Available online at www.sciencedirect.com
SCIENCE DIRECT
Protein Expression and Purification 47 (2006) 446–454
Protein Expression & Purification
www.elsevier.com/locate/yprep

An expression vector tailored for large-scale, high-throughput purification of recombinant proteins ☆

Mark I. Donnelly *, Min Zhou, Cynthia Sanville Millard [1], Shonda Clancy, Lucy Stols, William H. Eschenfeldt, Frank R. Collart, Andrzej Joachimiak

Biosciences Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received 5 October 2005, and in revised form 15 December 2005
Available online 30 January 2006

Abstract

Production of milligram quantities of numerous proteins for structural and functional studies requires an efficient purification pipeline. We found that the dual tag, his_6-tag–maltose-binding protein (MBP), intended to facilitate purification and enhance proteins' solubility, disrupted such a pipeline, requiring additional screening and purification steps. Not all proteins rendered soluble by fusion to MBP remained soluble after its proteolytic removal, and in those cases where the protein remained soluble, standard purification protocols failed to remove completely the stoichiometric amount of his_6-tagged MBP generated by proteolysis. Both liabilities were alleviated by construction of a vector that produces fusion proteins in which MBP, the his_6-tag and the target protein are separated by highly specific protease cleavage sites in the configuration MBP-site-his_6-site-protein. In vivo cleavage at the first site by co-expressed protease generated untagged MBP and his_6-tagged target protein. Proteins not truly rendered soluble by transient association with MBP precipitated, and untagged MBP was easily separated from the his-tagged target protein by conventional protocols. The second protease cleavage site allowed removal of the his_6-tag.

Keywords: High-throughput; Structural genomics; Maltose-binding protein; TVMV protease; Ligation-independent cloning

The burgeoning genomic information now available makes vast numbers of proteins accessible for structural and functional studies, and many large-scale projects have developed automated protocols for amplifying, cloning, and expressing genes, and for screening proteins for desirable properties [1–5]. Similar strides have been made in streamlining protein purification, but production of sufficient material for detailed structural and functional characterization remains labor-intensive and time-consuming [3,4,6,7]. Typically, purification is facilitated by fusing proteins to affinity tags, most commonly a his-tag, which allows purification by immobilized metal-ion affinity chromatography (IMAC, [8]). Additional tags are often attached to improve proteins' solubility, such as maltose-binding protein (MBP) [2–4,9,10]. In typical protein production pipelines, the resulting fusion proteins are first screened for solubility, then purified by semi-robotic protocols in which the tags are removed by a specific protease such as the tobacco etch virus (TEV) protease [4,6,11,12]. A second step, such as subtractive IMAC, then removes contaminating host proteins. When standard protocols of this design, as implemented by the Midwest Center for Structural Genomics (MCSG) [6], were applied to targets appended with N-terminal his_6–MBP tags, complications arose because of false positives (proteins scored as soluble

* Corresponding author. Fax: +1 630 252 7709.
E-mail address: mdonnelly@anl.gov (M.I. Donnelly).
[1] Present address. DeCode Chemistry, 2501 Davey Road, Woodridge, IL 60517, USA.

1046-5928/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.pep.2005.12.011

in screens of fusion proteins but insoluble after removal of MBP) and by failure of the secondary IMAC step to remove completely the his_6–MBP generated by TEV cleavage. Here we describe a new vector that alleviates these problems without modification of established screening and purification protocols.

Vector pMCSG19 (Fig. 1, Table 1) is derived from the simple his_6-tag–TEV-site vector, pMCSG7 [13], which has been used routinely for the production of proteins within the MCSG. The new vector applies strategies developed by Waugh and colleagues [14,15] to the problems outlined above. It encodes a leader sequence of MBP–TVMV-site–his_6-tag–TEV-site, where TVMV-site refers to the recognition sequence of tobacco vein mottling virus (TVMV) protease, another highly specific plant viral protease similar to TEV protease but with distinct specificity [16]. This configuration is distinct from the conventional arrangement used in most MBP fusions where the his-tag is not separated from MBP by a cleavage site, as occurs in vector pMCSG9 (Fig. 1B), which encodes the leader his_6–MBP–TEV-site. Expression

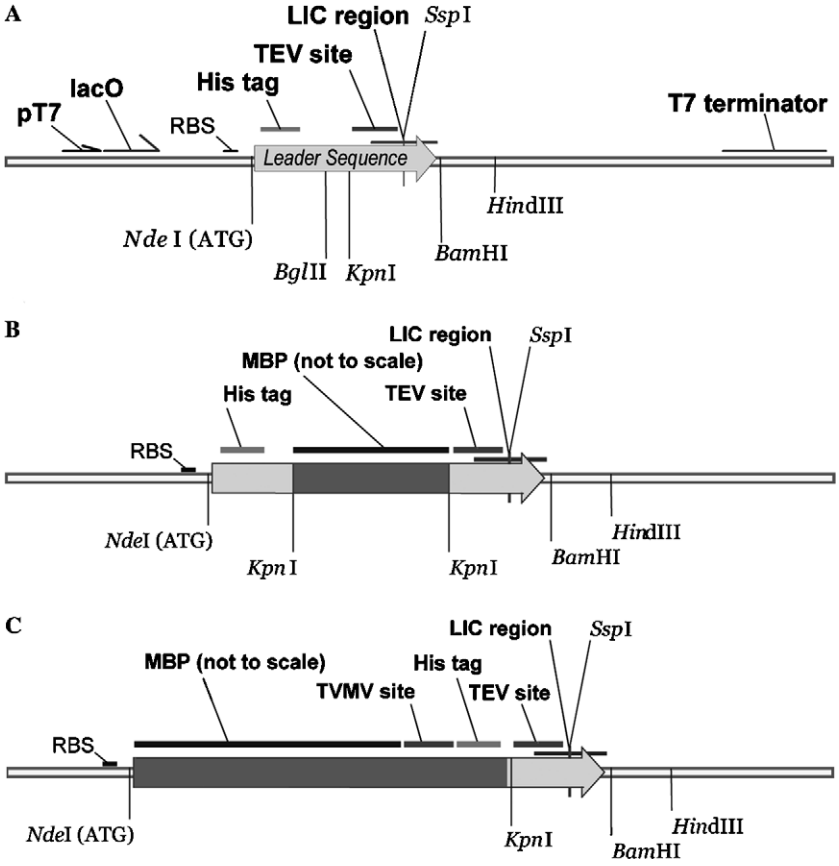

Fig. 1. Organization of pMCSG vectors. Vectors are based on the pET system of vectors (Novagen). All vectors contain the same LIC region and accept the same PCR products. (A) Expression region of pMCSG7. Following the T7 promoter, lac operator, and ribosome binding site (RBS) derived from pET21a, pMCSG7 encodes a leader sequence consisting of a his_6-tag, a spacer and the TEV protease recognition sequence followed by a LIC region based on a central *Ssp*I site [13]. Restriction sites within the leader sequence encoding region allow insertion of modules or replacement of sequences, indicated by the darker regions in B and C (see Materials and methods). (B) Leader sequence encoding region of pMCSG9. Expression of genes cloned into the LIC site generates his_6–MBP-target fusion proteins in which his_6–MBP can be released by TEV protease cleavage. (C) Leader sequence region of pMCSG19. Expression of cloned genes produces N-terminal fusions of target proteins in which MBP is separated from the his_6-tag by a TVMV protease recognition sequence. TVMV protease cleavage releases the his_6-tagged target protein, and TEV protease cleavage gives the untagged target protein.

Table 1
Properties of vectors

Vector	Encoded leader[a]	Leader MW[b] (Da)	Plasmid size (bp)
pMCSG7	his_6–TEV	2755	5286
pMCSG9	his_6–MBP–TEV	43713	6147
pMCSG19	MBP–TVMV–his_6–TEV	45050/2711[c]	6441

[a] *Abbreviations:* TEV, tobacco etch virus protease recognition sequence; MBP, maltose-binding protein; TVMV, tobacco vein mottling virus protease recognition sequence.

[b] Molecular weight of leader sequence appended to target proteins introduced into vectors by LIC. After cleavage with TEV protease, the residues SNA of the leader remain attached to the protein's N-terminus (MW = 289).

[c] First number refers to MW of entire leader, second to that remaining after cleavage by TVMV protease.

of proteins from pMCSG19 with co-expression of TVMV protease resulted in efficient removal of MBP and eliminated false positives that occurred from pMCSG9. Separation of the his_6-tag from MBP allowed established robotic purification protocols to purify the soluble proteins successfully without modification or addition of steps.

Materials and methods

Construction of pMCSG9 and pMCSG19

The vector pMCSG9 was constructed by inserting the gene encoding MBP into the *Kpn*I site of vector pMCSG7 [13]. The MBP encoding region was generated by PCR using plasmid pRK793 [17] as template (a generous gift from David Waugh) and the primers 5′-TTTTAGATCTG ATGTCCCCTATACTAGGTTATTGG and 5′-TTTTGG TACCTGGGATATCGTAATCATCCGATTTTGGAG GATGGT (purchased from the Howard Hughes Medical Institute-Keck Laboratory of Yale University, New Haven, CT). The vector was digested with *Kpn*I and dephosphorylated with calf intestinal phosphatase (Promega, Madison, WI), and ligated to the *Kpn*I-treated PCR product. The resulting plasmids were screened for orientation and the expression region of a positive candidate was sequenced to verify that the sequence of MBP matched that encoded by pRK793. Vector pMCSG19 was constructed by replacing the region encoding the his_6-tag in pMCSG7 (between *Nde*I and *Bgl*II, Fig. 1A) with a sequence encoding MBP–TVMV-site–his_6-tag. The MBP–TVMV-site portion of this region was amplified from vector pRK1035 [15] (Science Reagents, Inc.) by PCR using Platinum Pfx polymerase (Invitrogen) and the primers TTAAACATATGAAAATC GAAGAAGG and TTATAGGATCCACGCCAGAA GAGTGATGATGATGGTG (encoding the his_6-tag on its complement) in 2× strength reaction buffer with 1 mM Mg^{2+} for 25 cycles. The PCR product was cleaved with *Nde*I and *Bam*HI and ligated into pMCSG7 which had been treated with *Nde*I and *Bgl*II followed by calf intestinal phosphatase and gel purification. The amplified and flanking sequences of the resulting construct were verified by DNA sequencing.

Ligation-independent cloning into pMCSG vectors

Vectors were prepared for LIC by cleavage with *Ssp*I endonuclease, purification by agarose gel electrophoresis, and treatment with T4 DNA polymerase in the presence of dGTP. Fifteen micrograms of vector DNA, purified with a Qiagen Plasmid Midi kit (Qiagen, Valencia, CA), were incubated with 75 U high concentration *Ssp*I (New England Biolabs) at 37 °C for 2 h in a reaction volume of 60 μl, then purified following agarose gel electrophoresis using a QiaEx II gel extraction kit. The material was then treated with 40 U LIC-qualified T4 DNA polymerase (Novagen, Madison, WI) and 2.5 mM dGTP in a volume of 40 μl in 1× commercial buffer supplemented with 5 mM DTT. Genes were amplified by PCR with primers encoding the LIC overhang [13] (sense: TACTTCCAATCCAATGCX followed by the genes' N-terminal sequences; antisense: TTATCCACTTCCAATG followed by the complement of a stop codon and of the C-terminus of the gene), purified with a QIAQuick PCR purification kit (Qiagen), and treated with T4 polymerase as described above except in the presence of dCTP. Following annealing of 30–50 ng of this material with 15 ng LIC-prepared vector, the resulting plasmids were transformed into DH5α, and plasmids prepared from these transformants were introduced into BL21(DE3) containing the plasmid pRK1037 [15] (Science Reagents, Inc.). Transformants were isolated on LB plates containing 100 μg/ml ampicillin and 30 μg/ml kanamycin.

Expression and analysis of solubility

Cultures were grown at 37° in LB containing ampicillin and kanamycin (100 μg/μl and 30 μg/μl, respectively) to an OD_{600} of 0.5 at which time the temperature was dropped to 20 °C and protein synthesis was induced by addition of 1 mM IPTG. Cells were harvested the next morning, suspended in 0.1 M Tris/HCl, pH 8.0, incubated with lysozyme and DNase (rLysonase and Benzonase, respectively, Promega) for 30 min at room temperature, frozen briefly, then sonicated. Following centrifugation at 6000*g* for 15 min, the soluble and insoluble fractions were analyzed for protein by denaturing gel electrophoresis.

Production of selenomethionyl proteins in non-sterile enriched minimal medium in 2-liter plastic bottles

Selenomethionyl proteins were produced in BL21(DE3)—a strain not auxotrophic for methionine—using feedback inhibition of methionine biosynthesis [18,19]. Cultures were grown in 2-liter polyethylene terephthalate beverage bottles [20,21] containing one liter of non-sterile M9 salts supplemented with glucose, glycerol, amino acids, trace metals and vitamins to increase the cell yield [22–25]. Amendments were, per liter: glycerol, 5 g; glucose, 4.4 g; non-inhibitory amino acids (L-glutamate, L-aspartate, L-arginine, L-histidine, L-alanine, L-proline, L-glycine, L-serine, L-glutamine, L-asparagine, and L-tryptophan), 200 mg each; trace metal mixture (EDTA, 5 mg; $MgCl \cdot 6H_2O$, 430 mg; $MnSO_4 \cdot H_2O$, 5 mg; NaCl, 10 mg; $FeSO_4 \cdot 7H_2O$, 1 mg; $Co(NO_3)_2 \cdot 6H_2O$, 1 mg; $CaCl_2$, 11 mg; $ZnSO_4 \cdot 7H_2O$, 1 mg; $CuSO_4 \cdot 5H_2O$, 0.1 mg; $AlK(SO_4)_2$, 0.1 mg; H_3BO_3, 0.1 mg; $Na_2MoO_4 \cdot 2H_2O$, 0.1 mg; Na_2SeO_3, 0.01 mg; $Na_2WO_4 \cdot 2H_2O$, 0.1 mg; $NiCl_2 \cdot 6H_2O$, 0.2 mg); ampicillin, 50 mg; kanamycin, 30 mg; thiamine 1 μg; and vitamin B12, 2.7 μg. Media components other than glycerol were supplied as aliquots of mixed solids in foil packets or as concentrated stock solutions by Medicillin, Chicago, IL (catalog numbers MD045004A, MD045004B, MD045004C, and MD045004E). Cultures were grown at 37 °C to an $OD_{600} = 1 - 2$, when inhibitory amino acids (25 mg each of L-valine, L-isoleucine, L-leucine, L-lysine,

L-threonine, L-phenylalanine, and 15 mg of selenomethionine; Medicillin, Catalog No. MD045004D) and 1 mM isopropylthio-β-D-galactoside (IPTG) were added, and the temperature dropped to 20 °C. Cultures were incubated overnight, harvested the next morning, suspended in lysis buffer (50 mM Hepes, pH 7.8, containing 500 mM NaCl, 10 mM imidazole, 10 mM β-mercaptoethanol, and 5% glycerol), and lysed by sonication. Proteins were purified by established protocols [6].

Results

Salvaging poorly soluble proteins through MBP fusions—pMCSG9

Insertion of the gene encoding MBP, amplified by PCR from the vector pRK739 [17], into the leader sequence encoding region of pMCSG7 gave pMCSG9 (Fig. 1, Materials and methods). Resulting plasmids were screened for orientation and expression of a protein of the expected molecular weight of his-tagged MBP (the product of the vector before introduction of a target gene), and the sequence of the MBP gene and surrounding expression region was verified by DNA sequencing. During restriction analysis, we also discovered that a portion of the vector near the Ap^R gene was larger than anticipated, both in pMCSG9 and its parent, pMCSG7. Sequencing of this region revealed that a mutation in one of the *Ssp*I sites that were removed from the parent of pMCSG7, pET21a, during its construction resulted in retention of 129 additional bases of the parental vector. The mutation and retained bases appear not to affect expression of cloned genes: over 2000 proteins have been produced in good yield from pMCSG7, leading to the deposition of over 200 structures in the Protein Data Bank. Vector sequences are available at http://www.bio.anl.gov/terrestrialr/microbiology1.html.

As expected [9,10], fusion to MBP effectively enhanced the solubility of poorly soluble bacterial proteins. PCR products encoding 134 *S. typhimurium* proteins that were poorly soluble when produced from pMCSG7 were introduced into pMCSG9 and reevaluated (Table 2). These proteins were originally scored as poorly soluble (Solubility Score 1) when screened by robotic protocols [26]. Proteins in this category are visible on gels but only at an abundance similar to host proteins, and normally are not carried forward to purification in the structure determination pipeline. Fusion to MBP effectively redistributed these proteins in the spectrum of solubility scores, some appearing to become less soluble, but more improving in solubility (Table 2). Sixty-four of the proteins (46%) were improved to Solubility Score 2 or 3 (soluble or highly soluble, respectively) by fusion to MBP. Proteins of Solubility Score 2 are clearly visible on gels in amounts greater than host proteins, and those of Solubility Score 3 are abundant, at far higher amounts than host proteins. Proteins in these categories routinely proceed to purification. These results substantiate, with a large data set, the anticipated effectiveness of MBP in salvaging poorly soluble proteins and allowing them to reenter the purification pipeline. However, we found that proteins produced from pMCSG9 failed to give target protein of sufficient purity to proceed to crystallization trials after purification by semi-automated protocols that were highly effective for soluble his_6-tagged proteins [6]. In general, a minimum of 10 mg of protein of at least 95% purity is required. None of 38 poorly soluble proteins that were made soluble by fusion to MBP satisfied these criteria, either due to precipitation after removal of MBP or failure of the second subtractive IMAC to remove completely the stoichiometric amount of his_6-tagged MBP generated by TEV protease cleavage.

Table 2
Effect of fusion to MBP on the solubility of 134 proteins in robotic screening

Vector	Solubility Score[a]			
	0	1	2	3
pMCSG7	0	134	0	0
pMCSG9	28	44	36	26

[a] Solubility assessment was based on visual inspection gels of the soluble fraction of cell extracts. Solubility Scores are: 0, insoluble; 1, poorly soluble; 2, moderately soluble; and 3, highly soluble. Proteins in category 0 were not detected on gels of soluble fractions of cell lysates. Those in category 1 were present in amounts less than major host proteins. Category 2 proteins were more abundant than any host protein, and category 3 proteins dominated protein expression. Of 134 proteins that were poorly soluble when produced from pMCSG7 (his_6-tag–TEV-site leader), 62 (46%) were improved to Solubility Score 2 or 3 when produced from pMCSG9 with the leader his_6-tag–MBP–TEV-site.

In vivo cleavage to release untagged MBP from fusion proteins—pMCSG19

Rather than adapt screening and purification protocols to accommodate these limitations of MBP fusion proteins produced from pMCSG9, we modified the expression vector to bypass them. Strategies developed by Waugh and colleagues [14,15] were adapted to design pMCSG19 (Fig. 1, Table 1). To construct the vector, we replaced the region of pMCSG7 encoding the N-terminal his_6-tag with a region encoding MBP, a protease site, and a his_6-tag (Materials and methods). Sequencing of the resulting construct verified the sequence of the amplified fragment and surrounding components of the expression region. Expression of genes introduced into pMCSG19 by LIC generates target proteins fused to an N-terminal leader of untagged MBP followed, in order, by a TVMV protease recognition sequence, a his-tag, and a TEV protease recognition sequence. Cleavage of these proteins with TVMV protease generates untagged MBP and a target protein with a his_6–TEV-site leader identical to that produced from pMCSG7 except with an N-terminal serine instead of methionine preceding the his_6-tag. If produced in cells co-expressing TVMV protease, cleavage will occur in vivo [15].

Sixteen proteins, picked at random from the original set of 38, whose solubility was improved by fusion to MBP but

which failed to give pure protein after standard purification (Table 3), were used to evaluate pMCSG19. The available PCR products encoding the proteins were introduced into pMCSG19 by LIC and transformed into BL21(DE3) cells containing vector pRK1037[15]. This plasmid encodes TVMV protease under control of the P_L-tetO promoter. In hosts that do not produce the Tet repressor, such as BL21(DE3), the plasmid produces TVMV protease constitutively, and proteins produced from pMCSG19 are cleaved at the TVMV site in vivo. Following induction, cells were lysed, fractionated by centrifugation, and the soluble and insoluble fractions analyzed by polyacrylamide gel electrophoresis (Fig. 2).

The predominant protein present in all lanes of the soluble fraction (Fig. 2A) is MBP released by TVMV cleavage. The other predominant band below the MBP band in some

Table 3
Proteins expressed in pMCSG19

Lane[a]	APC number[b]	Source	Assignment	MW (Da)
1	22819	*B. cereus*	Hypothetical	13,279
2	22808	*B. cereus*	Hypothetical	10,039
3	23402	*S. typhimurium*	Cytoplasmic protein[c]	13,639
4	23431	*S. typhimurium*	Regulatory protein	25,524
5	22906	*S. typhimurium*	RNA ligase	19,633
6	23852	*S. typhimurium*	Cytoplasmic protein[c]	13,313
7	24034	*S. typhimurium*	Inner membrane protein[c]	19,687
8	24177	*S. typhimurium*	Inner membrane protein[c]	32,650
9	24238	*S. typhimurium*	Cytoplasmic protein[c]	17,470
10	24253	*S. typhimurium*	Hydrophilic protein	24,845
11	25385	*S. typhimurium*	Regulatory protein[c]	10,880
12	25420	*S. typhimurium*	SAM methyltransferase[c]	27,127
13	25436	*S. typhimurium*	Galactitol enzyme IIA	16,967
14	25439	*S. typhimurium*	Transport protein[c]	18,542
15	23650	*Staphylococcus aureus*	Hypothetical	20,083
16	23645	*Staphylococcus aureus*	Urease accessory protein	22,345

[a] Lane number in Fig. 2.
[b] Protein identification number: APC (Accelerated Protein Crystallography) number. Details available at http://www.mscg.anl.gov.
[c] Putative assignment to general class of protein.

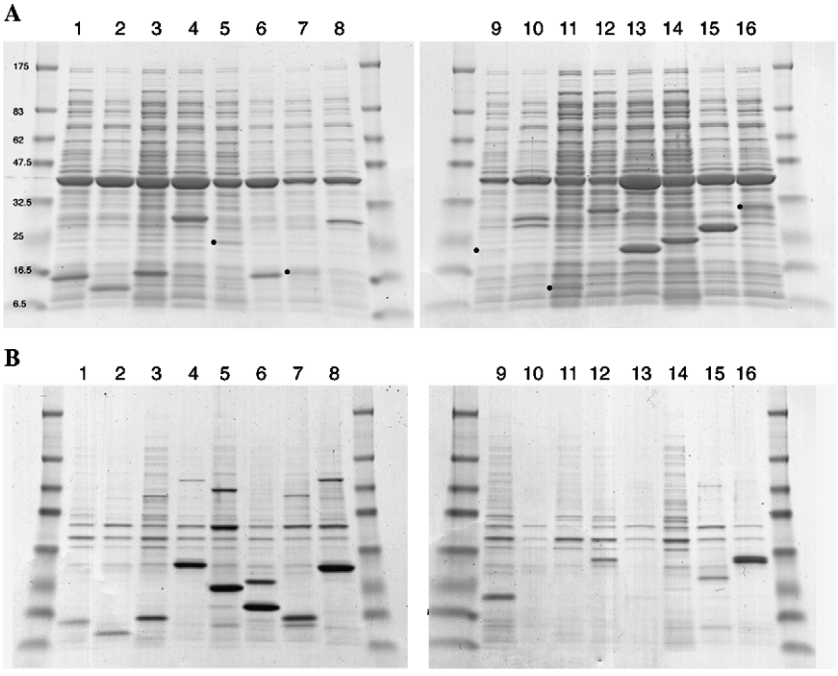

Fig. 2. Elimination of false positives by in vivo cleavage by TVMV protease. Soluble (A) and insoluble (B) fractions prepared from cells producing proteins from pMCSG19 in the presence of pRK1037. Genes encoding 16 poorly soluble proteins that were rendered soluble by fusion to MBP were introduced into pMCSG19 and evaluated for the production of soluble target proteins after in vivo removal of MBP by co-expressed TVMV protease. The predominant band in all lanes of the soluble fraction (A) is MBP released by in vivo cleavage. Where present, intense bands below MBP are soluble target proteins. Black dots indicate low abundance target proteins. Proteins in lanes 1–4 and 12–15 proceeded to large-scale production and purification. Prestained molecular weight markers (unlabeled lanes) are 175, 83, 62, 47.5, 32.5, 25, 16.5, and 6.5 kDa (Promega, Madison, WI). The identity of proteins 1–16 is given in Table 3.

of the lanes is target protein that remained soluble after cleavage from MBP. Analysis of the insoluble fraction (Fig. 2B) confirmed the expression of the target protein in those cases where little or no soluble target was observed in the soluble fraction. Based only on abundance in the soluble fraction, 11 of the 16 proteins (lanes 1–4, 6, 8, 10, and 12–15) would be scored as sufficiently soluble to proceed to purification. However, the large amount of insoluble protein for the targets in lanes 6 and 8 (Fig. 2B) would disqualify them, as would the doublet for the soluble target protein in lane 10. The remaining eight proteins were produced on a large scale and purified (see below). In summary, only half of these target proteins were deemed satisfactory for purification; the others represent false positives that were previously scored as satisfactory based on analysis of the his_6–MBP fusion proteins generated from pMCSG9. The MBP–his_6-target fusion proteins generated by pMCSG19 were also soluble (data not shown). Thus, in vivo cleavage of proteins produced from pMCSG19 effectively eliminated false positives without addition of a protease cleavage step to the screening protocols.

In those cases where fusion to MBP truly improved a protein's solubility, the his_6–MBP leader attached by production from pMCSG9 compromised purification by protocols that were highly effective for simple his_6-tagged proteins. Proteolytic cleavage with TEV generated a stoichiometric amount of his_6–MBP, which consistently failed to bind well to the second, subtractive IMAC of standard protocols, as illustrated in the purification of a representative target protein (Fig. 3A). For proteins produced from pMCSG9, the first IMAC yields the partially purified his_6–MBP-target fusion protein (Fig. 3A, lanes 1–4). Hydrolysis of this protein by TEV protease generates his_6–MBP (larger protein in Fig. 3A, lane 5) and the target protein (lower band). When this material is passed through the subtractive IMAC column, his-tagged MBP fails to bind well under the standard conditions, resulting in severe contamination of the target protein in the final eluted fraction (Fig. 3A, lanes 5–7). In contrast, when proteins are produced from pMCSG19 in a host expressing TVMV protease (Fig. 3B), MBP is cleaved away from the target protein in vivo (Fig. 3B, lane 1). Because of the design of the vector, this MBP is not his-tagged, and passes through the first IMAC column (Fig. 3B, lane 2). The target protein, which is directly his-tagged, is retained, and elutes in partially purified form (Fig. 3B, lane 4). Hydroylsis of this protein with TEV protease and subtractive IMAC chromatography generates target protein of sufficient purity to initiate crystallization trials (Fig. 3B, lanes 5–7). Production of proteins from pMCSG19 with in vivo co-expression of TVMV protease thereby resolved the purification problem without modification of protocols or addition of tertiary steps.

Production of selenomethionyl proteins from pMCSG19

Selenomethionyl derivatives of soluble proteins were produced by culturing cells in 2-liter polyethylene terephthalate beverage bottles [20] in 1 liter of non-sterile M9 salts supplemented with additional nutrients (Materials and methods). The medium and conditions were identical to those described previously [21] but with the addition of glycerol, non-inhibitory amino acids, trace metals and vitamins [18,22–25] to improve the yield of cells. Under these conditions, cell yields were two- to three-fold higher compared to those attained in unsupplemented medium, typically generating OD_{600} values of 4–12, depending on the protein expressed, with no detriment to expression or in vivo cleavage. Fig. 4 shows the partial purification of three proteins through the first IMAC step and is representative of the strong expression, efficient cleavage by co-expressed TVMV protease, and complete removal of MBP typically obtained. The three proteins were further purified by subtractive IMAC to remove trace *Escherichia coli* proteins [6], and analyzed for selenomethionine incorporation. Amino acid analysis failed to detect methionine in any of the three proteins, consistent with selenomethionine

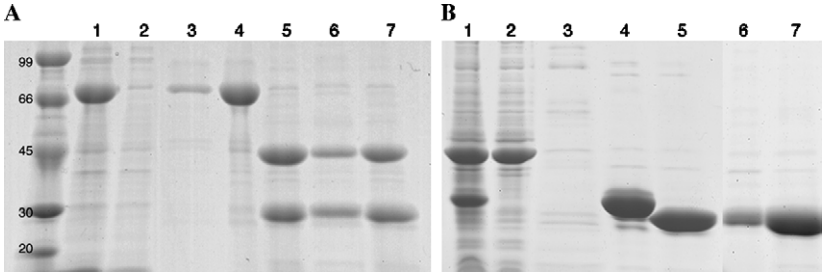

Fig. 3. Purification of target proteins produced from pMCSG9 and pMCSG19. A representative purification by standardized purification protocols [6] of APC25420 produced from (A) pMCSG9, and (B) pMCSG19. In both cases, numbered lanes contain (1) cell extract, (2–4) first IMAC flow-through, wash, and eluate, respectively, (5) TEV protease treated eluate, (6–7) subtractive IMAC flow-through and wash, respectively. For proteins produced from pMCSG9 (A), the first IMAC yields partially purified his_6–MBP-target fusion protein (lane 4). Cleavage with TEV protease generates untagged target protein (the smaller protein in this example, lane 5), and his_6–MBP. Under standard protocols, his_6–MBP fails to bind well to the subtractive IMAC and elutes with the target protein (lane 7). When proteins are produced from pMCSG19 in the presence of TVMV protease (B), untagged MBP generated by proteolysis (in this example, the larger protein in lane 1) passes through the first IMAC column unretarded (lane 2), and the partially purified target protein is free of MBP (lane 4). Following removal of the his_6-tag by TEV protease cleavage (lane 5) subtractive IMAC yields protein of sufficient purity for crystallization trials (lane 7). Molecular weight standards (unlabeled lane) are 99, 66, 45, 30, and 20 kDa, (Amersham Biosciences). Gels were stained with (A) SimplyBlue SafeStain (Invitrogen) or (B) Coomassie brilliant blue R.

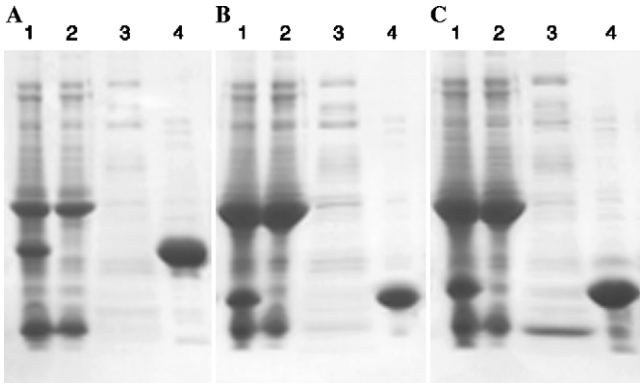

Fig. 4. Expression and partial purification of selenomethionyl proteins. Partial purification by Ni-IMAC of selenomethionyl proteins produced from pMCSG19 in enriched medium supplemented with selenomethionine (see Materials and methods) with in vivo cleavage by TVMV protease produced from plasmid pRK1037. (A–C) Correspond to proteins 12–14 of Fig. 2, which are APC25420, APC25436, and APC25439. In each panel, the lanes contain: (1) total soluble protein extract (loaded onto IMAC column), (2) unbound material that passed through the column, (3) column wash, and (4) eluted proteins. Subsequent cleavage by TEV protease followed by subtractive IMAC generated protein of >95% purity. Samples of these more highly purified proteins were analyzed for selenomethionine incorporation.

incorporation of 90% or more [21]. Yields per liter of purified target proteins produced in the amended medium were consistently two-fold or more higher than those obtained from unamended medium, allowing production of sufficient protein for crystallization trials from a single bottle, halving the number of cultures required to produce each protein. The eight soluble target proteins were produced on a large scale as their selenomethionyl derivatives and purified. Six generated greater than 10 mg of protein of greater than 95% purity (average yield for these 6 was 39 mg per liter of culture, range 14–84 mg). These were carried forward to crystallization trials; two gave crystals, and one was solved.

Discussion

Because many studies support the effectiveness of fusion to MBP for improving the solubility of proteins [2–4,9,10], we assessed the potential of using MBP fusions as a salvage pathway for poorly soluble proteins in a high-throughput protein production pipeline. A large set of *S. typhimurium* proteins that were poorly soluble when expressed with a simple his_6-tag was re-evaluated as MBP fusions (Table 2). Fusion to MBP in effect redistributed the proteins in the spectrum of solubility, generating a spread from completely insoluble to highly soluble fusion proteins, but with a strong bias toward improved solubility. Of the 134 proteins, only 28 (20%) became less soluble as MBP fusions, whereas the solubility of 62 (45%) improved upon fusion to MBP. Twenty-six proteins (19%) were deemed highly soluble and 36 (26%), moderately soluble based on the intensity of bands in protein gels of the soluble fraction of lysates. Purification of proteins in these two categories by standard protocols is considered likely to produce greater than 10 mg of pure protein (the amount needed for crystallization screening trials) and they, therefore, are passed forward into the labor-intensive large-scale purification phase of the pipeline. The results obtained from these 134 proteins justify incorporation of MBP fusions into high-throughput screening and purification pipelines for salvaging poorly soluble proteins.

The apparent solubility of proteins attached to MBP, however, was often transitory; in these cases, proteolytic removal of MBP resulted in aggregation or precipitation of the target protein. The large, highly soluble MBP protein apparently allowed intrinsically insoluble proteins to partition into the soluble fraction as long as they remain linked to MBP. If carried forward to purification protocols designed for truly soluble proteins, such proteins will fail to give enough pure material for structural or functional characterization, resulting in considerable wasted time and effort. Evaluation of solubility after proteolysis could detect false positives of this sort, but would require an additional step in screening protocols. For target proteins expressed from pMCSG19, in vivo removal of MBP by co-expressed TVMV protease effectively eliminated these false positives (Fig. 2). Of sixteen proteins, all of which were soluble as MBP fusions, only about half remained soluble after in vivo removal of MBP (Fig. 2A). The intensity of the target protein band (seen below the predominant band of MBP in all lanes) suggested eleven were rendered sufficiently soluble by transitory fusion to MBP (Solubility Score 2 or 3, see Table 2), but analysis of the insoluble fraction (Fig. 2B) provided additional insight into the suitability of the proteins for purification. The large amount of insoluble target protein in lanes 6 and 8 suggested these proteins would be more prone to aggregation and precipitation during processing, causing them to be rejected, as was the target in lane 10 because it produced a doublet. Purification of the remaining eight proteins supported the utility of evaluating the insoluble fraction as well as the soluble. Six of these proteins yielded enough pure protein to pass on to crystallization trials, but those that failed had partitioned evenly between the soluble and insoluble fractions in the solubility analysis (lanes 3 and 4 of Fig. 2). Of the six that entered crystallization trials, two crystallized, and one was solved. Whereas this data set is clearly too small to measure the effectiveness of pMSCG19 precisely, it strongly supports the potential utility of the vector in a salvage pathway for poorly soluble target proteins.

A second complication arose during purification of his_6–MBP-tagged proteins derived from pMCSG9. The stoichiometric amount of his_6–MBP produced by TEV protease cleavage of the fusion proteins was not removed effectively by the second, subtractive IMAC column of standard purification protocols, resulting in contamination of the final product with his_6–MBP (Fig. 3). Altered or additional purification steps, such as gradient elution, use of larger columns, or subsequent ion-exchange or gel-filtration steps, successfully separated target proteins from his-tagged

MBP, but each approach seriously disrupted the general laboratory workflow and increased the effort required to purify the individual proteins. Vector pMCSG19 eliminates his-tagged MBP by placing a second highly specific protease cleavage site, the TVMV-site, between an N-terminal, untagged MBP and the his-tag, which is followed by the standard TEV protease cleavage site (Fig. 1). Combined with in vivo cleavage at the TVMV site by co-expressed TVMV protease, pMCSG19 generates untagged MBP plus a simple his_6-tagged target protein identical to that produced from pMCSG7 except for the presence of an N-terminal serine instead of methionine. During standard purification, the untagged MBP passes through the initial IMAC column unretarded. The stoichiometric 'contaminant' is thus removed efficiently, and subsequent steps of the purification are identical to those used for simple his_6-tagged targets produced from pMCSG7.

Use of pMCSG19 and in vivo cleavage with TVMV protease provides a satisfactory system for producing most target proteins via a very streamlined screening and purification pipeline. The validation experiments described here were limited to small (<40 kDa) bacterial proteins, but are sufficiently compelling to justify the use of pMCSG19 in standard production protocols. Accordingly, we have initiated routine salvaging of poorly soluble proteins by reprocessing in pMCSG19. The approach represents an additional application of in vivo cleavage by highly specific proteases [14,15], and the strategy of incorporating a second high-specificity protease cleavage site to remove an untagged chaperone component in vivo could be of value in other protein production pipelines. Time and effort that might have been spent pursuing false positives or performing additional screening or purification steps can be spent on other crucial tasks. The process is also fully compatible with production of selenomethionyl derivatives for crystallography. High yields of target proteins with very efficient incorporation of selenomethionine were obtained in enriched defined medium using a non-auxotrophic host and commercially available, premixed medium components. For the six proteins successfully purified, the average yield was 39 mg per liter of culture. Such high-yields and incorporation can be obtained as well in similar production media employing autoinduction and incorporating isotopically labeled amino acids for nuclear magnetic resonance experiments [23–25]. In addition to the application described here, highly enriched defined media and vectors of the general configuration, tag1-site1-tag2-site2-protein, could be exploited for other purposes, including functional analyses of proteins.

Acknowledgments

We thank David Waugh for providing plasmids pRK793, pRK1035, and pRK1037, and for an informative web site (http://mcl1.ncifcrf.gov/waugh_tech.html). We also thank Pearl Quartey, Jerzy Osipiuk, Lour Volkart, and Hui Li for feedback on purifications of proteins, and Debbie Hanson, Phil Laible, and Marianne Schiffer for comments on the manuscript. Analysis of selenomethionine incorporation was carried out by Yale University's HHMI Keck Laboratories, and DNA sequencing by the University of Chicago Cancer Research Center's DNA Sequencing Facility. This work was supported by the National Institutes of Health Grant GM62414-01 and by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, under contract W-31-109-Eng-38.

References

[1] M. DiDonato, A.M. Deacon, H.E. Klock, D. McMullan, S.A. Lesley, A scaleable and integrated crystallization pipeline applied to mining the Thermotoga maritima proteome, J. Struct. Funct. Genomics 5 (2004) 133–146.
[2] P. Braun, Y. Hu, B. Shen, A. Halleck, M. Koundinya, E. Harlow, J. LaBaer, Proteome-scale purification of human proteins from bacteria, Proc. Natl. Acad. Sci. USA 99 (2002) 2654–2659.
[3] M.R. Dyson, S.P. Shadbolt, K.J. Vincent, R.L. Perera, J. McCafferty, Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression, BMC Biotechnol. 4 (2004) 32.
[4] W.B. Jeon, D.J. Aceti, C.A. Bingman, F.C. Vojtik, A.C. Olson, J.M. Ellefson, J.E. McCombs, H.K. Sreenath, P.G. Blommel, K.D. Seder, B.T. Burns, H.V. Geetha, A.C. Harms, G. Sabat, M.R. Sussman, B.G. Fox, G.N. Phillips Jr., High-throughput purification and quality assurance of *Arabidopsis thaliana* proteins for eukaryotic structural genomics, J. Struct. Funct. Genomics 5 (2004).
[5] A. Yee, K. Pardee, D. Christendat, A. Savchenko, A.M. Edwards, C.H. Arrowsmith, Structural proteomics: toward high-throughput structural biology as a tool in functional genomics, Acc. Chem. Res. 36 (2003) 183–189.
[6] Y. Kim, I. Dementieva, M. Zhou, R. Wu, L. Lezondra, P. Quartey, G. Joachimiak, O. Korolev, H. Li, A. Joachimiak, Automation of protein purification for structural genomics, J. Struct. Funct. Genomics 5 (2004) 111–118.
[7] P. Braun, J. LaBaer, High throughput protein production for functional proteomics, Trends Biotechnol. 21 (2003) 383–388.
[8] J. Porath, Immobilized metal ion affinity chromatography, Protein Expr. Purif. 3 (1992) 263–281.
[9] J.D. Fox, D.S. Waugh, Maltose-binding protein as a solubility enhancer, Methods Mol. Biol. 205 (2003) 99–117.
[10] 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.
[11] W.G. Dougherty, J.C. Carrington, S.M. Cary, T.D. Parks, Biochemical and mutational analysis of a plant virus polyprotein cleavage site, EMBO J. 7 (1988) 1281–1287.
[12] T.D. Parks, K.K. Leuther, E.D. Howard, S.A. Johnston, W.G. Dougherty, Release of proteins and peptides from fusion proteins using a recombinant plant virus proteinase, Anal. Biochem. 216 (1994) 413–417.
[13] L. Stols, M. Gu, L. Dieckman, R. Raffen, F.R. Collart, M.I. Donnelly, A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site, Protein Expr. Purif. 25 (2002) 8–15.
[14] R.B. Kapust, D.S. Waugh, Controlled intracellular processing of fusion proteins by TEV protease, Protein Expr. Purif. 19 (2000) 312–318.
[15] S. Nallamsetty, R.B. Kapust, J. Tozser, S. Cherry, J.E. Tropea, T.D. Copeland, D.S. Waugh, Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro, Protein Expr. Purif. 38 (2004) 108–115.
[16] H.Y. Yoon, D.C. Hwang, K.Y. Choi, B.D. Song, Proteolytic processing of oligopeptides containing the target sequences by the recombinant tobacco vein mottling virus NIa proteinase, Mol. Cells 10 (2000) 213–219.

[17] R.B. Kapust, J. Tozser, J.D. Fox, D.E. Anderson, S. Cherry, T.D. Copeland, D.S. Waugh, Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency, Protein Eng. 14 (2001) 993–1000.
[18] S. Doublie, Preparation of selenomethionyl proteins for phase determination, Methods Enzymol. 276 (1997) 523–530.
[19] G.D. Van Duyne, R.F. Standaert, P.A. Karplus, S.L. Schreiber, J. Clardy, Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin, J. Mol. Biol. 229 (1993) 105–124.
[20] C.S. Millard, L. Stols, P. Quartey, Y. Kim, I. Dementieva, M.I. Donnelly, A less laborious approach to the high-throughput production of recombinant proteins in Escherichia coli using 2-liter plastic bottles, Protein Expr. Purif. 29 (2003) 311–320.
[21] L. Stols, C.S. Millard, I. Dementieva, M.I. Donnelly, Production of selenomethionine-labeled proteins in two-liter plastic bottles for structure determination, J. Struct. Funct. Genomics 5 (2004) 95–102.
[22] W.B. Whitman, E. Ankwanda, R.S. Wolfe, Nutrition and carbon metabolism of *Methanococcus voltae*, J. Bacteriol. 149 (1982) 852–863.
[23] F.W. Studier, Protein production by auto-induction in high density shaking cultures, Protein Expr. Purif. 41 (2005) 207–234.
[24] R.C. Tyler, H.K. Sreenath, S. Singh, D.J. Aceti, C.A. Bingman, J.L. Markley, B.G. Fox, Auto-induction medium for the production of [U-^{15}N]- and [U-^{13}C, U-^{15}N]-labeled proteins for NMR screening and structure determination, Protein Expr. Purif. 40 (2005) 268–278.
[25] H.K. Sreenath, C.A. Bingman, B.W. Buchan, K.D. Seder, B.T. Burns, H.V. Geetha, W.B. Jeon, F.C. Vojtik, D.J. Aceti, R.O. Frederick, G.N. Phillips Jr., B.G. Fox, Protocols for production of selenomethionine-labeled proteins in 2-L polyethylene terephthalate bottles using auto-induction medium, Protein Expr. Purif. 40 (2005) 256–267.
[26] S. Moy, L. Dieckman, M. Schiffer, N. Maltsev, G.X. Yu, F.R. Collart, Genome-scale expression of proteins from *Bacillus subtilis*, J. Struct. Funct. Genomics 5 (2004) 103–109.