.J. 1994. High-resolution solution structures of oxidized and reduced *Escherichia coli* thioredoxin. *Structure* **2**: 853–868.
- Johnson, B.A. and Blevins, R.A. 1994. NMRView: A computer program for the visualization and analysis of NMR data. *J. Biol. NMR* **4**: 603–614.
- Kabsch, W. 1993. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.* **26**: 795–800.
- Katti, S.K., LeMaster, D.M., and Eklund, H. 1990. Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* **212**: 167–184.
- Kay, L.E., Torchia, D.A., and Bax, A. 1989. Backbone dynamics of proteins as studied by ^{15}N inverse detected heteronuclear NMR spectroscopy: Application to staphylococcal nuclease. *Biochemistry* **28**: 8972–8979.
- Ke, A. and Wolberger, C. 2003. Insights into binding cooperativity of MATA1/MAT α 2 from the crystal structure of a MATA1 homeodomain-maltose binding protein chimera. *Protein Sci.* **12**: 306–312.
- Kliche, W., Fujita-Becker, S., Kollmar, M., Manstein, D.J., and Kull, F.J. 2001. Structure of a genetically engineered molecular motor. *EMBO J.* **20**: 40–46.
- Kobe, B., Center, R.J., Kemp, B.E., and Pombourios, 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.
- Krissinel, E. and Henrick, K. 2007. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**: 774–797.
- Kuge, M., Fujii, Y., Shimizu, T., Hirose, F., Matsukage, A., and 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.
- Lally, J.M., Newman, R.H., Knowles, P.P., Islam, S., Coffey, A.I., Parker, M., and Freemont, P.S. 1998. Crystallization of an intact GST-estrogen receptor hormone binding domain fusion protein. *Acta Crystallogr. D Biol. Crystallogr.* **54**: 423–426.
- Lim, K., Ho, J.X., Keeling, K., Gilliland, G.L., Ji, X., Ruker, F., and Carter, D.C. 1994. Three-dimensional structure of *Schistosoma japonicum* glutathione S-transferase fused with a six-amino acid conserved neutralizing epitope of gp41 from HIV. *Protein Sci.* **3**: 2233–2244.
- Liu, J., He, L., Collins, I., Ge, H., Libutti, D., Li, J., Egly, J.M., and Levens, D. 2000. The FBP interacting repressor targets TFIID to inhibit activated transcription. *Mol. Cell* **5**: 331–341.
- Liu, Y., Manna, A., Li, R., Martin, W.E., Murphy, R.C., Cheung, A.L., and Zhang, G. 2001. Crystal structure of the SarR protein from *Staphylococcus aureus*. *Proc. Natl. Acad. Sci.* **98**: 6877–6882.
- McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. 2007. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**: 658–674.
- Murshudov, G.N., Vagin, A.A., and Dodson, E.J. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**: 240–255.
- Nauli, S., Farr, S., Lee, Y.J., Kim, H.Y., Faham, S., and Bowie, J.U. 2007. Polymer-driven crystallization. *Protein Sci.* **16**: 2542–2551.
- Niemann, H.H., Knetsch, M.L., Scherer, A., Manstein, D.J., and Kull, F.J. 2001. Crystal structure of a dynamin GTPase domain in both nucleotide-free and GDP-bound forms. *EMBO J.* **20**: 5813–5821.
- Page-McCaw, P.S., Amonlirdviman, K., and Sharp, P.A. 1999. PUF60: A novel U2AF65-related splicing activity. *RNA* **5**: 1548–1560.
- Polev, A., Hartmann, A., and Stamm, S. 2000. A trans-acting factor, isolated by the three-hybrid system, that influences alternative splicing of the amyloid precursor protein minigene. *Eur. J. Biochem.* **267**: 4002–4010.
- Rosenbaum, D.M., Cherezov, V., Hanson, M.A., Rasmussen, S.G., Thian, F.S., Kobilka, T.S., Choi, H.J., Yao, X.J., Weis, W.I., Stevens, R.C., et al. 2007. GPCR engineering yields high-resolution structural insights into β_2 -adrenergic receptor function. *Science* **318**: 1266–1273.
- Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. 2005. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**: 1173–1176.

- Sali, A. and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**: 779–815.
- Smyth, D.R., Mrozkiewicz, M.K., McGrath, W.J., Listwan, P., and Kobe, B. 2003. Crystal structures of fusion proteins with large-affinity tags. *Protein Sci.* **12**: 1313–1322.
- Stoll, V.S., Manohar, A.V., Gillon, W., MacFarlane, E.L., Hynes, R.C., and Pai, E.F. 1998. A thioredoxin fusion protein of VanH, a D-lactate dehydrogenase from *Enterococcus faecium*: Cloning, expression, purification, kinetic analysis, and crystallization. *Protein Sci.* **7**: 1147–1155.
- Tang, L., Guo, B., Javed, A., Choi, J.Y., Hiebert, S., Lian, J.B., van Wijnen, A.J., Stein, J.L., Stein, G.S., and Zhou, G.W. 1999. Crystal structure of the nuclear matrix targeting signal of the transcription factor acute myelogenous leukemia-1/polyoma enhancer-binding protein 2 α /core binding factor α 2. *J. Biol. Chem.* **274**: 33580–33586.
- Tjandra, N., Grzesiek, S., and Bax, A. 1996. Magnetic field dependence of nitrogen-proton *J* splittings in ^{15}N -enriched human ubiquitin resulting from relaxation interference and residual dipolar coupling. *J. Am. Chem. Soc.* **118**: 6264–6272.
- Van Buskirk, C. and Schupbach, T. 2002. Half pint regulates alternative splice site selection in *Drosophila*. *Dev. Cell* **2**: 343–353.
- Ware, S., Donahue, J.P., Hawiger, J., and Anderson, W.F. 1999. Structure of the fibrinogen γ -chain integrin binding and factor XIIIa cross-linking sites obtained through carrier protein driven crystallization. *Protein Sci.* **8**: 2663–2671.
- Zhan, Y., Song, X., and Zhou, G.W. 2001. Structural analysis of regulatory protein domains using GST-fusion proteins. *Gene* **281**: 1–9.
- Zhang, Z., Devarajan, P., Dorfman, A.L., and Morrow, J.S. 1998. Structure of the ankyrin-binding domain of α -Na,K-ATPase. *J. Biol. Chem.* **273**: 18681–18684.