Protein Expression and Purification 116 (2015) 42–49

Contents lists available at ScienceDirect

Protein Expression and Purification

journal homepage: www.elsevier.com/locate/yprep

Tandem SUMO fusion vectors for improving soluble protein expression and purification

Fernando Guerrero, Annika Ciragan, Hideo Iwaï *

Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, P.O. Box 65, Helsinki FIN-00014, Finland

ARTICLE INFO

Article history:
Received 22 June 2015
and in revised form 17 August 2015
Accepted 18 August 2015
Available online 20 August 2015

Keywords:
Fusion proteins
Solubility enhancement tag
SUMO system
Ulp1 protease
TonB
Scytovirin
Protein ligation

ABSTRACT

Availability of highly purified proteins in quantity is crucial for detailed biochemical and structural investigations. Fusion tags are versatile tools to facilitate efficient protein purification and to improve soluble overexpression of proteins. Various purification and fusion tags have been widely used for overexpression in *Escherichia coli*. However, these tags might interfere with biological functions and/or structural investigations of the protein of interest. Therefore, an additional purification step to remove fusion tags by proteolytic digestion might be required. Here, we describe a set of new vectors in which yeast SUMO (SMT3) was used as the highly specific recognition sequence of ubiquitin-like protease 1, together with other commonly used solubility enhancing proteins, such as glutathione S-transferase, maltose binding protein, thioredoxin and trigger factor for optimizing soluble expression of protein of interest. This tandem SUMO (T-SUMO) fusion system was tested for soluble expression of the C-terminal domain of TonB from different organisms and for the antiviral protein scytovirin.

1. Introduction

Although the use of recombinant proteins has been a valuable advance in recent times, the choice of the appropriate host and expression system needs to be optimized on a case-by-case basis according to the target protein [1,2]. Purification tags like polyhistidine-tag are indispensable for facilitating efficient protein purification of heterogeneous proteins overexpressed in *Escherichia coli* [3,4]. In addition, various proteins have been used as fusion tags in combination with purification tags for improving properties like solubility and expression levels of target proteins [5]. Glutathione S-transferase (GST) [6], thioredoxin (TRX) [7], DsbA [8], maltose binding protein (MBP) [9], trigger factor (TF) [10] and others have been often used [11]. Various vectors for fusion proteins are already commercially available. However, fusion tags might interfere with biological assays or structural investigations, making it necessary to remove them before carrying out such studies. Thus, additional steps of proteolytic cleavage and subsequent removal of the proteolytic enzyme and fusion tag are applied to produce tag-free target proteins of interest. Widely used and commercially available proteases for the specific cleavages are thrombin, factor Xa, enterokinase, TEV protease, and precision protease [12]. Due to their lower specificity and instability of the target proteins, undesired cleavages have been observed outside the expected cleavage site [13]. For example, thrombin cannot differentiate between Ser and Cys in their recognition [14]. Moreover, these commercial enzymes might not be cost-effective when large-scale protein production is required, restricting their industrial scale applications in biotechnology. An alternative might be the use of thiol-inducible self-cleavable intein tags, which do not require additional proteases [15]. Instead, it utilizes an autocatalytic self-cleavage reaction induced by thiol reagents to avoid this problem. However, premature cleavage has been observed, thereby reducing the purification efficiency, and they often require optimization of parameters such as expression temperature and junction sequences [16,17]. In addition, the reducing condition used for the cleavage might not be compatible with some target proteins bearing disulfide bridges.

Here we report a set of new vectors in which small ubiquitin-like modifier (SUMO or SUMO homologue, SMT3) from yeast is used as cleavage tag, in tandem with other fusion tags such as TRX, TF, MBP and GST for solubility and expression enhancement. These tandem fusion vectors utilize the high specificity of ubiquitin-like protease 1 (Ulp1), which recognizes the three-dimensional structure of SUMO domain and cleaves after di-glycine at the C-terminus [18]. We demonstrated soluble expression and purification of the

Abbreviations: HSQC, heteronuclear single quantum coherence NMR spectrum; RF cloning, restriction free cloning (cloning which is independent of restriction sites).

* Corresponding author.
E-mail address: hideo.iwai@helsinki.fi (H. Iwaï).

http://dx.doi.org/10.1016/j.pep.2015.08.019
1046-5928/

C-terminal domain (CTD) of TonB protein from three organisms, and a small lectin protein scytovirin for optimal choice of a tandem fusion vector.

2. Materials and methods

2.1. Construction of plasmids

The backbone plasmid used to construct the expression vectors was pHYRSF53 [19,20]. Four different fusion tags were inserted in frame upstream of the coding sequence of SMT3. Two PCR steps were necessary to add a hexa-histidine (H_6) tag at the N-terminus of the cloned fusion tag. In a first PCR step, the coding sequences of the fusion tags were amplified using synthetic oligonucleotides. In all cases the forward primer contained an overhang coding part of a H_6-tag. The full H_6-tag and an *Nco*I site for cloning were added in a second PCR reaction using the first PCR product as a template and the oligonucleotide HK683: 5′-TAC CATGGGCAGCAGCCATCATCATCATCACGG as a forward primer. The resulting amplicon containing the H_6-fusion tag was inserted into the vector pHYRSF53 using the restriction sites *Nco*I and *Spe*I to generate the tandem SUMO-fusion vectors. The backbone pHYRSF53 and the four vectors generated are shown in Fig. 2A. To obtain pLJSRSF3, the glutathione S-transferase (GST) coding region was amplified with the primers I399: 5′-CATCATCATCAT CACGGCTCCCCTATACTAGGTTATTG and I398: 5′-TTTACTAGTTTTTG GAGGATGGTCGCCACC using the vector pGEX-2TK (GE Healthcare) as a template and cloned into pHYRSF53 as described above. The vector pLJSRSF7 was generated in the same way, amplifying the gene of maltose binding protein (MBP) from the plasmid pTWIN-MBP1 (New England Biolabs) using the primers I397: 5′-CATCAT CATCATCACGGCAAAATCGAAGAAGGTAAAC and I395: 5′-AAACTAG TACCCGAATTAGTCTGCGCGTC. The plasmid pCARSF85 was created similarly, amplifying trigger factor (TF) directly from *E. coli* genomic DNA using the primers I09: 5′-ATCATCATCATCATCACGGT CAAGTTTCAGTTGAAACC and I08: 5′-AAACTAGTACCTCCACCCGCCT GCTGGTTCATCAGC. Finally, to generate the plasmid pCARSF63 thioredoxin (TRX) was also cloned directly from *E. coli* genomic DNA using the primers HK682: 5′-CATCATCATCATCACGGCAGCGA TAAAATTATTCACC and HK684: 5′-CCACTAGTTCCCGCCAGGT TAGCGTCGAGG, and inserted into pHYRSF53 after the second PCR reaction as described above. The plasmids pHYRSF53, pLJSRSF3, pLJSRSF7, pCARSF85 and pCARSF63 are deposited and available for academic researchers from Addgene (addgene.org) with the deposit numbers #64696, #64692, #64693, #64694 and #64695, respectively.

The construction of the plasmid pJDJRSF05 carrying SMT3 fused with the single chain *Npu*DnaE intein variant was described previously [21]. The linker region was shortened by one Gly residue compared to pJDJRSF05 by amplifying the sequence of an inactive variant of single chain *Npu*DnaE intein using the synthetic oligonucleotides HK202: 5′-GTGGATCCGGAGCTCTAAGCTATGAAACG and SK187: 5′-ATCAAGCTTAATTAGAAGCTATGAAGCC. The PCR product was digested with *Bam*HI and *Hin*dIII and cloned into the vector pHYRSF53 to generate the plasmid pJDJRSF04 (Fig. 3). Likewise, to lengthen the linker region by one Gly residue compared to pJDJRSF05, a similar approach was carried out using the synthetic oligonucleotides HK204: 5′-GTGGATCCGGAGGAGGAGCTCTAAGC TATGAAACG and SK187. The resulting plasmid was pJDJRSF06. These three constructs were used to assess the $Ulp1_{403–621}$ activity.

The expression vectors constructed allow the possibility to add another two features downstream of the protein of interest (POI): the N-term of the *Npu*DnaE split intein, and a chitin binding domain. In this work we inserted the POI into the sites *Bam*HI and *Hin*dIII to avoid modifications in the C-terminus (Fig. 2A). Four different proteins were used in this study: the C-terminal domain (CTD) of TonB from three different organisms: *E. coli* (*Ec*TonB), *Pseudomonas aeruginosa* PAO1 (*Ps*TonB), and *Helicobacter pylori* (*Hp*TonB), and the antiviral lectin scytovirin (SVN) from *Scytonema varium*. The coding sequences of TonB variants were amplified by PCR from purified genomic DNA using synthetic oligonucleotides. The fragment containing the last 89 residues of *Ec*TonB (*Ec*TonB-89) was cloned using I470: 5′-AAGGATCCGGACCACGCGCAT TAAGCCG and SK009: 5′-TACAAGCTTACTGAATTTCGGTGGTGCCG. The amplified PCR fragment was *Bam*HI-*Hin*dIII digested and inserted into the *Bam*HI-*Hin*dIII digested expression vectors pHYRSF53, pLJSRSF3, pLJSRSF7, pCARSF85, pCARSF63 to generate the plasmids pFGRSF01, pFGRSF02, pFGRSF03, pFGRSF04, pFGRSF05, respectively. For *Ps*TonB two versions with different lengths were generated: the last 77 residues of TonB (*Ps*TonB-77) and the last 96 residues (*Ps*TonB-96). To clone the last 77 residues (*Ps*TonB-77) we used the synthetic oligonucleotides HK062: 5′-A AGGATCCCGGATGGCCCAGGCGCGGCG and HK063: 5′-TACAAGCT

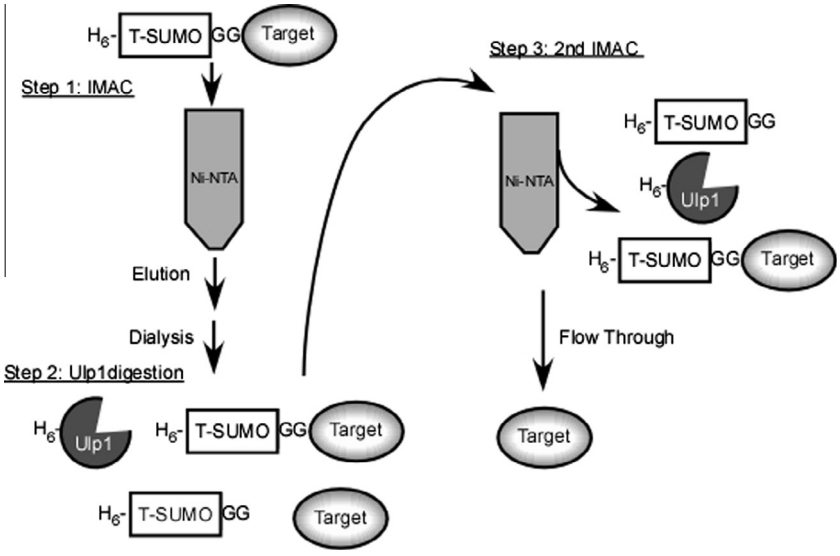

Fig. 1. Overview of the purification procedure of the target protein using the tandem SUMO vectors. T-SUMO: tandem fusion bearing yeast SMT3 as protease recognition site. Ulp1: protease domain (residues 403–621) of the *Saccharomyces cerevisiae* ubiquitin-like-specific protease 1.

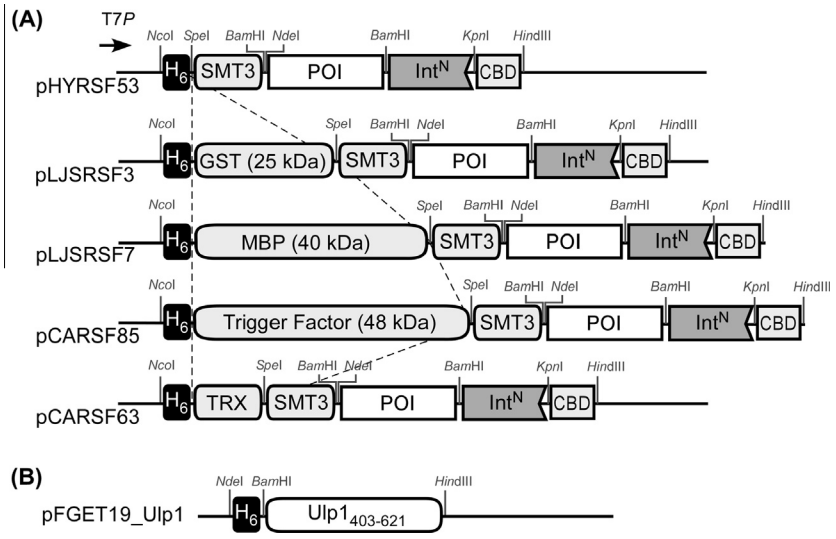

Fig. 2. Schematic representation of the vectors developed in this study. (A) Expression vectors with tandem fusion tags used in this work. The blocks represent the coding sequences, being H_6: hexahistidine tag; POI: protein of interest; SMT3: *Saccharomyces cerevisiae* SUMO homogues SMT3; Int^N: N-terminus of the *Npu*DnaE split intein; CBD: chitin binding domain; GST: glutathione S-transferase; MBP: maltose binding protein; TRX: thioredoxin. Restriction sites *Nco*I, *Spe*I, *Bam*HI, *Nde*I, *Kpn*I and *Hin*dIII are indicated. (B) Vector generated for the production of the specific SUMO protease, containing the proteolytic region (residues 403–621) of the *Saccharomyces cerevisiae* Ulp1 protease.

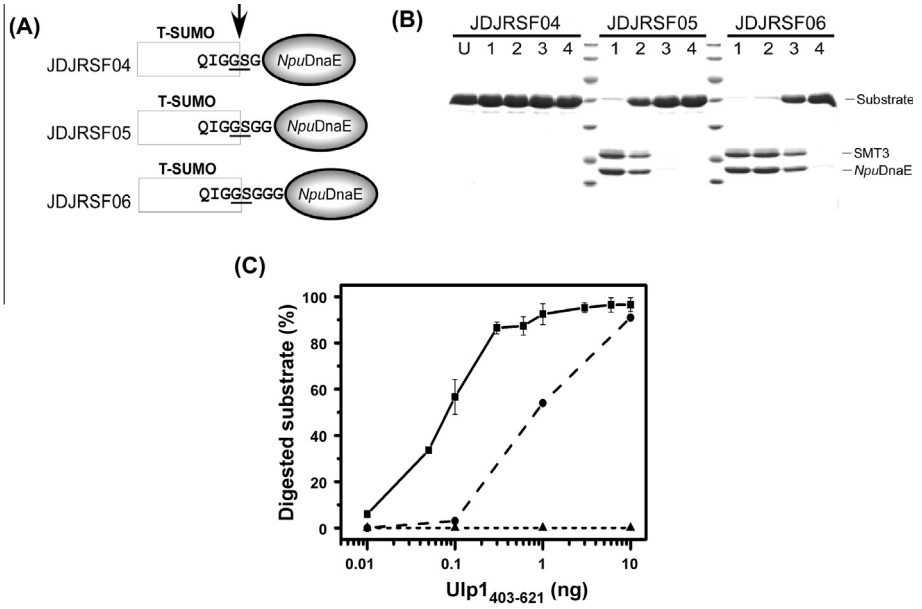

Fig. 3. Characterization of the $Ulp1_{403-621}$ activity. (A) Schematic representation of the three different linkers used in this study. The arrow represents the cleavage site. (B) 18% SDS–PAGE of the $Ulp1_{403-621}$ digestion reactions of the substrates containing the three different linkers: SG (JDJRSF04) on the left, SGG (JDJRSF05) in the center, and SGGG (JDJRSF06) on the right. Substrate was always present in a quantity of 2 μg per reaction. Lines are U: undigested (without $Ulp1_{403-621}$); 1: 10 ng of $Ulp1_{403-621}$ added to 2 μg substrate (1:200 enzyme/substrate molar ratio); 2: 1 ng of $Ulp1_{403-621}$ added (1:2000); 3: 0.1 ng of $Ulp1_{403-621}$ added (1:20,000); 4: 0.01 ng of $Ulp1_{403-621}$ added (1:200,000). (C) Graphical representation of the percentage of digested substrate against the quantity of $Ulp1_{403-621}$ added. GG/SG linker (JDJRSF04) is represented as triangles, GG/SGG (JDJRSF05) as circles, and GG/SGGG (JDJRSF06) as squares.

TAGCGGCGCTTCTCGATCTTGAAG to amplify the coding sequence of *Ps*TonB-77 from the genomic DNA. The PCR fragment was inserted into *Bam*HI-*Hin*dIII digested pCARSF63 and pLJSRSF7 to generate pFGRSF12 and pFGRSF13, respectively. A longer version containing the last 96 residues of the same protein (*Ps*TonB-96) was constructed similarly but using the two primers HK210: 5′-A CATATGGGCAGCCTCAACGACAGCG and HK063. The amplified PCR product was inserted into *Bam*HI-*Hin*dIII digested pCARSF63 and pLJSRSF7 to generate pFGRSF14 and pFGRSF15, respectively. *Hp*TonB-92 consisted of the C-terminal 92 residues from TonB. The coding sequence was amplified using the primers I471: 5′-A ACGGATCCAACGAATTTTTAATGAAGATCCAAAC and I472: 5′-TTAA AGCTTAGTCTTCTTTCAAGCTATAAGCGATAG. This PCR product was inserted between *Bam*HI and *Hin*dIII restriction sites in pHYRSF53 and pLJSRSF7 to generate the plasmids pACRSF01 and pACRSF02, respectively. For the cloning of SVN we used as a template the vector pET-32c-TRX-SVN [22] and the primers I223: 5′-AAG GATCCGGTCCGACCTACTGCTG and HK968: 5′-CTAAAGCTTACG CAGCCGCGTGACCCG. These *Bam*HI-*Hin*dIII digested PCR products were inserted into the *Bam*HI-*Hin*dIII digested expression vectors

pHYRSF53, pLJSRSF3, pLJSRSF7, pCARSF85, pCARSF63 to generate the plasmids pFGRSF07, pFGRSF08, pFGRSF09, pFGRSF10, pFGRS F11, respectively. The DNA sequences of all the constructed vectors were verified by DNA sequencing.

To overexpress and purify the C-terminal catalytic domain of Ulp1 protease (residues 403–621, $Ulp1_{403-621}$), a new plasmid with kanamycine resistance was constructed transferring from the vector pHYRS52 (deposited in Addgene, plasmid #31122) the *Xba*I-*Eco*RI Shine-Dalgarno region along with the coding sequence of the $Ulp1_{403-621}$ into pET-28b(+) (Novagen). Due to the presence of a second *Xba*I site in the vector, partial digestion was carried out to select the appropriate DNA fragment. The resulting vector was called pFGET19_Ulp1 (Fig. 2B). The plasmid pFGET19_Ulp1 is deposited and available from Addgene (addgene.org) with the deposit number #64697.

2.2. Protein expression and purification

E. coli ER2566 competent cells bearing a chromosomal T7 RNA polymerase under the control of a lac operon were transformed for the protein expression with the different expression vectors: pFGRSF01 (H_6-SMT3-*Ec*TonB-89), pFGRSF02 (H_6-GST-SMT3-*Ec*TonB-89), pFGRSF03 (H_6-MBP-SMT3-*Ec*TonB-89), pFGRSF04 (H_6-TF-SMT3-*Ec*TonB-89), pFGRSF05 (H_6-TRX-SMT3-*Ec*TonB-89), pFGRSF07 (H_6-SMT3-SVN), pFGRSF08 (H_6-GST-SMT3-SVN), pFGRSF09 (H_6-MBP-SMT3-SVN), pFGRSF10 (H_6-TF-SMT3-SVN), pFGRSF11 (H_6-TRX-SMT3-SVN), pFGRSF12 (H_6-TRX-SMT3-*Ps*TonB-77), pFGRSF13 (H_6-MBP-SMT3-*Ps*TonB-77), pFGRSF14 (H_6-TRX-SMT3-*Ps*TonB-96), pFGRSF15 (H_6-MBP-SMT3-*Ps*TonB-96), pACRSF01 (H_6-SMT3-*Hp*TonB-92), pACRSF02 (H_6-MBP-SMT3-*Hp*TonB-92), pJDJRSF04 (H_6-SMT3-SG-*Npu*DnaE), pJDJRSF05 (H_6-SMT3-SGG-*Npu*DnaE), pJDJRSF06 (H_6-SMT3-SGGG-*Npu*DnaE). For every experiment, an overnight grown cell colony from LB-agar plates was transferred to liquid LB medium and grown at 37 °C (unless otherwise stated) in the presence of 25 μg ml^{-1} kanamycin to avoid the loss of the plasmid during the growth. The volumes were 5 ml for small scale purification, and 1 l for large scale purification (2 l when using ^{15}N labelled M9 minimal medium). The cell suspension was allowed to reach an OD_{600} of ≈0.5 before the addition of isopropyl-β-D-thiogalactoside (IPTG) at 1 mM final concentration. Cells were then grown either for 4 h at 37 °C, 5 h at 30 °C, or overnight (≈16 h) at 25 °C. After the induction time, the cells were collected by centrifugation and either immediately used or stored frozen at −70 °C.

Total cell lysates were prepared by suspending the cells in B-PER reagent (Thermo scientific) and incubated at 25 °C and 1000 rpm shaking for 10 min. Soluble and insoluble fractions were separated by centrifugation. To compare soluble and insoluble protein fractions, supernatant and pellet were separated and dissolved in an equal volume of SDS-buffer. The protein content was analyzed in a Coomassie-stained 18% SDS–PAGE.

For large scale purification, whole-cell pellets were resuspended in wash buffer (50 mM sodium phosphate, 300 mM NaCl, pH 8.0). The cells were lysed either by ultrasonication or in an EmulsiFlex C5 homogenizer device for 10 min at 15,000 PSI. The cell debris was removed from the protein solution by centrifugation at 18,000*g* for 45 min. The entire amount of the supernatant was loaded on a 5 ml HP HisTrap column (Qiagen) previously equilibrated with the wash buffer. The bound protein was eluted from the Ni-NTA column with elution buffer (50 mM sodium phosphate, 300 mM NaCl, 250 mM imidazole, pH 8.0). The protein was dialyzed against standard phosphate-buffered saline (PBS) buffer. In the case of $Ulp1_{403-621}$, after this step it was stored frozen at −70 °C in storage buffer (50% glycerol v/v, 25 mM DTT) until further use. For the rest of the constructs, the dialyzed protein was digested with 0.1% $Ulp1_{403-621}$ (v/v) plus 1 mM DTT, and the digestion was loaded on a Ni-NTA spin column (Qiagen) equilibrated with wash buffer. The flow-through contained the purified tag-free target protein, which was dialyzed against 20 mM sodium phosphate buffer pH 6, and concentrated using centrifugal concentrators (5000 MWCO Vivaspin Turbo 15, Sartorius).

2.3. $Ulp1_{403-621}$ activity assay

To check the $Ulp1_{403-621}$ protease activity we performed 5 μl volume reactions containing 2 μg of substrate in PBS buffer, 1 mM DTT, and a known quantity of $Ulp1_{403-621}$. Reactions were carried out for 1 h at 30 °C, and then the whole reaction was analyzed in an 18% SDS–PAGE. We found that the purified $Ulp1_{403-621}$ protein was functional, having around 1000 U μl^{-1}. Unit definition: 1 U of $Ulp1_{403-621}$ protease is defined as the amount of enzyme needed to cleave 85% of 2 μg of control substrate (JDJRSF06, 28.5 kDa) in 1 h at 30 °C.

2.4. NMR measurements

For nuclear magnetic resonance (NMR) studies the protein samples were prepared from cultures grown in two liters of M9 medium containing $^{15}NH_4Cl$ as the sole nitrogen source, and purified as described above. NMR measurements were performed at the ^1H frequency of 600 MHz on Bruker Avance III HD spectrometer equipped with a triple resonance cryogenic probe. The [^1H,^{15}N]-heteronuclear single quantum coherence (HSQC) spectra were recorded at 298 K with ≈0.5 mM samples in 200 μl volume.

3. Results

3.1. Design of the tandem fusion vectors and the efficient purification procedure

SUMO fusion system has been used to improve overexpression and solubility of the protein of interest [2,23–25]. We decided to use yeast SMT3 protein (ubiquitin-like protein of the SUMO family) as the specific recognition domain for the proteolytic cleavage together with the N-terminal H_6-tag for convenient protein purification using Immobilized Metal Ion Affinity Chromatography (IMAC). In addition, a fragment comprising residues 403–621 of ubiquitin-like-specific protease 1 ($Ulp1_{403-621}$) with an N-terminal H_6-tag was expressed and purified to serve as a specific protease to cleave the SMT3 tag. Both, the fusion protein and $Ulp1_{403-621}$ protease contain the N-terminal H_6-tag. Therefore, undigested fusion protein, cleaved SMT3, and $Ulp1_{403-621}$ protease can be efficiently removed from the target protein in the 2nd IMAC (Fig. 1). The purification procedure of the SMT3 fusion protein/protease system is straightforward as follows: (1) IMAC purification of the fusion protein from cell lysate. (2) Dialysis of the purified fusion protein in phosphate-buffered saline (PBS); (3) Digestion by $Ulp1_{403-621}$ protease; (4) Purification of the target protein by IMAC by removing $Ulp1_{403-621}$ and H_6-SMT3 and the undigested precursor (Fig. 1).

This purification system has been used in our laboratory to purify various proteins for structural studies and also produced several high-quality crystals without further purification steps [19,26–29]. However, in several cases we found that the fusion proteins were insoluble even with SMT3 fusion partner, which has often been used as a solubility enhancer (Fig. 4A) [2]. It seems that T7 promoter in a high-copy number plasmid with RSF origin used in our plasmids [30] tends to direct the fusion proteins to insoluble fraction. For example, even though *E. coli* TonB protein has previously been expressed successfully [31], the SUMO fusion protein using our plasmid was not soluble. Thus, it was of our practical interests to create a new set of vectors for testing the solubility

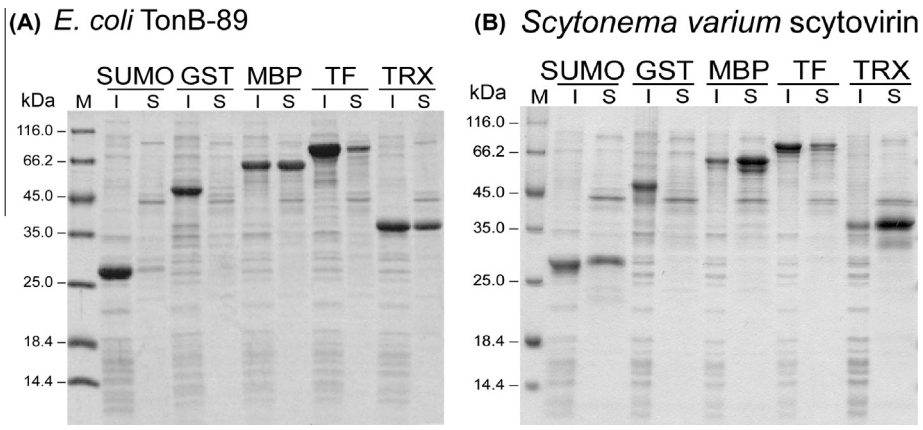

Fig. 4. (A) Coomassie stained 18% SDS–PAGE gel from the expression in *E. coli* ER2566 of the different fusion tags fused to the *Ec*TonB-89. The gel shows the separation of the insoluble fraction (I) and the soluble fraction (S). M: molecular weight marker. (B) Coomassie stained 18% SDS–PAGE gel from the expression in *E. coli* ER2566 of the different fusion tags fused to SVN. I: insoluble membrane fraction. S: soluble fraction. M: molecular weight marker.

in the same plasmid backbone by fusing several commonly used fusion tags (e.g. GST, MBP, TF and TRX) to the POI, yet utilizing the SUMO domain for the cleavage. Fig. 2A depicts five schematic vector maps of the newly constructed tandem-fusion vectors and our previous vector with SMT3 [19].

The vector has been originally designed for protein ligation by split *Npu*DnaE intein by protein trans-splicing (PTS) [19]. The split fragments required for protein ligations by PTS can be less soluble than the original target protein [26]. The newly developed tandem fusion vectors can be not only useful for protein purification (POI can be cloned between *Bam*HI and *Hin*dIII), but they can also be used for protein ligation by protein trans-splicing with enhanced solubility due to solubility enhancement tag (POI can be cloned using *Bam*HI or by restriction-free (RF) cloning) [32] (Fig. 2A).

3.2. Linker length required for Ulp1 protease digestion

The SUMO system relies on highly pure Ulp1 protease to cleave the SMT3 tag from the POI even though Ulp1 is very specific to SMT3 domain. With the newly constructed plasmid, the expression and purification of the $Ulp1_{403-621}$ catalytic domain yielded up to 87 mg of the highly pure enzyme per liter. Generally the insertion of various fusion tags (GST, MBP, TF and TRX) at the N-terminus of SUMO domain did not interfere with the activity of $Ulp1_{403-621}$ protease for cleavage of our fusion proteins because we could digest the fusion proteins equally well when the linker after the C-terminal di-glycine of SMT3 is sufficiently long (data not shown). This is presumably because the N-terminus of SMT3 is distantly located from $Ulp1_{403-621}$ as observed in the crystal structure of SMT3/Ulp1 complex [18]. Therefore, the SUMO domain could serve not only as a solubility enhancer but also as a general cleavage site. However, the linker after di-glycine peptide of the C-terminus of SMT3 domain influenced the activity of Ulp1 protease drastically (Fig. 3). We systematically analyzed the activity of $Ulp1_{403-621}$ with different lengths for the linker connecting SMT3 and the target protein. We used an inactive variant of the *Npu*DnaE intein (C1A) as a POI. We tested three different linker lengths by inserting SG, SGG or SGGG between the C-terminal di-glycine of SMT3 and the first residue of *Npu*DnaE intein (C1A). When only two residues were inserted at the front of the *Npu*DnaE intein (C1A), the fused protein could not be digested by $Ulp1_{403-621}$ at all (Fig. 3). Since the first residue of *Npu*DnaE intein is already integral part of the three-dimensional structure of *Npu*DnaE intein [33], it is thus likely that the short linker to di-glycine residue of SMT3 inhibits $Ulp1_{403-621}$ to access the cleavage site. Extending the linker length by glycine residues improved the cleavage considerably, indicating that $Ulp1_{403-621}$ requires at least three flexible residues between a structured globular domain and di-glycine peptide of SMT3 for cleaving the protein of interest from SMT3. It might require even a longer linker for larger target proteins, since *Npu*DnaE intein is a small protein (15.8 kDa).

3.3. Comparison of the five tandem SUMO fusion vectors for soluble protein expression

Even though the C-terminal domain (CTD) of TonB from *E. coli* has been previously overexpressed and purified [31,34,35], the fusion protein with SMT3 domain bearing CTD of *E. coli* TonB consisting of 89 residues (*Ec*TonB-89) was mostly insoluble using our vector of a high-copy number plasmid with T7 promoter (Fig. 4A). We compared the solubility of various fusion proteins using the tandem SUMO fusion vectors bearing an additional fusion tag: GST, MBP, TF, or TRX. The open reading frame of *Ec*TonB-89 was inserted between *Bam*HI and *Hin*dIII sites of the newly constructed tandem fusion vectors for comparison (Fig. 2A). All *Ec*TonB-89 fusion proteins were highly expressed in *E. coli*. The fusion proteins in soluble fraction were compared (Fig. 4A).

The improved solubility was particularly observed for TRX and MBP fusion tags. The solubility when using TF as a fusion tag was slightly improved. Similar results were observed when we tested the expression of scytovirin within the set of the five different tandem vectors (Fig. 4B), being MBP and TRX as the fusion partners that improved more the solubility of the expressed protein. Interestingly, in the case of scytovirin, TRX fusion improved the solubility more than *Ec*TonB-89.

*Ec*TonB-89 was successfully purified from either MBP-fusion or TRX-fusion after Ulp1 digestion (Fig. 5A). The well-dispersed NMR signals of *Ec*-TonB-89 purified from MBP-fusion in the [^{1}H,^{15}N]-HSQC spectrum indicate the well-ordered structure of *Ec*TonB-89 (Fig. 5B), which is in agreement with the previously published resonance assignments [34]. Fig. 5B shows that the majority of the chemical shift positions of *Ec*TonB-89 match with the *Ec*TonB resonance assignments published earlier ((BMRB ID 6375), suggesting that *Ec*TonB-89 is identical with that of *Ec*TonB produced without any fusion tag [34].

3.4. Self-contained domain of Pseudomonas TonB

Next, we used these vectors for the production of CTD of one of the TonB proteins from *P. aeruginosa* (*Ps*TonB), which is annotated as TonB in UniProt database with no available structural informa-

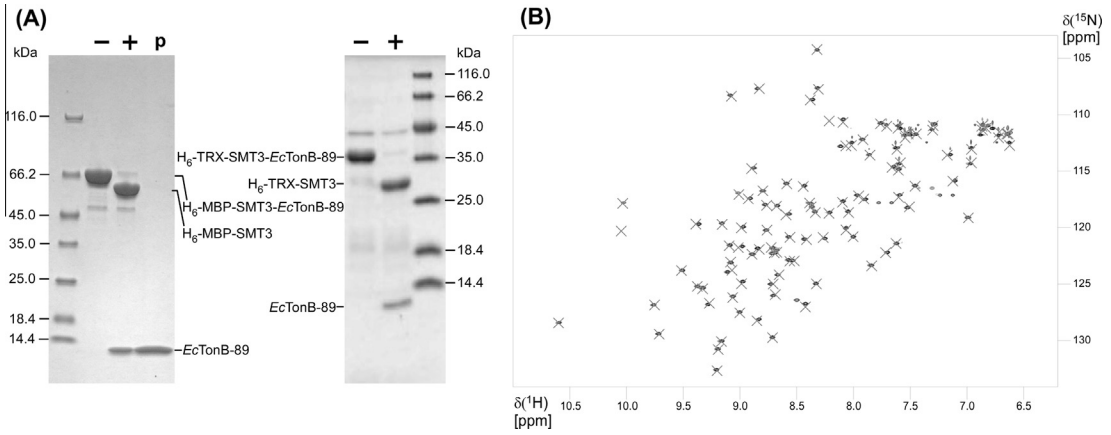

Fig. 5. Release of *E. coli* TonB-89 after digestion with Ulp1. (A) SDS–PAGE gel analysis after purification of the fusion proteins H_6-MBP-SMT3-*Ec*TonB-89 (left) and H_6-TRX-SMT3-*Ec*TonB-89 (right) before (−) and after (+) the digestion. The solubility tag has been removed by a second affinity chromatography purification. Purified *Ec*TonB-89 (p) from MBP-fusion has been used for NMR studies. (B) Two-dimensional [^{1}H,^{15}N]-HSQC spectrum of purified 0.5 mM *Ec*TonB-89 in 20 mM sodium phosphate buffer at pH 6.0 at 298 K. The spectrum was recorded at ^{1}H frequency of 600 MHz. The crosses show the chemical shift positions of the previously published resonance assignments of *Ec*TonB (BMRB ID 6375) [34].

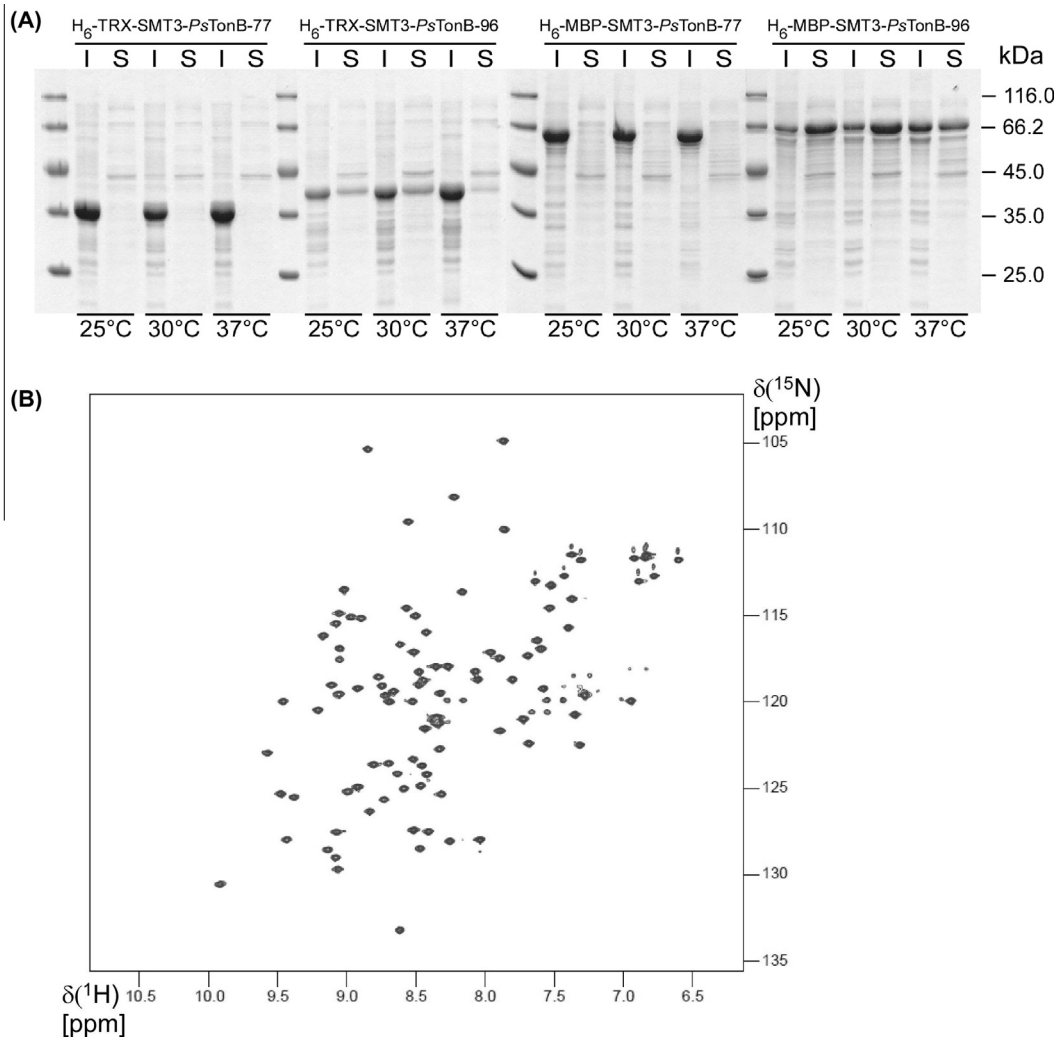

Fig. 6. Purification of *Pseudomonas* TonB-96. (A) Coomassie stained 18% SDS–PAGE gel analysis of the purification of *Ps*TonB using the short version (*Ps*TonB-77) or the long version (*Ps*TonB-96) of the POI fused to the tags H_6-TRX-SMT3 (indicated as H_6-TRX-SMT3-*Ps*TonB-77 or H_6-TRX-SMT3-*Ps*TonB-96, respectively) or to H_6-MBP-SMT3 (right, indicated as H_6-MBP-SMT3-*Ps*TonB-77 or H_6-MBP-SMT3-*Ps*TonB-96, respectively). The gel shows the separation of the insoluble fraction (I) and the soluble fraction (S) at different induction temperatures (indicated below the lanes). (B) Two-dimensional [^{1}H,^{15}N]-HSQC spectrum of purified 1 mM *Ps*TonB-96 in 20 mM sodium phosphate buffer at pH 6.0 at 298 K. The spectrum was recorded at the ^{1}H frequency of 600 MHz.

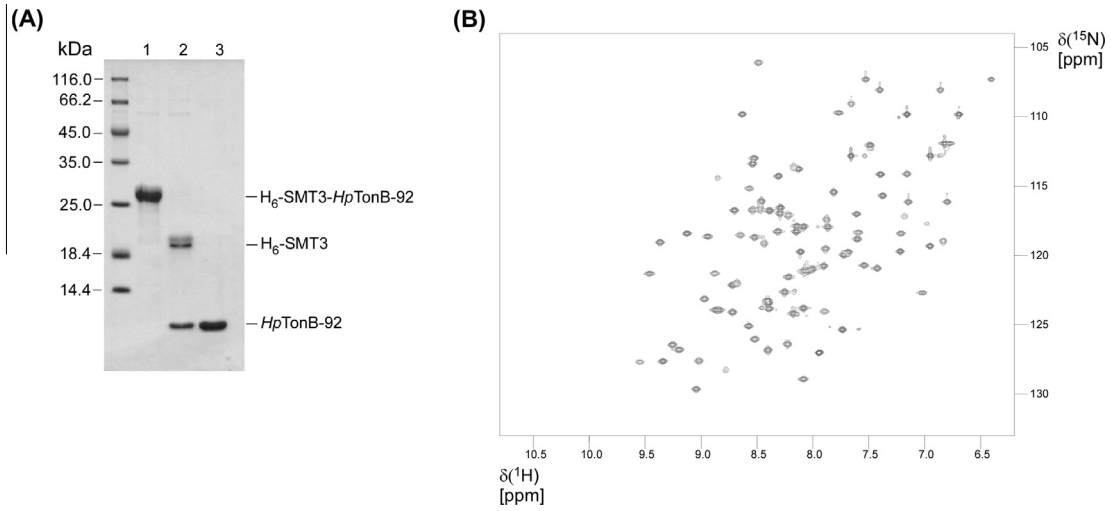

Fig. 7. (A) Coomassie stained 18% SDS–PAGE gel of $Ulp1_{403–621}$ digestion of H_6-SMT3-*Hp*TonB-92. H_6-SMT3-*Hp*TonB-92 was purified using affinity chromatography (Lane 1). After $Ulp1_{403–621}$ digestion (Lane 2), a second purification step was performed to collect *Hp*TonB-92 (lane 3). (B) [^{1}H,^{15}N]-HSQC spectrum of *Hp*TonB-92 in 20 mM potassium phosphate buffer, pH 6.0, 300 K, recorded at 600 MHz ^{1}H frequency.

tion (UniProt: Q51368). Because the boundary of the self-contained CTD of *Ps*TonB was unclear from the sequence alignments with TonB, we constructed two variants for CTD of *Ps*TonB with two different lengths, which contain either the last 77 residues (*Ps*TonB-77) or the last 96 residues (*Ps*TonB-96). We inserted these two variants into the vectors that had shown better solubility in the case of *Ec*TonB-89, which were the MBP tandem fusion (pLJRSF7) and the TRX tandem fusion (pCARSF63). *Ps*TonB-96 could be expressed with both MBP-tandem vector and TRX-SMT3 vector in soluble form and purified successfully (Fig. 6A), although the fusion with MBP yielded more soluble protein. Soluble fraction of TRX-SMT3 fusion of *Ps*TonB-96 was slightly increased at lowering the expression temperature from 37 °C to 25 °C. Whereas the well-spread NMR signals of *Ps*TonB-96 in the [^{1}H,^{15}N]-HSQC spectrum indicate a folded globular structure (Fig. 6B), it was not possible to obtain *Ps*TonB-77 in soluble fraction neither by changing the fusion partners nor by lowering the expression temperature (Fig. 6A). This result indicates that *Ps*TonB-77 probably cannot exist as a self-contained domain. It also indicates that the new tandem vectors might be only useful for expression of self-contained domains and not for partial fragments of small proteins.

3.5. CTD of TonB from H. pylori (HpTonB-92)

In the case of *Hp*TonB-92 from *H. pylori* the protein was soluble with both SMT3 tag and MBP-SMT3 tag. Although expression together with MBP-SMT3 tag was more soluble, SMT3 tag was used for the production of ^{15}N-labeled *Hp*TonB-92 due to the larger molecular weight of MBP, resulting in a better yield of the target protein (Fig. 7A). The NMR signals of *Hp*TonB-92 in the [^{1}H,^{15}N]-HSQC spectrum are well separated, indicating a globular structure of the protein (Fig. 7B).

4. Discussion

Structural investigation such as by NMR spectroscopy typically requires a large amount of highly pure protein. Due to stable isotope-labeling, it is crucial to achieve a high yield (>10 mg/l). Therefore, the use of the strong T7 promoter in high copy number plasmids might be preferable for overexpression of exogenous proteins in *E. coli* to fulfill the sample requirements. Here, we constructed a set of new fusion vectors in which SUMO domain (yeast SMT3) was used as a highly specific recognition site of Ulp1 protease together with solubility enhancement proteins (GST, TF, MBP, TRX), that are commonly used as fusion tags. None of the fusion tags in these tandem SUMO fusion proteins disturbed the cleavage at the C-terminus of SUMO domain by Ulp1 protease when at least three flexible residues were introduced between a globular domain of POI and SMT3. This confirms that SMT3 domain could be used as a highly specific recognition sequence of $Ulp1_{403–621}$ protease, which requires a very small enzyme/substrate ratio (1:10,000) for cleavage. As demonstrated with CTDs of TonB protein from three different organisms and scytovirin, an additional solubility enhancement protein in form of a fusion tag, could indeed increase the amount of the fusion protein in the soluble fraction when SMT3 alone was not sufficient for improving solubility. However, enhancement of protein solubility was not achieved for all the constructs. While none of the *Ps*TonB-77 fusion proteins were soluble, *Ps*TonB-96 was found to be soluble when expressed with MBP-SMT3 and TRX-SMT3 fusion tags (Fig. 6A). This result suggests that these fusion tags might not necessarily improve protein solubility of the target protein, when the target protein is truncated within the self-contained domain. High specificity of Ulp1 protease is highly desirable for the production of unstable proteins or unfolded protein fragments, because unfolded fragments and unstable proteins are more prone to undesired proteolytic digestion. Therefore, the tandem fusion proteins bearing SMT3 as a highly specific recognition sequence of Ulp1 protease might be more advantageous than other commercial proteolytic enzymes that are commonly used.

Moreover, the newly constructed plasmids with the tandem SUMO fusion are designed so that POI can be fused with the N-terminal fragment of a naturally occurring *Npu*DnaE split intein (Int^N), which could be subsequently used for protein ligation by protein *trans*-splicing. Because it is crucial to express the precursor protein in soluble fraction for *in vivo* protein ligation using the time-delayed dual over-expression [26], various cleavable solubility enhancement tags can be important for *in vivo* protein ligation [36]. Split intein precursors, particularly artificial split inteins, can be less soluble [28] even though naturally split inteins seems to be soluble in many cases [37,38]. Thus, the set of new plasmids using tandem SUMO fusion system could be of practical importance.

5. Conclusions

We demonstrated that tandem-SUMO fusion vectors bearing five different fusion tags could improve solubility of otherwise poorly soluble protein fragments. In order to achieve soluble protein expression, SMT3 (yeast SUMO homologue) was used as the highly specific cleavage site of Ulp1 protease, together with a solubility enhancing fusion tag. The proteolytic activity of Ulp1 protease (residue 403–621) remained undisturbed, when a three-residue linker was introduced between SMT3 and the target protein. Significant differences regarding the solubility of target proteins could be observed between the five constructs, proving that the choice of an ideal solubility tag is essential for soluble expression of otherwise poorly soluble proteins.

Acknowledgments

This work is supported by the grants from the Academy of Finland (137995) and Sigrid Juselius Foundation. A. C. acknowledges the National Doctoral Programme in Informational and Structural Biology (ISB) for financial support. The NMR facility at the Institute of Biotechnology is supported by Biocenter Finland. The authors thank C. Albert, L. Sipilä and J. Djupsjöbacka for constructing vectors and Drs. A. Wlodawer and B.R. O'Keefe for providing the plasmid containing SVN.

References

[1] K. Terpe, Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems, Appl. Microbiol. Biotechnol. 72 (2006) 211–222.
[2] J.G. Marblestone, S.C. Edavettal, Y. Lim, P. Lim, X. Zuo, T.R. Butt, Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO, Protein Sci. 15 (1) (2006) 182–189.
[3] E. LaVallie, J. McCoy, Gene fusion expression systems in *Escherichia coli*, Curr. Opin. Biotechnol. 6 (1995) 501–506.
[4] D. Esposito, D.K. Chatterjee, Enhancement of soluble protein expression through the use of fusion tags, Curr. Opin. Biotechnol. 17 (2006) 353–358.
[5] M.R. Bell, M.J. Engleka, A. Malik, J.E. Strickler, To fuse or not to fuse: what is your purpose?, Protein Sci 22 (2013) 1466–1477.
[6] D.B. Smith, K.S. Johnson, Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase, Gene 67 (1988) 31–40.
[7] E.R. Lavallie, E.A. DiBlasio, S. Kovacic, K.L. Grant, P.F. Schendel, J.M. McCoy, A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm, Biotechnology 11 (1993) 187–193.
[8] L.A. Collins-Racie, J.M. McColgan, K.L. Grant, E.A. DiBlasio-Smith, J.M. McCoy, E. R. LaVallie, Production of recombinant bovine enterokinase catalytic subunit in Escherichia coli using the novel secretory fusion partner DsbA, Biotechnology 11 (2) (1993) 187–193.
[9] R.B. Kapust, D.S. Waugh, Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused, Protein Sci. 8 (1999) 1668–1674.
[10] A. Basters, L. Ketscher, E. Deuerling, C. Arkona, J. Rademann, K. Knobeloch, G. Fritz, High yield expression of catalytically active USP18 (UBP43) using a Trigger Factor fusion system, BMC Biotechnol. 12 (2012). 56-56.
[11] C.L. Young, Z.T. Britton, A.S. Robinson, Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications, Biotechnol. J. 7 (2012) 620–634.
[12] D.S. Waugh, An overview of enzymatic reagents for the removal of affinity tags, Protein Expr. Purif. 80 (2011) 283–293.
[13] R.J. Jenny, K.G. Mann, R.L. Lundblad, A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa, Protein Expr. Purif. 31 (1) (2003) 1–11.
[14] M. Gallwitz, M. Enoksson, M. Thorpe, L. Hellman, The extended cleavage specificity of human thrombin, PLoS ONE 7 (2012).
[15] S. Chong, F.B. Mersha, D.G. Comb, M.E. Scott, D. Landry, L.M. Vence, F.B. Perler, J. Benner, R.B. Kucera, C.A. Hirvonen, J.J. Pelletier, H. Paulus, M.Q. Xu, Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element, Gene 192 (2) (1997) 271–281.
[16] S. Chong, G.E. Montello, A. Zhang, E.J. Cantor, W. Liao, M.Q. Xu, J. Benner, Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step, vol. 26, pp. 5, Nucleic acids Res. 26 (1998) 5109–5115.
[17] Y. Minato, T. Ueda, A. Machiyama, I. Shimada, H. Iwaï, Segmental isotopic labeling of a 140 kDa dimeric multi-domain protein CheA from Escherichia coli by expressed protein ligation and protein trans-splicing, J. Biomol. NMR 53 (3) (2012) 191–207.
[18] E. Mossessova, C.D. Lima, Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast, Mol. Cell 5 (2000) 865–876.
[19] M. Muona, A.S. Aranko, H. Iwai, Segmental isotopic labelling of a multidomain protein by protein ligation by protein trans-splicing, ChemBioChem 9 (2008) 2958–2961.
[20] A.S. Aranko, S. Züger, E. Buchinger, H. Iwaï, In vivo and in vitro protein ligation by naturally occurring and engineered split DnaE inteins, PLoS One 4 (2009) e5185.
[21] K. Heinämäki, J.S. Oeemig, K. Pääkkönen, J. Djupsjöbacka, H. Iwaï, NMR resonance assignment of DnaE intein from Nostoc punctiforme, Biomol. NMR Assign. 3 (1) (2009) 41–43.
[22] C. Xiong, B.R. O'Keefe, I. Botos, A. Wlodawer, J.B. McMahon, Overexpression and purification of scytovirin, a potent, novel anti-HIV protein from the cultured cyanobacterium *Scytonema varium*, Protein Expr. Purif. 46 (2) (2006) 233–239.
[23] C.D. Lee, H.C. Sun, S.M. Hu, C.F. Chiu, A. Homhuan, S.M. Liang, C.H. Leng, T.F. Wang, An improved SUMO fusion protein system for effective production of native proteins, Protein Sci. 17 (7) (2008) 1241–1248.
[24] T.R. Butt, S.C. Edavettal, J.P. Hall, M.R. Mattern, SUMO fusion technology for difficult-to-express proteins, Protein Expr. Purif. 43 (2005) 1–9.
[25] M.P. Malakhov, M.R. Mattern, O.A. Malakhova, M. Drinker, S.D. Weeks, T. Butt, SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins, J. Struct. Funct. Genomics 5 (1–2) (2004) 75–86.
[26] M. Muona, A.S. Aranko, V. Raulinaitis, H. waï, Segmental isotopic labeling of multi-domain and fusion proteins by protein trans-splicing in vivo and in vitro, Nat. Protoc. 5 (2010) 574–587.
[27] J.S. Oeemig, D. Zhou, T. Kajander, A. Wlodawer, H. Iwaï, NMR and crystal structures of the *Pyrococcus horikoshii* RadA intein guide a strategy for engineering a highly efficient and promiscuous intein, J. Mol. Biol. 421 (1) (2012) 85–99.
[28] A.S. Aranko, J.S. Oeemig, D. Zhou, T. Kajander, A. Wlodawer, H. Iwaï, Structure-based engineering and comparison of novel split inteins for protein ligation, Mol. BioSyst. 10 (5) (2014) 1023–1034.
[29] A.S. Aranko, J.S. Oeemig, T. Kajander, H. Iwaï, Intermolecular domain swapping induces intein-mediated protein alternative splicing, Nat. Chem. Biol. 9 (2013) 616–622.
[30] T. Som, J. Tomizawa, Origin of replication of *Escherichia coli* plasmid RSF 1030, Mol. Gen. Genet. 187 (3) (1982) 375–383.
[31] J. Ködding, F. Killig, P. Polzer, S.P. Howard, K. Diederichs, W. Welte, Crystal structure of a 92-residue C-terminal fragment of TonB from *Escherichia coli* reveals significant conformational changes compared to structures of smaller TonB fragments, J. Biol. Chem. 280 (4) (2005) 3022–3028.
[32] F. van den Ent, J. Löwe, RF cloning: a restriction-free method for inserting target genes into plasmids, J. Biochem. Biophys. Methods 67 (1) (2006) 67–74.
[33] J.S. Oeemig, A.S. Aranko, J. Djupsjöbacka, K. Heinämäki, H. Iwaï, Solution structure of DnaE intein from *Nostoc punctiforme*: structural basis for the design of a new split intein suitable for site-specific chemical modification, FEBS Lett. 583 (9) (2009) 1451–1456.
[34] R.S. Peacock, A.M. Weljie, S.P. Howard, F.D. Price, H.J. Vogel, The solution structure of the C-terminal domain of TonB and interaction studies with TonB box peptides, J. Mol. Biol. 345 (2005) 1185–1197.
[35] C. Chang, A. Mooser, A. Plückthun, A. Wlodawer, Crystal structure of the dimeric C-terminal domain of TonB reveals a novel fold, J. Biol. Chem. 276 (29) (2001) 27535–27540.
[36] S. Züger, H. Iwai, Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies, Nat. Biotechnol. 23 (6) (2005) 736–740.
[37] H. Wu, Z. Hu, X.-Q. Liu, Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 9226–9231.
[38] S.B. Kim, T. Ozawa, S. Watanabe, Y. Umezawa, High-throughput sensing and noninvasive imaging of protein nuclear transport by using reconstitution of split *Renilla luciferase*, Proc. Natl. Acad. Sci. U.S.A. 101 (32) (2004) 11542–11547.