J Biomol NMR (2010) 46:23–31
DOI 10.1007/s10858-009-9371-6

PERSPECTIVE

Overcoming the solubility limit with solubility-enhancement tags: successful applications in biomolecular NMR studies

Pei Zhou · Gerhard Wagner

Received: 11 June 2009 / Accepted: 18 August 2009 / Published online: 3 September 2009

Abstract Although the rapid progress of NMR technology has significantly expanded the range of NMR-trackable systems, preparation of NMR-suitable samples that are highly soluble and stable remains a bottleneck for studies of many biological systems. The application of solubility-enhancement tags (SETs) has been highly effective in overcoming solubility and sample stability issues and has enabled structural studies of important biological systems previously deemed unapproachable by solution NMR techniques. In this review, we provide a brief survey of the development and successful applications of the SET strategy in biomolecular NMR. We also comment on the criteria for choosing optimal SETs, such as for differently charged target proteins, and recent new developments on NMR-invisible SETs.

Keywords NMR · Solubility enhancement tag · Protein aggregation · Protein GB1 · Protein · Protein stability enhancement

Introduction

The advancement of NMR instrumentation and methodology has made solution NMR spectroscopy an increasingly powerful tool for investigations of protein structure and dynamics under physiological conditions, and for studies of ligand binding and reaction mechanisms in solution. However, the inherent sensitivity limitation of NMR requires protein samples to be stable at high concentrations (>100 μM for structural studies) for an extended period (typically over a couple of days). Unfortunately, an estimated 75% of soluble proteins and many biologically important macromolecules are characterized by low solubility and instability (Christendat et al. 2000). Therefore, preparation of well-behaved, non-aggregated samples at sufficiently high protein concentrations remains a serious challenge for structural and dynamic studies by NMR.

Numerous efforts have been devoted to overcoming the solubility and sample stability issues. For example, extensive buffer screening (Bagby et al. 1997; Lepre and Moore 1998), addition of charged amino acids (Golovanov et al. 2004), or introduction of point mutants (Huang et al. 1996; Ito and Wagner 2004; Sun et al. 1999) have been successfully utilized to increase the solubility of the target proteins. However, these methods are often protein specific, largely based on trial and error, and may not be easily applicable to other systems. To overcome these issues and develop a generic approach, we introduced the concept of non-cleavable solubility-enhancement tags (SETs) for studies of poorly behaved proteins by solution NMR (Zhou et al. 2001b). Since then, this strategy has found wide applications in the NMR community, and has been used to improve the solubility and sample stability of ~30 proteins. For many of these examples, the SET approach has enabled successful determination of high-resolution solution structures. Here, we give a brief overview of the initial development, the theory and the successful application of the SET strategy in biomolecular NMR studies, and we comment on recent improvements of the SET strategy. We refer readers to the excellent review by Waugh (2005) for applications of protein tags in a non-NMR setting.

P. Zhou
Department of Biochemistry, Duke University Medical Center, 242 Nanaline Duke Building, Research Drive, Durham, NC 27710, USA

G. Wagner (✉)
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Building C1, Room 112, 240 Longwood Avenue, Boston, MA 02115, USA
e-mail: wagner@hms.harvard.edu

Development and application of SET

Protein tags such as GST and MBP have been widely used as affinity tags for purifying recombinant proteins (di Guan et al. 1988; Smith and Johnson 1988). It was frequently observed that these fusion proteins overexpress better and exhibit enhanced solubility and sample stability compared to their untagged counterparts. This observation has prompted the search of new fusion tags to improve the soluble expression of target proteins in *E. coli* (Davis et al. 1999; DelProposto et al. 2009; Forrer and Jaussi 1998; Huth et al. 1997; LaVallie et al. 2000; Pilon et al. 1996; Samuelsson et al. 1994; Zuo et al. 2005, 2008; reviewed by Waugh 2005). Due to the size limit of routine NMR techniques ($\sim$30 kDa), it is preferable to remove the protein tag before subsequent NMR studies. Unfortunately, once the fusion tag is cleaved by proteolytic digestion, the target protein often becomes unstable again and precipitates within hours, thereby prohibiting further NMR studies.

Because it is only the size limit that restricts the use of protein tags in solution NMR studies, we reasoned that a highly soluble and stable protein that is also sufficiently small can be used as a non-cleavable tag for NMR studies. Several small protein tags, such as protein G B1 domain (GB1, 56 residues; Huth et al. 1997), protein D (110 residues; Forrer and Jaussi 1998), the Z domain of Staphylococcal protein A (58 residues; Samuelsson et al. 1994) and thioredoxin (109 residues; LaVallie et al. 2000), have been shown to increase the yield of soluble proteins. We chose the smallest tag, GB1 as the solubility-enhancement tag for further evaluation. In our study of the DFF40/45 N-terminal CIDE domain complex, attachment of the non-cleavable GB1 tag to DFF45 not only increased the solubility of the DFF40/45 complex from 0.2 to 0.6 mM, but also increased the sample stability from 5 days to over a month at 23°C (Zhou et al. 2001b). The use of the solubility-enhancement tag has resulted in a dramatic improvement of spectral quality (Fig. 1) and has enabled subsequent structure determination of the DFF40/45 CIDE domain complex by NMR (Zhou et al. 2001a). To our knowledge, this is the first demonstration of using non-cleavable solubility-enhancement tags to overcome sample solubility and stability issues for structural studies by NMR.

Since the initial demonstration and application of the SET strategy to NMR structure determination (Zhou et al. 2001a, b), this fusion tag approach has found wide applications in the NMR community. Approximately 30 examples have now been reported in the literature, which show significant enhancement of protein solubility and/or sample stability using SETs (Table 1). Additionally, in many cases, the creation of SET-fusion proteins also significantly improved protein overexpression levels in *E. coli* and the final yields of the purified proteins. These target proteins cover a wide range of structural topologies and biological functions, which truly demonstrate the generality of the SET approach in biomolecular NMR studies.

Choice of SETs

Although GB1 has been a highly successful solubility-enhancement tag, other highly soluble and stable small protein domains can also serve similar functions. Unfortunately, how the SET enhances the solubility of a target protein remains poorly understood, and comparative proteomic studies have not revealed a universally good tag for all protein targets (Hammarström et al. 2002, 2006). Based on a thermodynamic analysis, we suggest here the following criteria for choosing a solubility-enhancement tag.

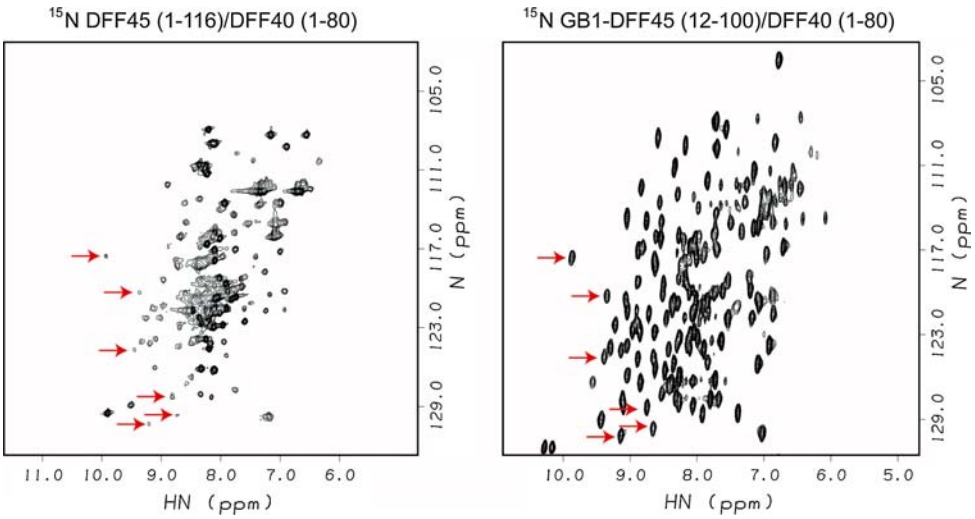

Fig. 1 HSQC spectra of ^{15}N-labeled DFF45 N-terminal CIDE domain in complex with unlabeled DFF40 (1–80). Attachment of the GB1 tag significantly increased the solubility and stability of the DFF40/45 complex and generated superior NMR spectra. *Arrows* indicate distinct resonances from DFF45 in the DFF40/45 complex (reprinted with permission from Fig. 1bc of (Zhou et al. 2001b), Journal of Biomolecular NMR)

Table 1 Examples of NMR studies using the SET approach

Tag	Target protein and property	Reported effect(s)	Notes and references
GB1 (6.2 kDa; pI = 4.5)	Mouse prion protein (mPrP)	Increased the expression yield and solubility	GB1 was used to enhance the expression yield and the solubility of selected mPrP constructs (Hornemann et al. 2009)
$GB1^{basic}$ (6.2 kDa; pI = 8.0)	HPV16 E6 constructs (1) E6 (17 kDa; pI = 9.0) (2) E6 N (8.9 kDa, pI = 6.7) (3) E6C (7.5 kDa; pI = 9.7)	Improved solubility and sample stability of HPV-16 E6 protein	$GB1^{basic}$ is a GB1-mutant (D22N, D36R, and E42K) with a pI of 8.0. It was used to express basic target proteins to avoid aggregation. Use of the $GB1^{basic}$ tag allowed preparation of stable NMR samples of E6N and E6C at 2 mM, and E6 at 0.2 mM. The intrinsic solubility of E6N (after removing the GST-tag from GST-E6N) was in the range of hundreds of μM (Liu et al. 2009)
GB1 (invisible C-terminal tag; 6.2 kDa; pI = 4.5)	Vav C terminus SH3 (7.5 kDa; pI = 6.4)	Enhanced the solubility of VcSH3 by more than tenfold	The target protein was initially expressed as an N-terminal GB1-fusion construct. A sortase-mediated protein ligation method was used to ligate a second, unlabeled GB1 to the C terminus of the target protein. The N-terminal GB1 tag was subsequently removed by protease cleavage. The Vav C terminus SH3 was almost insoluble at physiological pH. Using invisible C-terminal GB1 tag enabled preparation of stable NMR samples at 0.6 mM and subsequent structural determination (PDB: 2KBT) (Kobashigawa et al. 2009)
GB1 (6.2 kDa; pI = 4.5)	Borealin (1) Full length (31.3 kDa; pI = 9.84) (2) residues 13–92 (9.4 kDa; pI = 5.48)	Significantly improved the protein yield in the soluble fractions	(Zhou et al. 2009)
Calmodulin (CaM; invisible tag; 16.8 kDa, pI = 4.1)	Sterile alpha motif (SAM) from p63 (7.5 kDa; pI = 5.8)	Enhanced solubility by over 20-fold	The target protein was inserted between GST and the calmodulin binding peptide (CBP). The unlabeled calmodulin, which serves the role of a solubility-enhancement tag, was added to form a CBP-calmodulin complex. The N-terminal GST-tag was then removed by protease cleavage (Durst et al. 2008)
GB1 (6.2 kDa; pI = 4.5)	17β-hydroxysteroid dehydrogenase type 1 (HSD17β1) (homodimer with a molecular weight of 70 kDa)	Increased sample stability	The fusion protein formed soluble aggregates at high concentrations, but maintained enzymatic activity to allow NMR-based inhibitor studies (Ludwig et al. 2008)
GB1 (6.2 kDa; pI = 4.5)	Potassium channel-interacting protein 4a (KChIP4a, residues 1–34; 3.7 kDa; pI = 4.0)	Enhanced solubility	(Schwenk et al. 2008)
GB1 (C-terminal tag; 6.2 kDa; pI = 4.5)	CK2 substrate (XT111–132; 2.4 kDa; pI = 8.2)	Enhanced solubility of fused peptide in live cells	GB1 was used as a soluble carrier of a phosphorylation site and provided the solubility needed for recording spectra in live cells (Selenko et al. 2008)
GB1 (6.2 kDa; pI = 4.5)	mRNA-decapping enzyme Dcp2 Nudix domain (17.3 kDa; pI = 8.5)	Enhanced solubility	The untagged Nudix domain was only marginally soluble. The GB1-tagged protein (in the presence of Arg/Glu additives) was stable at 0.5 mM for several weeks (PDB: 2JVB) (Deshmukh et al. 2008)
GB1 (6.2 kDa; pI = 4.5)	Eukaryotic translation initiation factor eIF5 (residues 241–405; 19.3 kDa; pI = 5.2)	Enhanced solubility	(Reibarkh et al. 2008)

Table 1 continued

Tag	Target protein and property	Reported effect(s)	Notes and references
GB1 (6.2 kDa; pI = 4.5)	Parkin ubiquitin like domain mutant ($Ubld^{R42P}$) (8.8 kDa; pI = 6.7)		The GB1 tag was used to overcome the poor expression and degradation of the $Ubld^{R42P}$ mutant; without the GB1 tag, the $Ubld^{R42P}$ could not be isolated (Safadi and Shaw 2007)
GB1 (6.2 kDa; pI = 4.5)	A ubiquitin variant found at the N terminus of S27a in *Giardia lamblia* ($GlUb_{S27A}$; 7.0 kDa; pI = 4.7)	Enhanced solubility/sample stability	No protein expression was observed with the His- or HA-tagged constructs. The GB1-tagged $GlUb_{S27A}$ was stable at 1 mM for about a week at 25°C (Catic et al. 2007)
GB1 (6.2 kDa; pI = 4.5)	Fas death domain (Fas-DD; 9.9 kDa; pI = 8.7)	Increased sample stability/ solubility	The untagged Fas-DD had an intrinsic tendency to form soluble aggregates at physiological pH (Ferguson et al. 2007)
GB1 (6.2 kDa; pI = 4.5)	Inositol 1,4,5-trisphosphate receptor (IP3R) intraluminal loop L3-2 (2.3 kDa; pI = 6.3)	Increased sample stability/ solubility	No protein expression was observed with a His-tagged construct (Kang et al. 2007)
GB1 (6.2 kDa; pI = 4.5)	Murine eIF4E (25 kDa; pI = 5.8)	Greatly enhanced solubility	(Untagged) mammalian eIF4E behaved poorly in solution (Moerke et al. 2007)
Poly Arg or Lys peptide tags	BPTI-22 (a BPTI variant containing 22 alanines)	Enhanced solubility by fourfold to sixfold	Kato et al. (2007)
GB1 (6.2 kDa; pI = 4.5)	SRp20 RNA recognition motif (RRM; 9.6 kDa; pI = 6.6)	Enhanced solubility	Poor solubility of the untagged protein prevented NMR studies. The GB1-SRp20 RRM was stable at 1 mM, which enabled structural studies (PDB: 2I38 & 2I2Y) (Hargous et al. 2006)
GB1 (6.2 kDa; pI = 4.5)	9G8 RNA recognition motif (9G8 RRM; 11.3 kDa; pI = 9.6)	Enhanced solubility	Poor solubility of the untagged protein prevented NMR studies. The GB1-9G8 RRM (in the presence of Arg/Glu additives) was stable at 1 mM (Hargous et al. 2006)
GB1 (6.2 kDa; pI = 4.5)	UBA domain of human bone marrow stromal cells ubiquitin-like protein (BMSC-UbP; 4.8 kDa; pI = 4.0)	Dramatically enhanced the solubility	The untagged UBA domain readily precipitated in solution. The GB1-UBA was stable at 1 mM (PDB: 2CWB) (Chang et al. 2006)
GB1 (6.2 kDa; pI = 4.5)	Rat ADAR2 double-stranded RNA binding domain (dsRBD; 24.3 kDa; pI = 6.2)	Improved protein expression and solubility	The untagged rat ADAR2 dsRBD12 (74–301) had low solubility in common NMR buffers. The GB1-fusion protein was stable at 0.8 mM (Stefl et al. 2005, 2006)
GB1 (invisible tag)	Chitin-binding domain	Not reported	Used an intein-based strategy to incorporate the unlabeled GB1 tag into isotopically labeled proteins (Züger and Iwai 2005)
GB1 (6.2 kDa; pI = 4.5)	Eukaryotic translation initiation factor 2 gamma (eIF2γ; 51 kDa; pI = 8.7)	Enhanced solubility	GB1 was used to enhance the solubility of eIF2γ to enable studies of its interaction with eIF2α (Ito et al. 2004)
GB1 (6.2 kDa; pI = 4.5)	Mutant myotoxin a (MyoP20G; 4.7 kDa; pI = 9.5)	Increased the expression yield and enhanced the refolding efficiency	Untagged protein refolded poorly. The GB1 tag was removed after refolding (Cheng and Patel 2004)

Table 1 continued

Tag	Target protein and property	Reported effect(s)	Notes and references
GB1 (6.2 kDa; pI = 4.5)	Human Ki67 FHA domain (hNIFK; 5 kDa; pI = 4.5)	Increased the protein yield and sample stability	Li et al. (2004)
GB1 (6.2 kDa; pI = 4.5)	NALP1 pyrin domain (10 kDa; pI = 5.9)	Enhanced solubility by ~100-fold	Untagged protein aggregated at concentrations above ~10 μM. The GB1-tagged protein was stable at 1 mM (PDB: 1PN5) (Hiller et al. 2003)
GB1 (6.2 kDa; pI = 4.5)	Human T-cell leukemia virus 1 (HTLV-1) Tax40N (4.3 kDa; pI = 6.0)	Not reported	Li et al. (2003)
GB1 (6.2 kDa; pI = 4.5)	eIF5B-CTD (16.7 kDa; pI = 8.7)	Enhanced solubility	Marintchev et al. (2003)
MBP (40.7 kDa; pI = 5.2)	Integrin $\alpha_{IIb}\beta_3$ (MW of $\beta 3$ is 5.5 kDa; pI = 9.2)	Enhanced solubility	(PDB: 1M8O) (Vinogradova et al. 2002)

The SET should not interact with the target protein or protein complex

Ideally, a solubility-enhancement tag should be "transparent" to the target protein, i.e., it should not perturb the structure or function of the target protein. In the absence of such prior knowledge, proper control experiments must be included to demonstrate the "inertness" of the solubility-enhancement tag for functional assays. Likewise, the lack of perturbations of tag resonances in the fusion protein provides a compelling argument that the solubility-enhancement tag does not interact with the target protein and is unlikely to alter its structure.

In this regard, GB1 appears to be remarkably "transparent" as demonstrated in a variety of GB1-fusion proteins in NMR studies (Table 1). Interestingly, many examples of the GB1-fusion proteins in NMR studies also display better sample stability at high concentrations (μM–mM). Because the "passive" GB1 tag is unlikely to alter the *thermal* stability of the target protein, the improved sample stability presumably results from the enhanced solubility and reduced aggregation of the fusion protein.

Because GB1 is slightly acidic (pI = 4.5), it may cause non-specific electrostatic interactions when fused to proteins with basic pI values. To avoid these non-specific interactions, we created a GB1 mutant ($GB1^{basic}$) by mutating D22N, D36R, and E42K, which increased the pI of GB1 to 8.0 (Zhou and Wagner, unpublished). This basic GB1 tag was successfully utilized to prepare highly soluble HPV16 E6 samples and prevent non-specific electrostatic interactions between the tag and the target protein (Liu et al. 2009). Without the tag, the solubility of the E6 constructs was too low to record spectra (J. Baleja, private communication). Consistent with this notion of choosing a SET based on matching its charge state with that of the target protein, Harrison et al. showed in their statistical model that avoidance of charge neutralization increases the probability of producing soluble proteins in *E. coli* (Davis et al. 1999; Wilkinson and Harrison 1991).

It should be noted that an "active" fusion tag can also be highly effective. For example, Mal et al. fused the TAF N-terminal domain 1 and 2 (TAND12) with its binding partner TATA-binding protein (TBP) to form a stable protein complex, which displayed enhanced solubility and sample stability (Mal et al. 2007). This is also called single-chain approach and has been used frequently, such as for NMR studies of receptor dimers (Sun et al. 2001). However, such an "active" fusion tag is target specific and cannot be easily applied to other proteins.

The SET should be highly soluble

Assuming that (1) there is no interaction between the tag and the target protein, (2) there is no structural change of

either the tag or the target in the fusion protein, and (3) the contribution of the linker can be neglected, we give an estimation of the solubility-enhancement effect based on a simple thermodynamic model. Although the analysis below focuses on fusion proteins containing a single tag, it is straightforward to extend such an analysis to fusion proteins with multiple tags.

The free energies of individually transferring A (the tag) and B (the target protein) from the solid state to the solution state are given by:

$$\Delta G_A = \Delta G_A^\circ + RT\ \ln([A]_{\text{solution}}/[A]_{\text{solid}})$$
$$\Delta G_B = \Delta G_B^\circ + RT\ \ln([B]_{\text{solution}}/[B]_{\text{solid}}). \quad (1)$$

At equilibrium (i.e., at saturation), the free energy of transferring the A and B from the solid state to the solution state is zero. Therefore one has:

$$0 = \Delta G_A^\circ + RT\ \ln([A]^{\text{saturation}}_{\text{solution}}/[A]_{\text{solid}})$$
$$0 = \Delta G_B^\circ + RT\ \ln([B]^{\text{saturation}}_{\text{solution}}/[B]_{\text{solid}}), \quad (2)$$

which can be re-arranged to give

$$-RT\ \ln([A]^{\text{saturation}}_{\text{solution}}) = \Delta G_A^\circ - RT\ \ln([A]_{\text{solid}})$$
$$-RT\ \ln([B]^{\text{saturation}}_{\text{solution}}) = \Delta G_B^\circ - RT\ \ln([B]_{\text{solid}}). \quad (3)$$

With Eq. 3, one can rewrite Eq. 1 as

$$\Delta G_A = RT\ \ln\left([A]_{\text{solution}}/[A]^{\text{saturation}}_{\text{solution}}\right)$$
$$\Delta G_B = RT\ \ln\left([B]_{\text{solution}}/[B]^{\text{saturation}}_{\text{solution}}\right). \quad (4)$$

If there is no interaction between A and B, we can conceptually describe the transfer of the fusion protein A–B from the solid state to the solution state as two separate processes: transferring A_{solid} to A_{solution} and transferring B_{solid} to B_{solution}. The free energy of such a combined transfer is zero at equilibrium.

$$0 = \Delta G^{\text{saturation}}_{A-B} = \Delta G_A^{(\text{saturation in}\cdot A-B)} + \Delta G_B^{(\text{saturation in}\cdot A-B)}$$
$$= RT\ \ln\left([A]^{(\text{saturation in}\cdot A-B)}_{\text{solution}}/[A]^{\text{saturation}}_{\text{solution}}\right)$$
$$+ RT\ \ln\left([B]^{(\text{saturation in}\cdot A-B)}_{\text{solution}}/[B]^{\text{saturation}}_{\text{solution}}\right) \quad (5)$$

Because the covalent linker requires

$$[A]^{(\text{in}\cdot A-B)}_{\text{solution}} = [B]^{(\text{in}\cdot A-B)}_{\text{solution}} = [A-B]_{\text{solution}}, \quad (6)$$

by substituting $[A]^{(\text{saturation in}\cdot A-B)}_{\text{solution}}$ and $[B]^{(\text{saturation in}\cdot A-B)}_{\text{solution}}$ with $[A-B]^{\text{saturation}}_{\text{solution}}$, we can rewrite Eq. 5 as

$$0 = RT\ \ln\left([A-B]^{\text{saturation}}_{\text{solution}}/[A]^{\text{saturation}}_{\text{solution}}\right) + RT\ \ln\left([A-B]^{\text{saturation}}_{\text{solution}}/[B]^{\text{saturation}}_{\text{solution}}\right)$$
$$= RT\ \ln\left(\frac{[A-B]^{\text{saturation}}_{\text{solution}} \times [A-B]^{\text{saturation}}_{\text{solution}}}{[A]^{\text{saturation}}_{\text{solution}} \times [B]^{\text{saturation}}_{\text{solution}}}\right), \quad (7)$$

which requires

$$\frac{\left([A-B]^{\text{saturation}}_{\text{solution}}\right)^2}{[A]^{\text{saturation}}_{\text{solution}} \times [B]^{\text{saturation}}_{\text{solution}}} = 1 \quad (8)$$

Therefore, we have the saturation concentration of the fusion protein as:

$$[A-B]^{\text{saturation}}_{\text{solution}} = \sqrt{[A]^{\text{saturation}}_{\text{solution}} \times [B]^{\text{saturation}}_{\text{solution}}} \quad (9)$$

We note that the above analysis does not account for changes of solid or solution state compositions, nor does it take into consideration of intermediate species (such as $A_{\text{solid}} - B_{\text{solution}}$ and $A_{\text{solution}} - B_{\text{solid}}$) of the solvation process. The latter approximation, in particular, can introduce a very large error in the solubility estimation of the fusion protein. Finally, strictly speaking, the concentration terms of Eq. 9 should be effective concentrations (i.e., activities), which may deviate from the apparent protein concentrations. This effect is expected to be larger at higher concentrations, which can result in an overestimation of the effective tag concentration at saturation. Because of these limitations, Eq. 9 can only be used in a qualitative way. It nevertheless gives a useful evaluation of the beneficial effect brought by a solubility-enhancement tag.

To give an example, we were able to make 15–20 mM GB1 solutions routinely without any noticeable precipitations. Using these numbers as the solubility of GB1, we estimate that the SET approach yields a saturation concentration of 1.2–1.4 or 0.38–0.44 mM for a target protein with inherent solubility of 0.1 or 0.01 mM respectively, corresponding to a $\sim$10- to 40-fold enhancement of the solubility over the untagged protein! Experimentally, approximately 3- to 100-fold enhancements of solubility have been reported for GB1-fusion proteins (Hiller et al. 2003; Kobashigawa et al. 2009; Zhou et al. 2001b). The largest effect was reported for the pyrin domain of NALP1, which saw its solubility increased from $\sim$10 μM to 1 mM (Hiller et al. 2003).

Equation 9 argues that proteins with higher intrinsic solubility, but not with larger molecular weights, function as better tags. Although this conclusion may seem counterintuitive, several large-scale solubility studies have consistently categorized the small GB1 tag (5.6 kDa) as one of the most effective tags to use (Hammarström et al. 2002, 2006). For example, Hammarström compared the

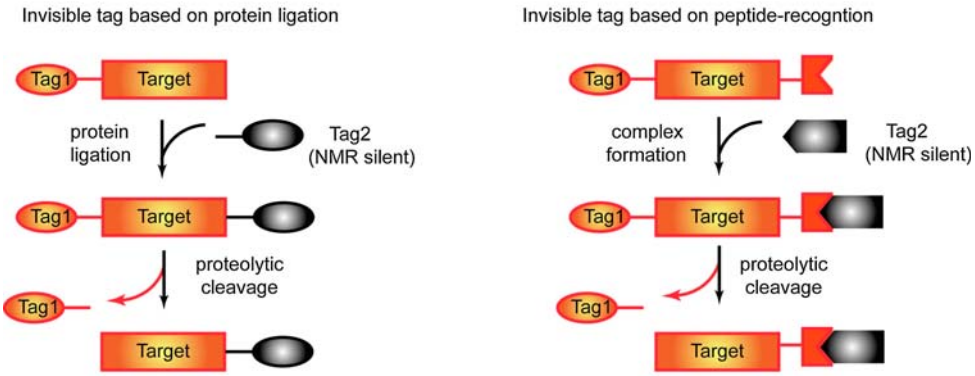

Fig. 2 NMR-invisible solubility-enhancement tags

effect of different tags on the solubility of 27 small- to medium-sized human proteins, and ranked GB1, MBP and thioredoxin as the best tags (Hammarström et al. 2002). The authors concluded that the there was no statistical difference of GB1, MBP and thioredoxin in their ability to enhance the solubility of a target protein. It is important to note that in most of the studies, the solubility (often reported as gel intensity) reflects the mass yield of the fusion proteins, but not the untagged target proteins. This could lead to an overestimation of the solubility-enhancement effect for large tags such as MBP or NusA. After correcting for the molecular weight contributions from different tags, Hammarstrom et al. (2006) concluded that GB1 gave a significantly larger amount of soluble target proteins for the 45 human proteins tested.

Finally, we would like to emphasize that Eq. 9 is based on a thermodynamic analysis. It assumes no interaction between the tag and the target protein and requires the solvation process to be fully reversible. Several protein tags have been shown to facilitate protein folding in *E. coli* by promoting disulfide bond formation (Stewart et al. 1998), by serving as a molecular chaperone (Bach et al. 2001; Kapust and Waugh 1999) or by enhancing transcription pausing (Davis et al. 1999). In these scenarios, the significantly better "solubilizing" effect of the "active" tags over "passive" tags may reflect the benefit of folding kinetics, but not thermodynamics.

The SET should be highly stable

Because NMR experiments are performed under a variety of pH, temperature and buffer conditions, a good solubility-enhancement tag should be stable under these conditions. The rapid two-state refolding property of a tag can also be highly beneficial. For example, in the study of mutant myotoxin a (MyoP20G), Cheng and Patel (2004) reported that GB1 appears to increase protein (re)folding efficiency, which likely comes from the enhanced solubility (and reduced aggregation) of the denatured fusion protein.

The SET can increase the overexpression level and yield of the target protein

As reported in early literature, a successful solubility-enhancement tag often enhances protein overexpression levels and increases the yields of the purified proteins. Some tags, such as MBP and thioredoxin, have been suggested to serve as chaperones to promote proper folding of target proteins (Bach et al. 2001; Kapust and Waugh 1999; Kern et al. 2003). Although similar benefits in protein expression levels and yields have been observed for GB1-fusion proteins (Table 1; also see studies by Hammarström et al. 2002, 2006), the experimental evidence for the chaperone activity of GB1 is lacking. It should be noted that such effects do not have to derive from the chaperone activity. The enhanced solubility of the fusion protein itself is expected to facilitate protein folding and overexpression in vivo and increase the yield of protein purification in vitro by reducing protein aggregation and precipitation.

Several studies reported diminished effects of SETs on the *E. coli* expression of large proteins (>25–30 kDa) in soluble fractions (Hammarström et al. 2002, 2006). Because large proteins frequently require chaperones or binding partners to fold properly, it is likely that these observations reflect an intrinsic folding (kinetic) problem of the large proteins, rather than the ineffectiveness of SETs.

Invisible SETs

Despite the success of the SET approach, it still brings a sizeable amount of extra signals from the protein tag. For a target protein of 10–20 kDa, inclusion of a small GB1 tag (56 residues) easily adds about a quarter to a half of "extra" signals to those from the untagged protein. Although the excellent signal dispersion and the lack of resonance perturbation make the tag signals easy to identify, they nevertheless bring extra burden and complexity for resonance assignment.

Recently, two types of NMR-invisible tags have used to overcome this issue (Fig. 2; Durst et al. 2008; Kobashigawa et al. 2009; Züger and Iwai 2005). Both approaches start from an isotopically enriched fusion protein containing a cleavable solubility tag. A second and unlabeled solubility tag—which is invisible by NMR—is then introduced to maintain solubility. The isotopically labeled tag is subsequently removed to generate the final form of the NMR sample.

The two approaches differ in how the NMR-invisible tag was introduced. In the first approach, the unlabeled GB1 tag was attached to the isotopically labeled chitin-binding domain or the Vav C terminus SH3 domain using either an intein-based or a sortase-mediated protein ligation strategy (Kobashigawa et al. 2009; Züger and Iwai 2005). Because the yield of the final fusion protein depends on the ligation efficiency, optimization of the ligation condition is critical for the general application of this approach. In the second approach, a calmodulin-binding peptide (CBP, 23 residues) was included in the construct of the GST-tagged target protein (Durst et al. 2008). The unlabeled calmodulin, which binds the CBP, was added to the solution. After formation of the calmodulin/CBP complex, the isopotically labeled GST-tag was removed by proteolytic cleavage, and the unlabeled calmodulin served as the NMR-invisible solubility-enhancement tag. Because the latter approach bypasses the protein ligation step completely, it is more convenient to use. However, there is no reason why one should be restricted to the CBP tag of 23 residues; systems using shorter peptides and the corresponding high-affinity binding partners are likely to emerge in the future.

Conclusion

The preparation of highly soluble and stable samples represents a significant challenge for solution NMR studies of proteins with inherent poor solubility and stability. The use of solubility-enhancement tags has been demonstrated to overcome sample solubility and stability barriers and has enabled detailed structural analyses of many poorly-behaving proteins. The recent development of NMR-invisible tags promises to further expand the application of the SET strategy in biomolecular NMR.

Acknowledgements This work was supported by NIH (grants GM47467 to GW and GM079376 to PZ). PZ would like to thank Prof. Terrence G. Oas (Department of Biochemistry, Duke University Medical Center) for critical reading of the manuscript.

References

Bach H, Mazor Y, Shaky S, Shoham-Lev A, Berdichevsky Y, Gutnick DL, Benhar I (2001) Escherichia coli maltose-binding protein as a molecular chaperone for recombinant intracellular cytoplasmic single-chain antibodies. J Mol Biol 312:79–93

Bagby S, Tong KI, Liu D, Alattia JR, Ikura M (1997) The button test: a small scale method using microdialysis cells for assessing protein solubility at concentrations suitable for NMR. J Biomol NMR 10:279–282

Catic A, Sun ZY, Ratner DM, Misaghi S, Spooner E, Samuelson J, Wagner G, Ploegh HL (2007) Sequence and structure evolved separately in a ribosomal ubiquitin variant. EMBO J 26:3474–3483

Chang YG, Song AX, Gao YG, Shi YH, Lin XJ, Cao XT, Lin DH, Hu HY (2006) Solution structure of the ubiquitin-associated domain of human BMSC-UbP and its complex with ubiquitin. Protein Sci 15:1248–1259

Cheng Y, Patel DJ (2004) An efficient system for small protein expression and refolding. Biochem Biophys Res Commun 317:401–405

Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I et al (2000) Structural proteomics of an archaeon. Nat Struct Biol 7:903–909

Davis GD, Elisee C, Newham DM, Harrison RG (1999) New fusion protein systems designed to give soluble expression in *Escherichia coli*. Biotechnol Bioeng 65:382–388

DelProposto J, Majmudar CY, Smith JL, Brown WC (2009) Mocr: a novel fusion tag for enhancing solubility that is compatible with structural biology applications. Protein Expr Purif 63:40–49

Deshmukh MV, Jones BN, Quang-Dang DU, Flinders J, Floor SN, Kim C, Jemielity J, Kalek M, Darzynkiewicz E, Gross JD (2008) mRNA decapping is promoted by an RNA-binding channel in Dcp2. Mol Cell 29:324–336

di Guan C, Li P, Riggs PD, Inouye H (1988) Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. Gene 67:21–30

Durst FG, Ou HD, Lohr F, Dotsch V, Straub WE (2008) The better tag remains unseen. J Am Chem Soc 130:14932–14933

Ferguson BJ, Esposito D, Jovanovic J, Sankar A, Driscoll PC, Mehmet H (2007) Biophysical and cell-based evidence for differential interactions between the death domains of CD95/Fas and FADD. Cell Death Differ 14:1717–1719

Forrer P, Jaussi R (1998) High-level expression of soluble heterologous proteins in the cytoplasm of *Escherichia coli* by fusion to the bacteriophage lambda head protein D. Gene 224:45–52

Golovanov AP, Hautbergue GM, Wilson SA, Lian LY (2004) A simple method for improving protein solubility and long-term stability. J Am Chem Soc 126:8933–8939

Hammarström M, Hellgren N, van Den Berg S, Berglund H, Hard T (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. Protein Sci 11:313–321

Hammarström M, Woestenenk EA, Hellgren N, Hard T, Berglund H (2006) Effect of N-terminal solubility enhancing fusion proteins on yield of purified target protein. J Struct Func Genom 7:1–14

Hargous Y, Hautbergue GM, Tintaru AM, Skrisovska L, Golovanov AP, Stevenin J, Lian LY, Wilson SA, Allain FH (2006) Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. EMBO J 25:5126–5137

Hiller S, Kohl A, Fiorito F, Herrmann T, Wider G, Tschopp J, Grutter MG, Wuthrich K (2003) NMR structure of the apoptosis- and inflammation-related NALP1 pyrin domain. Structure 11:1199–1205

Hornemann S, Christen B, von Schroetter C, Perez DR, Wüthrich K (2009) Prion protein library of recombinant constructs for structural biology. FEBS J 276:2359–2367

Huang B, Eberstadt M, Olejniczak ET, Meadows RP, Fesik SW (1996) NMR structure and mutagenesis of the Fas (APO-1/CD95) death domain. Nature 384:638–641

Huth JR, Bewley CA, Jackson BM, Hinnebusch AG, Clore GM, Gronenborn AM (1997) Design of an expression system for detecting folded protein domains and mapping macromolecular interactions by NMR. Protein Sci 6:2359–2364

Ito T, Wagner G (2004) Using codon optimization, chaperone co-expression, and rational mutagenesis for production and NMR assignments of human eIF2 alpha. J Biomol NMR 28:357–367

Ito T, Marintchev A, Wagner G (2004) Solution structure of human initiation factor eIF2alpha reveals homology to the elongation factor eEF1B. Structure 12:1693–1704

Kang J, Kang S, Yoo SH, Park S (2007) Identification of residues participating in the interaction between an intraluminal loop of inositol 1, 4, 5-trisphosphate receptor and a conserved N-terminal region of chromogranin B. Biochim Biophys Acta 1774:502–509

Kapust RB, Waugh DS (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. Protein Sci 8:1668–1674

Kato A, Maki K, Ebina T, Kuwajima K, Soda K, Kuroda Y (2007) Mutational analysis of protein solubility enhancement using short peptide tags. Biopolymers 85:12–18

Kern R, Malki A, Holmgren A, Richarme G (2003) Chaperone properties of *Escherichia coli* thioredoxin and thioredoxin reductase. Biochem J 371:965–972

Kobashigawa Y, Kumeta H, Ogura K, Inagaki F (2009) Attachment of an NMR-invisible solubility enhancement tag using a sortase-mediated protein ligation method. J Biomol NMR 43:145–150

LaVallie ER, Lu Z, Diblasio-Smith EA, Collins-Racie LA, McCoy JM (2000) Thioredoxin as a fusion partner for production of soluble recombinant proteins in *Escherichia coli*. Methods Enzymol 326:322–340

Lepre CA, Moore JM (1998) Microdrop screening: a rapid method to optimize solvent conditions for NMR spectroscopy of proteins. J Biomol NMR 12:493–499

Li J, Li H, Tsai MD (2003) Direct binding of the N-terminus of HTLV-1 tax oncoprotein to cyclin-dependent kinase 4 is a dominant path to stimulate the kinase activity. Biochemistry 42:6921–6928

Li H, Byeon IJ, Ju Y, Tsai MD (2004) Structure of human Ki67 FHA domain and its binding to a phosphoprotein fragment from hNIFK reveal unique recognition sites and new views to the structural basis of FHA domain functions. J Mol Biol 335:371–381

Liu Y, Cherry JJ, Dineen JV, Androphy EJ, Baleja JD (2009) Determinants of stability for the E6 protein of papillomavirus type 16. J Mol Biol 386:1123–1137

Ludwig C, Michiels PJ, Lodi A, Ride J, Bunce C, Gunther UL (2008) Evaluation of solvent accessibility epitopes for different dehydrogenase inhibitors. Chem Med Chem 3:1371–1376

Mal TK, Takahata S, Ki S, Zheng L, Kokubo T, Ikura M (2007) Functional silencing of TATA-binding protein (TBP) by a covalent linkage of the N-terminal domain of TBP-associated factor 1. J Biol Chem 282:22228–22238

Marintchev A, Kolupaeva VG, Pestova TV, Wagner G (2003) Mapping the binding interface between human eukaryotic initiation factors 1A and 5B: a new interaction between old partners. Proc Natl Acad Sci USA 100:1535–1540

Moerke NJ, Aktas H, Chen H, Cantel S, Reibarkh MY, Fahmy A, Gross JD, Degterev A, Yuan J, Chorev M et al (2007) Small-molecule inhibition of the interaction between the translation initiation factors eIF4E and eIF4G. Cell 128:257–267

Pilon AL, Yost P, Chase TE, Lohnas GL, Bentley WE (1996) High-level expression and efficient recovery of ubiquitin fusion proteins from *Escherichia coli*. Biotechnol Prog 12:331–337

Reibarkh M, Yamamoto Y, Singh CR, del Rio F, Fahmy A, Lee B, Luna RE, Ii M, Wagner G, Asano K (2008) Eukaryotic initiation factor (eIF) 1 carries two distinct eIF5-binding faces important for multifactor assembly and AUG selection. J Biol Chem 283:1094–1103

Safadi SS, Shaw GS (2007) A disease state mutation unfolds the parkin ubiquitin-like domain. Biochemistry 46:14162–14169

Samuelsson E, Moks T, Nilsson B, Uhlen M (1994) Enhanced in vitro refolding of insulin-like growth factor I using a solubilizing fusion partner. Biochemistry 33:4207–4211

Schwenk J, Zolles G, Kandias NG, Neubauer I, Kalbacher H, Covarrubias M, Fakler B, Bentrop D (2008) NMR analysis of KChIP4a reveals structural basis for control of surface expression of Kv4 channel complexes. J Biol Chem 283:18937–18946

Selenko P, Frueh DP, Elsaesser SJ, Haas W, Gygi SP, Wagner G (2008) In situ observation of protein phosphorylation by high-resolution NMR spectroscopy. Nat Struct Mol Biol 15:321–329

Smith DB, Johnson KS (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. Gene 67:31–40

Stefl R, Skrisovska L, Xu M, Emeson RB, Allain FH (2005) Resonance assignments of the double-stranded RNA-binding domains of adenosine deaminase acting on RNA 2 (ADAR2). J Biomol NMR 31:71–72

Stefl R, Xu M, Skrisovska L, Emeson RB, Allain FH (2006) Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. Structure 14:345–355

Stewart EJ, Aslund F, Beckwith J (1998) Disulfide bond formation in the Escherichia coli cytoplasm: an in vivo role reversal for the thioredoxins. EMBO J 17:5543–5550

Sun ZY, Dötsch V, Kim M, Li J, Reinherz EL, Wagner G (1999) Functional glycan-free adhesion domain of human cell surface receptor CD58: design, production and NMR studies. EMBO J 18:2941–2949

Sun ZJ, Kim KS, Wagner G, Reinherz EL (2001) Mechanisms contributing to T cell receptor signaling and assembly revealed by the solution structure of an ectodomain fragment of the CD3 epsilon gamma heterodimer. Cell 105:913–923

Vinogradova O, Velyvis A, Velyviene A, Hu B, Haas T, Plow E, Qin J (2002) A structural mechanism of integrin alpha(IIb)beta(3) "inside-out" activation as regulated by its cytoplasmic face. Cell 110:587–597

Waugh DS (2005) Making the most of affinity tags. Trends Biotechnol 23:316–320

Wilkinson DL, Harrison RG (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. Bio/technology (Nature Publishing Company) 9:443–448

Zhou P, Lugovskoy AA, McCarty JS, Li P, Wagner G (2001a) Solution structure of DFF40 and DFF45 N-terminal domain complex and mutual chaperone activity of DFF40 and DFF45. Proc Natl Acad Sci USA 98:6051–6055

Zhou P, Lugovskoy AA, Wagner G (2001b) A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. J Biomol NMR 20:11–14

Zhou L, Li J, George R, Ruchaud S, Zhou HG, Ladbury JE, Earnshaw WC, Yuan X (2009) Effects of full-length Borealin on the composition and protein-protein interaction activity of a binary chromosomal passenger complex. Biochemistry 48:1156–1161

Zou Z, Cao L, Zhou P, Su Y, Sun Y, Li W (2008) Hyper-acidic protein fusion partners improve solubility and assist correct folding of recombinant proteins expressed in *Escherichia coli*. J Biotechnol 135:333–339

Züger S, Iwai H (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. Nat Biotechnol 23:736–740

Zuo X, Mattern MR, Tan R, Li S, Hall J, Sterner DE, Shoo J, Tran H, Lim P, Sarafianos SG et al (2005) Expression and purification of SARS coronavirus proteins using SUMO-fusions. Protein Expr Purif 42:100–110