Journal of Structural and Functional Genomics (2006) 7:1–14
DOI 10.1007/s10969-005-9003-7

Effect of N-terminal solubility enhancing fusion proteins on yield of purified target protein

Martin Hammarström[1,3], Esmeralda A. Woestenenk[1], Niklas Hellgren[1], Torleif Härd[2] & Helena Berglund[1,3,]*

*[1]Department of Biotechnology, Royal Institute of Technology (KTH), SE-106 91, Stockholm, Sweden; [2]Department of Medical Biochemistry, Göteborg University, SE-405 30, Göteborg, Sweden; [3]Present address: Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE-171 77, Stockholm, Sweden; *Author for correspondence (e-mail: helena.berglund@mbb.ki.se; fax +46-08-524-86-868)*

Received 8 April 2005; accepted in revised form 21 September 2005

Key words: *Escherichia coli*, fusion protein, high throughput, His tag, solubility, structural genomics

Abstract

We have studied the effect of solubilising N-terminal fusion proteins on the yield of target protein after removal of the fusion partner and subsequent purification using immobilised metal ion affinity chromatography. We compared the yield of 45 human proteins produced from four different expression vectors: three having an N-terminal solubilising fusion protein (the GB1-domain, thioredoxin, or glutathione S-transferase) followed by a protease cleavage site and a His tag, and one vector having only an N-terminal His tag. We have previously observed a positive effect on solubility for proteins produced as fusion proteins compared to proteins produced with only a His tag in *Escherichia coli*. We find this effect to be less pronounced when we compare the yields of purified target protein after removal of the solubilising fusion although large target-dependent variations are seen. On average, the GB1 + His fusion gives significantly higher final yields of protein than the thioredoxin + His fusion or the His tag, whereas GST + His gives lower yields. We also note a strong correlation between solubility and target protein size, and a correlation between solubility and the presence of peptide fragments that are predicted to be natively disordered.

Abbreviations: eGFP – enhanced green fluorescent protein; GST – glutathione S-transferase; IMAC – immobilised metal ion affinity chromatography; MBP – maltose binding protein; MES – 2-(*N*-morpholino) ethanesulfonic acid; Ni-NTA – Ni^{2+}-nitrilotriacetic acid; ORF – open reading frame; RBS – ribosome binding site.

Introduction

Low solubility is one of the most frequently encountered problems when using *Escherichia coli* as a host for production of recombinant proteins. One approach to promote accumulation of soluble protein is to produce the protein as a fusion together with a solubilising fusion partner (reviewed in [1]). Several individual proteins or protein domains have been described and used as solubility enhancing fusion partners and recently some large scale comparative studies have been reported. We previously confirmed a clear positive effect on the solubility of recombinant proteins expressed with an N-terminal fusion partner as compared to an N-terminal His tag [2] and others have reported similar results [3–5]. However, several of these studies utilised expression vectors with variations in vector elements other than the fusion coding sequence that might influence the outcome. We recently described such effects in vectors producing proteins with an N-terminal His tag [6].

In the present study, we assess the utility of different fusion partners for production of a number of human proteins in *E. coli*. As we predominately work with smaller target proteins and domains we chose to include GST [7], thioredoxin [8], and the GB1 domain [9] as solubilising fusion partners in this study rather than the larger MBP [10, 11] and NusA fusions [12]. Thioredoxin from *E. coli* and the GB1 domain of protein G from *Streptococcus aurorus* are both small fusion partners that showed excellent solubilising properties in our previous work [2]. To specifically isolate the influence of the fusion partner we use expression vectors with identical backbones. We also extend our study to include removal of the solubilising fusion partner and purification of the isolated target proteins. Even if produced fusions appear soluble when expressed, there are several examples where proteins show characteristics classifying them as aggregated in analytical ultracentrifugation [13] or dynamic light scattering [14] experiments, or that proteins simply precipitate when the solubilising fusion partner is removed [15–18]. In this study we aim to assess not only the solubilising effect of fusion partners in *E. coli* but also how, or if, the solubilising effect influences the final yield of the purified target protein.

To enable easy and standardised purification in a 96-well plate format with identical conditions for proteins produced with and without the solubilising fusion partners we chose to include a His tag in all constructs. The use of a His tag is also justified by the fact that almost all high throughput protein production projects rely on this tag for affinity purification [19] and thus the comparison is of greater relevance than if it was to be done against the proteins in their native form and without IMAC purification. Although the addition of a His tag in the linker between the fusion partner and the target protein could affect the solubility of the fusion protein, this potential influence should be similar for the different fusion partners. The His tag enables us to use a single purification protocol to evaluate the utility of the fusion partners for improving solubility and to remove the fusion partner to compare the yields of the isolated target proteins.

We have used the methodology of a high throughput project in this study. Structural genomics and related projects that require large amounts of soluble and functional protein depend on standardised methods for protein production in order to handle the high number of protein targets. The use of standardised methodology of high throughput projects will enable evaluation of a large number of different factors for protein production, either by expanded studies for a limited number of proteins [2–5, 20–24] or by comparison of the results from large-scale projects employing different strategies [25]. Such studies are beneficial not only to the field of structural genomics but also to the biochemical field as a whole.

Material and methods

Vector construction

The vectors used in this study (Figure 1) are named pTH27 (histidine tag only), pTH28 (thioredoxin + His tag), pTH29 (GST + His tag) and pTH34 (GB1 domain + His tag). They are intended for use with the Gateway system (Invitrogen) for recombinational cloning and are based on the vector pTH18 (see below) that has the vector backbone of pET-21a. The His tag originates from the vector pDEST17 (Invitrogen). The thioredoxin fusion originates from the vector pDEST16 (Invitrogen) but has the enterokinase protease cleavage site replaced with a PreScission protease cleavage site where the linker sequence between the fusion partner and the protease site is identical to the original fusion construct designed by LaVallie and co-workers [8]. The GST fusion originates from the vector pDEST-TH6 [18] and the linker sequence is derived from the pGEX-2TK expression vector (GE healthcare) except for the final proline residue. The GB1 fusion is based on the vector pTH18 that in turn was constructed from the vector pDEST-TH3 [2] by replacing the tobacco etch virus protease cleavage site with a PreScission protease cleavage site.

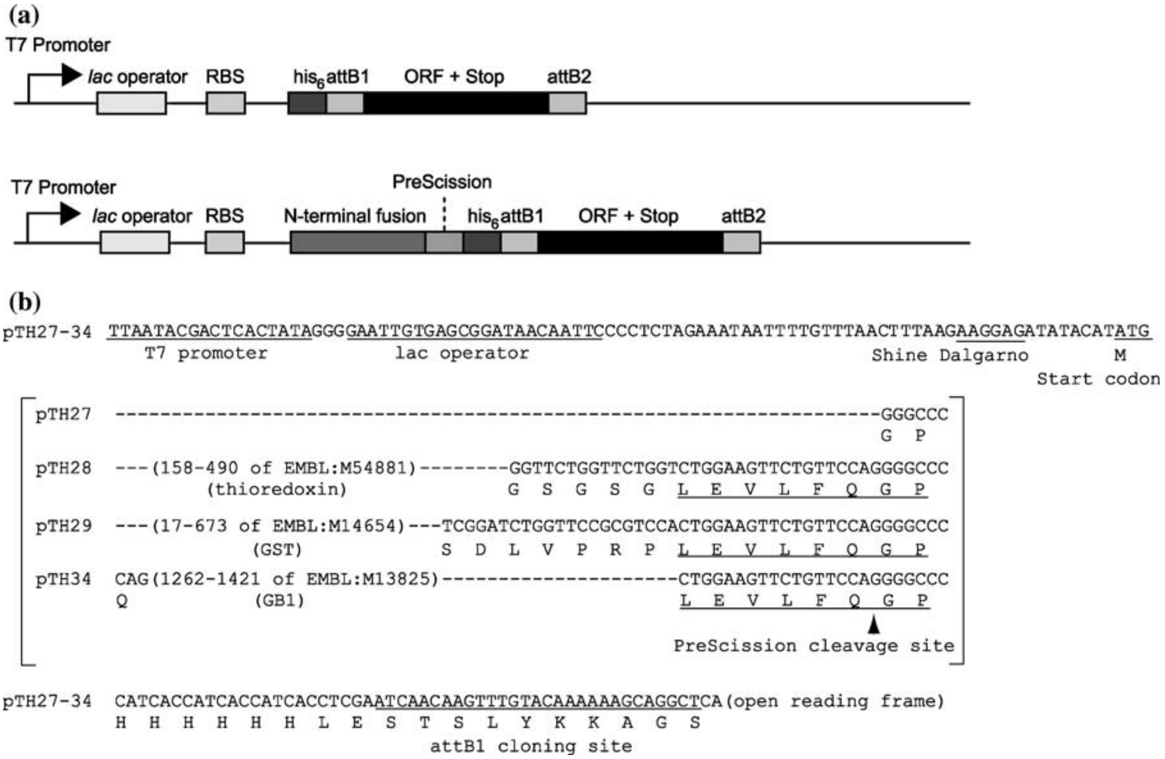
(a)
T7 Promoter
lac operator
RBS
his6attB1
ORF + Stop
attB2
T7 Promoter
lac operator
RBS
N-terminal fusion
PreScission
his6attB1
ORF + Stop
attB2
(b)
pTH27-34 TTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATG
T7 promoter
lac operator
Shine Dalgarno
M
Start codon
pTH27 --GGGCCC
G P
pTH28 ---(158-490 of EMBL:M54881)--------GGTTCTGGTTCTGGTCTGGAAGTTCTGTTCCAGGGGCCC
(thioredoxin) G S G S G L E V L F Q G P
pTH29 ---(17-673 of EMBL:M14654)---TCGGATCTGGTTCCGCGTCCACTGGAAGTTCTGTTCCAGGGGCCC
(GST) S D L V P R P L E V L F Q G P
pTH34 CAG(1262-1421 of EMBL:M13825)---------------------CTGGAAGTTCTGTTCCAGGGGCCC
Q (GB1) L E V L F Q G P
PreScission cleavage site
pTH27-34 CATCACCATCACCATCACCTCGAATCAACAAGTTTGTACAAAAAAGCAGGCTCA(open reading frame)
H H H H H H L E S T S L Y K K A G S
attB1 cloning site

Figure 1. Expression vectors used in this study with (a) schematic representation and (b) detailed sequence information for relevant elements. The pTH27 vector, encoding only an N-terminal His tag, is shown in the upper part of (a) and the vectors pTH28 (thioredoxin), pTH29 (GST), and pTH34 (GB1), which also include an N-terminal fusion partner and a PreScission protease cleavage site are shown in the lower part. The similar and within hard brackets, dissimilar parts of the vector sequences are shown in (b) with relevant DNA or protein sequence elements underlined. The coding sequences of the fusion partners are given by their EMBL accession numbers and specified base-pairs.

The first step in the vector conversion was to cut out the XbaI-NcoI fragment containing the ribosome binding site, start codon, fusion coding sequence including protease cleavage site, and the attR1 recombination cloning site from the different original vectors. These fragments containing the different fusion coding sequences were subsequently ligated into the corresponding NcoI-XbaI vector backbone fragment of pTH18, thus creating a set of vectors in which a T7*lac* promoter controls the expression and which are identical with exception for the N-terminal fusions they encode. These intermediate vectors are named pTH37 (His tag), pTH36 (thioredoxin) and pTH35 (GST). The second step was to introduce the His tag in the linker between the protease cleavage site and the attR1 cloning site, or in the case of the His tag vector to change the amino acids in front of the His tag so that they are identical to the other vectors. These changes were made by site-specific mutagenesis following the protocols of Wang and Malcolm [26]. The pTH37 vector was mutated with primers 5′CTTTAAGAAG GAGATATACA TATGGGTCCG CATCACCATC ACCATCACCT CGAATCAAC3′ and 5′GTTGATTCGA GGTGATGGTG ATGGTGATGC GGACCCATAT GTATATCTCC TTCTTAAAG3′ thus changing the initial amino acids from Met-Ser-Tyr-Tyr to Met-Gly-Pro (underlined bases). The pTH18, pTH35 and pTH36 fusion protein vectors were mutated with the primers 5′GTTCCAGGGG CCCCATCACC TCACCATCAC TCGAATC CAAGTTTGT AC3′ and 5′GTACAAACTT GTTGATTCGA GGTGATGGTG TGGTGATGG GGCCCCTGGA AC3′ to introduce the His tag (underlined bases) between the protease cleavage site and the attR1 recombination cloning site. The mutated XbaI-NcoI fragments were cleaved and ligated into the original NcoI-XbaI vector backbone fragment to eliminate the risk of including PCR-derived mutations in the vector backbone. The N-terminal fragment remaining on the produced proteins (GPHHH HHHLE STSLY KKAGS) after removal of the fusion proteins are identical in the three pTH28, pTH29 and pTH34 vectors, and they only differ by an N-terminal methionine to the fragment obtained with pTH27. All batches of destination vectors are tested for retained activity of the *ccdB* gene as described in the Gateway manual (Invitrogen) and also for giving a sufficient number of colonies (> 50) after transformation when used in a 5 μl LR cloning reaction.

Target proteins

The target proteins are listed in Table 1. The 45 genes are of human origin and can be divided into three different categories. Seven genes (2–28) in the first category are selected from our previous study [2] for producing soluble protein when expressed with an N-terminal His tag. The second category contains 29 genes (66–159) selected by the criteria of being cancer or disease related, not having sequence homology to any protein of known structure and not belonging to any PfamA family [27]. A majority of the genes in this category are known to express well with an N-terminal His tag while the behaviour of others is unknown. The third category contains nine gene fragments (202–215) corresponding to protein domains that are cancer or disease related and that belong to PfamA families without homology to proteins of known structure. The molecular mass of the produced proteins varies from 7 to 90 kDa with an average of 23 kDa. Three controls were included. As a convenient and colourful positive control of cloning, expression and purification we chose to include the enhanced green fluorescent protein (eGFP) [28] originally from *Aequorea victoria*. To verify that the destination vectors gave no false positive colonies in the cloning reaction a negative control lacking an entry clone was used. This control was also used as a blank during purification. To estimate the *E. coli* background in the purification step a control with the pUC-18 vector instead of a destination vector was used.

Cloning, transformation and expression

All genes were available in-house as sequenced Gateway entry clones in the pDONR-201 vector. Recombination cloning reactions using the Gateway cloning system (Invitrogen) were set up in a total volume of 3.75 μl with 0.75 μl each of LR clonase enzyme mix and LR cloning buffer, 1.0 μl destination vector (70 ng/μl), and 1.25 μl entry clone (50 ng/μl). After incubation at 25 °C for 2 h the reactions were transformed into 40 μl of Ca^{2+}-competent Rosetta (DE3) pLysS *E. coli* cells (Novagen) and plated on LB agar plates with 100 μg/ml ampicillin and 34 μg/ml chloramphenicol. All cloning and transformation reactions were performed in 96-well plates but following transformation, the reactions were plated on agar plates. Ten colonies were picked and pooled to make glycerol stocks for each expression clone. The glycerol stocks were used to start overnight cultures in 1 ml of Luria-Bertini broth with 100 μg/ml ampicillin and 34 μg/ml chloramphenicol in 96*2 ml plates. Two hundred microlitres of the overnight cultures were used to inoculate 5 ml expression cultures in Luria-Bertini broth with 100 μg/ml ampicillin and 34 μg/ml chloramphenicol in 24*10 ml polypropylene plates (Qiagen). Cells were grown for 2.5 h at a temperature of 37 °C and expression was induced by addition of isopropyl-*β*-D-thiogalactopyranoside (GE healthcare) to 1 mM. After 3 h of expression at 37 °C, the cells were harvested using centrifugation at 1500 × *g* for 20 min and cell pellets were frozen at −20 °C.

Purification

The frozen cell pellets were thawed in 270 μl lysis buffer (300 mM NaCl, 50 mM NaH_2PO_4, and 10 mM imidazole at pH 8.3) containing 2 mM $MgSO_4$, 60 U benzonase (Merck), and 40 μg hen egg white lysozyme (Boehringer Mannheim) and incubated at a temperature of 37 °C for 25 min. Two additional freeze-thawing cycles were performed to ensure efficient lysis. Soluble and insoluble fractions were separated by centrifugation at 2500 × *g* for 30 min. The insoluble fractions were resuspended in a volume equal to that of the soluble fractions and 12 μl samples were stored for gel analysis.

Table 1. Proteins with short descriptions, cDNA entry and molecular masses.

Number	Protein/domain name	cDNA clone (genebank)	Amino acids	Molecular mass[a] (kDa)
2	Over-Expressed Breast Tumor Protein[d]	N31559	2–74	8.0
7	T-Cell Activation Protein	AA913112	2–127	14.5
13	Heat Shock Factor Binding Protein	AI889952	2–76	8.4
16	Protein Kinase Inhibitor Alpha	AW163048	2–75	7.9
19	Translationally Controlled Tumor Protein	AI721229	2–172	19.5
27	Augementer of Liver Regeneration	AI992142	2–125	14.9
28	B-Cell Translocation Gene 1 Protein	AI479605	2–171	19.1
66	Cytochrome C Oxidase Copper Chaperone	BC010933	1–63	6.9
67	Transcription Factor BTF3 Homolog 2	BC008062	1–73[c]	8.2[c]
68	Leukemia Associated Protein 1	BC020692	1–78[c]	8.2[c]
69	CATR Tumorigenic Conversion 1 Protein	BC008502	1–99	10.7
71	Protein AF1Q	BC021703	1–90	10.1
74	C-Myc Binding Protein	BC008686	1–103	12.0
75	T-Cell Leukemia Translocation-Associated Gene Protein	BC005157	1–103	11.3
76	P75NTR-Associated Cell Death Executor	BC003190	1–111	13.0
77	SH3 Domain-Binding Glutamic Acid-Rich-Like Protein	BC016709	1–114	12.8
81	Melanoma Antigen P15	BC000507	1–128	14.9
84	G Antigen Family D 2 Protein	BC009538	1–81	9.1
90	NADH-Ubiquinone Oxidoreductase 19 kDa Subunit	BC001016	1–172	20.1
93	Tumor Protein D52 (N8 Protein)	BC018117	1–184	19.9
97	Tumor Protein D53 (HD53) (D52-Like 1)	BC002375	1–155[c]	17.4[c]
98	Neighbour of COX4	BC007445	1–210	23.8
99	CBP/P300-Interacting Transactivator 2[d]	BC004377	1–270	28.5
100	Ubiquinone Biosynthesis Protein COQ7 Homolog	BC003185	1–217	24.3
104	Thiopurine S-Methyltransferase	BC009596	1–245	28.2
105	Breast Cancer Metastasis-Suppressor 1	BC009834	1–246	28.6
107	Myeloid Leukemia Factor 2	BC000898	1–248	28.1
108	Sperm-Specific Antigen 2 (Cleavage Signal-1 Protein)	BC012947	1–267	28.9
113	Cytosolic Ovarian Carcinoma Antigen 1 (APK1 Antigen)	BC019254	1–317	36.9
120	Nipsnap1 Protein	BC002371	1–284	33.3
121	Interferon-Induced 35 kDa Protein[d]	BC001356	1–288	31.8
123	N-Myc-Interactor	BC021987	1–307	35.1
129	P53-Induced Protein 8	BC002390	1–340	39.0
147	Melanoma Antigen Preferentially Expressed In Tumors (OIP4)	BC014074	1–509	57.9
157	Restricted Expression Proliferation Associated Protein 100	BC020207	1–747	85.6
159	Colorectal Mutant Cancer Protein (MCC Protein)	BC009279	1–840[c]	94.2[c]
202	Anti_proliferat domain of BTG3	BC011957	1–203	23.6
203	HMG14_17 domain of HMG14	BC023984	2–96	10.2
204	PWP2 domain of PWP2H	BC013309	765–882	13.6
208	DCX domain of DCX	BC027925[b]	70–134	7.4
209	mbt domain of SCML2	BC051913[b]	67–140	8.2
210	PUA domain of DKC1	BC009928[b]	295–370	8.2
213	DUF51 domain of AMMECR1	AK091430[b]	129–303	20.5
214	UFD1 domain of UFD1L (short isoform)	BC001049[b]	14–194	20.5
215	GTP_CDC domain of PNUTL1	BC025261[b]	41–321	32.4

[a] Protein in its native form.

[b] Amplified from mRNA pool but with sequences identical to the specified cDNA-clones.

[c] Sequencing revealed differences in the 3′ primer region as compared to the cDNA entry resulting in extensions or deletions of 5–10 amino acids. The mass given in the table is the size of the resulting protein. Care should be taken when relating the expression and solubility results for these proteins to those of other studies.

[d] These genes contain single non-silent mutations as compared to database entry.

Purification of the soluble fractions was carried out after addition of 2-mercaptoethanol to 10 mM in all wells and four units of PreScission protease (GE healthcare) where required for fusion cleavage. Proteolysis was carried out at 37 °C for 2 h. The protein samples were loaded onto 100 μl of pre-equilibrated Ni-NTA superflow resin (Qiagen) in a 96*1 ml well filter plate (Pall) by centrifugation at 200 × *g* for 2 min and washed twice with 600 μl wash buffer (300 mM NaCl, 50 mM NaH_2PO_4, and 20 mM imidazole at pH 8.3). Initially filter plates with a pore size of 0.22 μm were used, but this resulted in problems with clogging of wells. Consequently we switched to plates with a pore size of 1 μm. Elution was performed in three separate steps with 140, 280, and 560 μl elution buffer (300 mM NaCl, 50 mM NaH_2PO_4, and 250 mM imidazole at pH 8.3).

Detection

The purity of the eluted proteins was evaluated by running 12 μl from the first 140 μl eluate on Criterion XT 4–12% gradient Bis–Tris gels (Biorad) using a 2-(*N*-Morpholino) ethanesulfonic acid (MES) based running buffer (50 mM MES, 50 mM Tris, 0.5% w/v SDS, and 1 mM EDTA at pH 7.3). Gels were stained with Coomassie stain and in some cases subsequently stained with silver stain (Biorad) for detection of proteins at lower concentrations. SDS-PAGE gel band intensities were measured using the Quantity One software (Biorad) on images from a Fluor-S MultiImager gel camera (Biorad). Protein size markers (low molecular weight, GE healthcare) were used to normalise band intensities from different gels. In cases where the proteolysis was incomplete as judged from the SDS-PAGE gel results, the expression and purification procedure was repeated.

For all proteins produced from each of the four vectors, 30 μl from each of the three elution steps in the purification were used for Bradford assay [29] in a 96 well format. The amount of protein in each eluted fraction was calculated from the measured absorbance in the Bradford assay using bovine serum albumin as a standard and the total amount of each eluted protein was calculated.

Mass spectrometry

Mass spectrometry was performed with a Micromass Q-TOF2 (Waters Corporation, Micromass MS Technologies) with a nanoflow electrospray ionisation source. A three-pump Waters CapLC system with autosampler and Stream Select module was used for on-line desalting of protein samples using a C4 μ-Precolumn Cartridge (300 Å, 300 μm × 5 mm, LC Packings). Samples were eluted using a gradient of acetonitrile in water containing 0.1% formic acid. Raw data were deconvoluted using the MaxEnt1 algorithm in the MassLynx software package (Micromass) to produce zero-charge spectra.

Results and discussion

Expression

Average expression and solubility data for the 45 proteins produced from the four different expression vectors are compiled in Figure 2 and examples of SDS-PAGE gels are shown in Figure 3. The results for all protein constructs can be found in the supplementary material. Of the 45 genes, 43 are expressed in at least one of the vectors and a majority are expressed in all four vectors. Even though we find a slight variation in cell density at harvest between cultures producing different proteins and also between different experiments, we see no variation in average cell density at harvest between the different vectors. We attribute the even growth to the identical backbones of the vectors. There are variations in the levels of both total expression and soluble expression depending on the fusion encoded by the vector. The average expression levels are almost twice as high for the thioredoxin + His and GB1 + His fusions as for the GST + His fusions and the proteins having only the His tag (Figure 2a).

Since the expression vectors are identical except for the N-terminal fusions, the overall differences in total expression levels between the vectors is likely to depend on the intrinsic properties of these coding sequences. Translational efficiency is known to depend on the mRNA sequence around the ribosome binding site including the 5′ end of the coding sequence [30, 31]

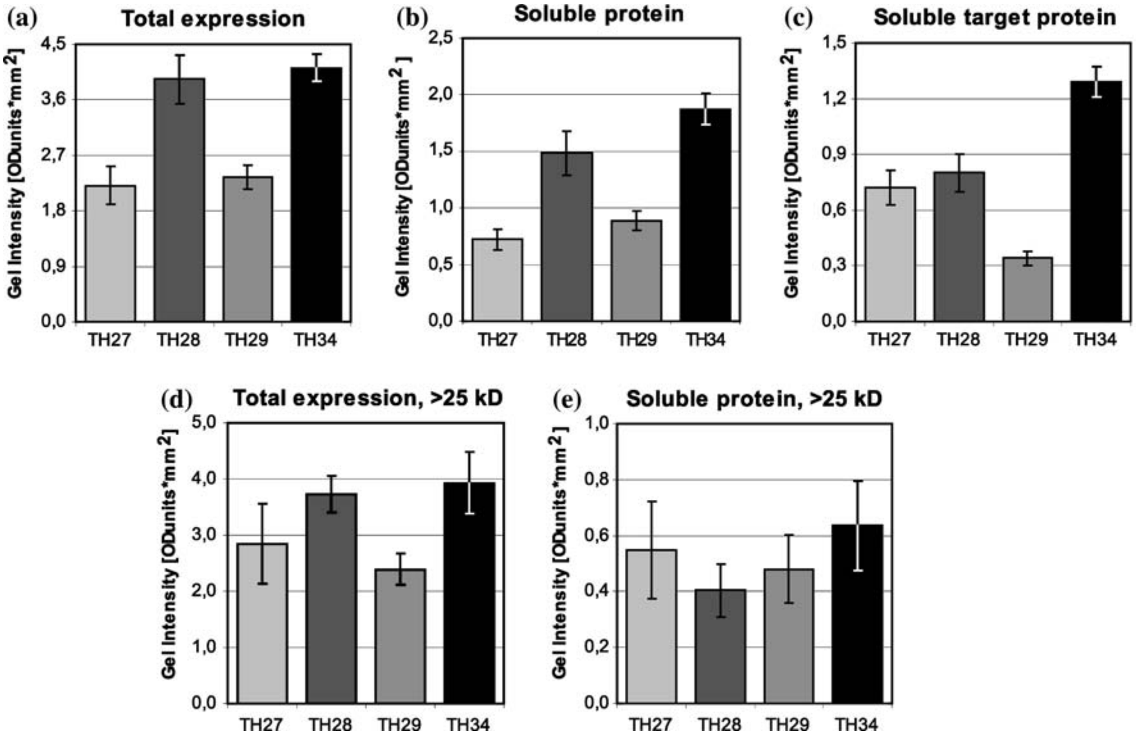

Figure 2. Average total (a) and soluble (b) expression based on gel band intensity data. For each of the expression vectors (TH27) His tag only, (TH28) thioredoxin + His tag, (TH29) GST + His tag, and (TH34) GB1 + His tag, the averages are taken over all target proteins including the eGFP positive control. All cultures were performed in triplicate and the error bars represent the standard deviation for the three expression experiments. In (c) the soluble expression has been reduced to compensate for the contribution of the fusion partners. In addition, the average total and soluble expression taken only over the target proteins with molecular weights above 25 kD are shown in (d) and (e).

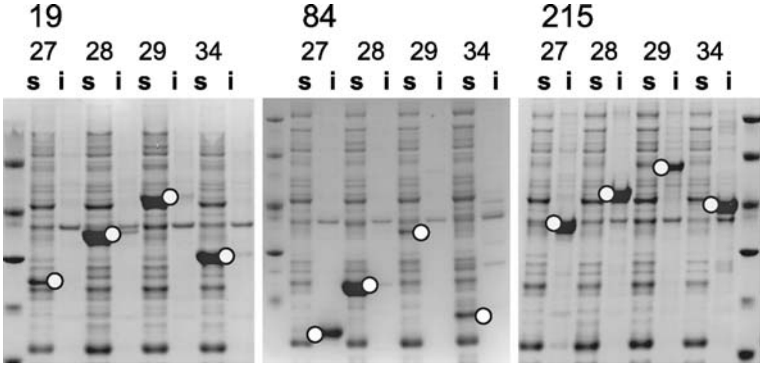

Figure 3. Examples of SDS-PAGE gels with soluble (s) and insoluble (i) fractions following lysis. The results when produced from the four different expression vectors (27: His tag only; 28: thioredoxin + His tag; 29: GST + His tag; 34: GB1 + His tag) are shown for three different target proteins. Target protein 19 is expressed to high levels and as soluble, target protein 84 has vector dependant variations in expression level and solubility, and target protein 215 is expressed to high levels, but as insoluble. A white circle indicates the position of each protein. Unlabeled lanes are protein size marker with protein masses: 97, 66, 45, 30, 20, and 14 kDa, respectively.

which varies in our four vectors. Specifically, we note a correlation between the overall expression levels for the different fusions and the suggested influence of the second codon on expression levels [31]. Both the GB1 and the thioredoxin fusions have moderately favourable second codons (CAG and AGC, respectively), whereas the GST fusion and the His tag both have less favourable second codons (UCC and GGU, respectively). Additionally, within 60 bases of the 5′ end of their coding sequences, both the GST fusion and the His tag have a cluster of low frequency codons which are not compensated for by the *E. coli* strain Rosetta (DE3) pLysS. It is conceivable that such a cluster of several consecutive uncompensated rare codons will affect the translation negatively. Thus, optimisation of the coding sequence to ensure high translational efficiency could probably improve the production yields for both the GST + His and the His tag fusion vectors.

Solubility

Of the 43 genes that are expressed, 14 are produced as predominantly insoluble (>90%) in all four different fusion constructs, highlighting the difficulty of producing eukaryotic proteins in *E. coli*. Only six proteins are produced as predominantly soluble (>90%) in all four different expression vectors. The average amounts of soluble protein produced from the four vectors follow the same pattern as the total expression levels (Figure 2b). The amounts of soluble protein are higher for the GB1 + His and thioredoxin + His fusion proteins than for the GST + His fusion proteins or the His-tagged proteins. Looking at individual targets there is little or no difference between the alternative fusions constructs for the highly soluble target proteins. For the 23 target proteins with intermediate solubility it is usually the GB1 + His and the thioredoxin + His fusions that give the highest amount of soluble protein, corresponding to a solubility enhancement by the added fusion partner as seen for approximately 50% of the target proteins in this group. Finally, for a few of these target proteins there is a drastic increase in amounts of soluble protein when expressed as a fusion protein as compared to His-tagged, corresponding to a solubilisation of the target protein. We also note that the solubility enhancing effect depends on the size of the target protein. For the larger target proteins, above 25 kDa, the amounts of soluble protein decrease for the thioredoxin + His and GB1 + His fusion partners even though the total expression levels are unchanged (Figure 2d, e). Thus the solubility enhancement of these fusion partners is less pronounced for larger target proteins.

In spite of the use of solubilising fusion partners, a substantial number of proteins still resist soluble expression. The experience from structural genomics projects on thermophilic organisms is that approximately 40% of the selected target proteins can be produced in a soluble form when expressed as His-tagged proteins in *E. coli* [32, 33]. For proteins of eukaryotic origin, this number drops considerably [34]. Comparing the results for our set of target proteins to related studies, we see several reoccurring patterns that contribute to the low success-rate for soluble expression: As the size of the target protein increases, the probability for soluble expression decreases as seen in the present study and several others [3, 5, 32]. Furthermore, one third of the selected proteins in the present study do not belong to any PfamA family [27], which has proven to be another negative determinant for expression of recombinant protein in a soluble form in *E. coli* [25]. Nevertheless we include this type of proteins to represent proteins with little functional annotation for which the structure could contribute novel information. Such proteins represent an important class of proteins in a structural genomics initiative and cannot be neglected as potential target proteins. In addition, several of the smaller proteins that are expressed as predominantly insoluble in this study are from the category with protein domains rather than full-length proteins. Expression and solubility of protein domains are known to depend strongly on a correct choice of domain boundaries of the expression construct and at least one of the protein domains that were produced as insoluble in this study has been expressed in a soluble form with a different choice of domain boundaries [35]. Finally, two thirds of our selected proteins include amino acid sequences longer than 30 amino acids that are predicted to be disordered by the disEMBL algorithm [36]. A large content of disordered segments is reported to impair soluble expression of the protein in *E. coli* [5]. Intriguingly, when we use the criteria set by Uversky and co-workers [37] for prediction of intrinsically unfolded proteins based on the amino acid composition of the entire polypeptide chain, we find that a majority of the 14 proteins that match these criteria are soluble in *E. coli*. It appears that proteins with no or very low structural content often can be produced in *E. coli* whereas proteins containing both structured and large unstructured regions can be problematic.

Purification

A frequently raised concern regarding the use of solubilising fusion partners is the potential aggregation of the target protein upon proteolytic removal of the fusion partner. There is also a risk that the protease cleavage site is made inaccessible due to an aggregated, but soluble, state of the fusion protein. We therefore extended this study to include proteolytic removal of the fusion partners and subsequent affinity purification to examine if the differences in expression and solubility also are retained in the final yield of purified target protein. This is under the assumptions that: (1) the presence of soluble aggregates before removal of the fusion will inhibit proteolysis to such an extent that the incomplete cleavage will be detectable on gel or that the aggregation will cause both proteolysis and purification to fail and (2) purification will be prevented if the target protein aggregates when the fusion partner is removed. No reproducible examples of incomplete proteolysis are observed for any of the target proteins in this study. Some individual cleavage reactions had to be repeated due to incomplete proteolysis, but in no case did this involve more than one of the reactions in each triplicate.

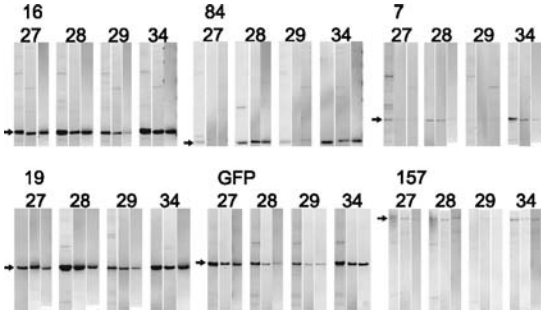

Figure 4. Examples of triplicate SDS-PAGE gel strips for six proteins with varying yields detected on Coomassie-stained gels after purification. The results when produced from the four different expression vectors (27: His tag only; 28: thioredoxin + His tag; 29: GST + His tag; 34: GB1 + His tag) are shown for the target proteins: 7, 16, 19, 84, 157 and the positive control eGFP. The arrows to the left of the gels indicate the position of the correct band.

After proteolytic removal of the fusion partners and subsequent IMAC purification of the target proteins, 26 out of 45 proteins are detected in the Bradford assay and 29 proteins are detected on SDS-PAGE protein gels, six of which are only detected by the more sensitive silver stain. The amounts of purified protein range up to 200 μg and the background level of contaminating *E. coli* proteins is approximately 10 μg as estimated from the pUC-18 negative control. Although cell densities at harvest are low with an optical density at 600 nm of about 0.8 absorbance units, the amounts of purified target protein theoretically correspond to up to 40 mg of protein per litre cell culture. Figure 4 shows examples of gel strips from Coomassie stained SDS-PAGE gels of samples from the first elution step. When comparing the average yield of purified protein produced from our four expression vectors (Figure 5) we find it to be highest when produced from the GB1 + His vector which gives about 25% more than the His tag and the thioredoxin + His vectors. The GST + His vector yields approximately 50% less purified protein than the three other vectors. A statistical analysis using Student's *t*-test with paired data on the Bradford results show that the differences between the expression vectors are significant at the 99% confidence level. All targets that generally yield a high fraction of soluble protein in the expression screen are also purified to high yields. For the target proteins with intermediate solubility the

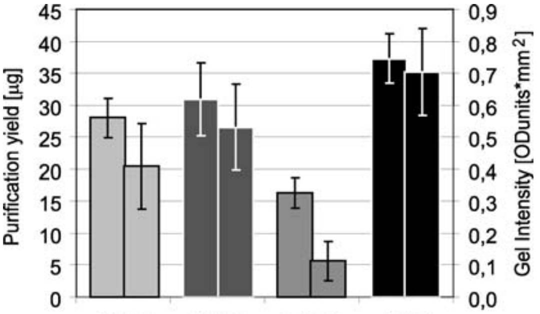

Figure 5. Amount of purified target protein based on Bradford data (leftmost bar, major axis) and SDS-PAGE image analysis (rightmost bar, minor axis). For each of the four expression vectors encoding respectively an N-terminal His-tag (pTH27), a thioredoxin + His tag (pTH28), a GST + His tag (pTH29), and a GB1 + His tag (pTH34), the data reflect average purification yield of all 45 target proteins including the positive control eGFP when performed in triplicate. The error bars represent the standard deviation for the three purification experiments.

purification yields are generally lower and in some cases purification failed. For the proteins that only were solubilised by the addition of a fusion partner the purification yields are generally low or non-detectable with the exception being target 84, which is successfully purified after removal of the fusion partner. The purification yields for all protein constructs can be found in the supplementary material.

Based on the relative sizes of the fusion partners and the target proteins, the contribution of the fusion protein can be subtracted and the amount of soluble target protein available for purification calculated (Figure 2c). Theoretically this corresponds to a proteolytic removal of the fusion partner and indicates the expected amount of soluble target protein available for purification, provided that proteolysis is complete and that no aggregation occurs. For example it can be seen that for the GST fusion partner, the combination of low expression levels and large size of the fusion partner result in low levels of soluble target protein available for purification. As the GB1 fusion partner is small, the yield of soluble target protein is still expected to be high. A comparison of the expected amounts of soluble protein available before purification (Figure 2c) and the yield after purification (Figure 5) shows that the increased amount of soluble protein created by e.g. the GB1 fusion as seen in the expression screen also is reflected in higher yield of purified target protein after removal of the fusion partner. This good quantitative agreement indicates that proteolysis and purification have worked without extensive sample loss, or at least that any potential losses are roughly equal for the different fusion partners. It should be clarified, however, that the actual amounts are not comparable between the two figures, as gel sample preparations and gel-to-gel normalisations are different.

An average increase in purification yield in the order of 25% as seen for the GB1-His constructs can be very beneficial for high throughput applications as a higher number of proteins can be produced in sufficient quantities (Figure 6). In our case, 13 target proteins are purified to a yield above 50 μg using the GB1 + His tag fusion, whereas for the thioredoxin + His tag fusion or the His-tagged proteins, nine target proteins are purified to similar levels. For the GST + His tag fusion no target proteins are purified to such high levels. Thus we conclude that there can be a positive effect on the total yield of purified target protein when using solubilising fusion partners, even after removal of the fusion partner. However, the effect is not general for all fusion partners as exemplified by the GST protein. One should also bear in mind that soluble expression is no guarantee for structured well behaving target protein but it will facilitate, and in many cases, enable further investigations of the protein.

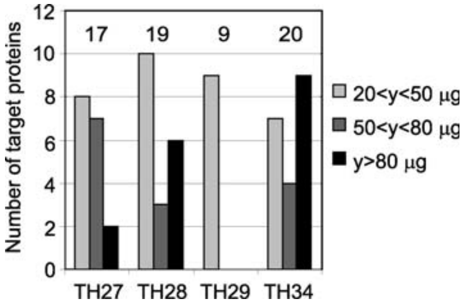

Figure 6. The number of target proteins that are purified to low, intermediate or high yield when produced from each of our vector: (TH27) N-terminal His-tag, (TH28) thioredoxin + His tag, (TH29) GST + His tag, and (TH34) GB1 + His tag. The total number of target proteins that are purified, irrespective of yield, is shown above each respective vector.

Fusion partners

The mechanism by which the fusion partners exert their solubilising function is not clear and possibly differs between the fusion proteins. There are now several comparative studies on the solubilising effect of fusion proteins [2–5] and even if there are differences in vector backbones and promoters, the presence of linkers, protease site or additional tags, the choice of host strain for expression, cultivation temperature, lysis condition, purification, proteolytic removal of fusion partners, detection methods and thresholds for scoring expression, solubility, or purification as successful between the studies, some general trends can be seen. First and foremost, although some fusion partners are generally good at increasing the yield of soluble protein, there are often variations in which fusion partner that give the highest yield for a specific target protein. Thus, it can be worthwhile to screen several different fusion partners in order to optimise the yield for a specific target protein. However it is important that the sequences encoding the N-terminal fusions provide good translation initiation properties in order to generate high expression levels. Another general observation is that soluble expression often is difficult to achieve for N-terminally His-tagged proteins from vectors with a T7 promoter such as the pDEST-17 vector used in [2, 3, 5, 6], but can be markedly improved using a T7*lac* promoter as in the present study and also in [5, 6].

Larger fusion proteins such as MBP and NusA are consistently good at producing soluble fusion protein [4, 5, 38], but it should be noted that the large size of this fusion could make the interpretation over-optimistic if it is included in the quantification and that large fusion partners pose a high metabolic cost on the cells. The utility of GST as a fusion partner has been questioned as it in some comparative studies has shown a less pronounced solubilising effect [4, 38]. GST has also been reported to be susceptible for *in vivo* degradation in *E. coli* [3, 5] and an additional concern is that the dimerisation of GST interferes with proteolytic release of the target protein. However, inspection of the crystal structure of the *S. japonicum* GST dimer (PDB entry 1M9B) reveals that the C-terminus of each polypeptide chain is positioned on opposite sides of the dimer and directed away from the dimeric interface and no obvious steric hindrance of the proteolytic cleavage is likely to be derived from the dimerisation. In accordance with the structural data we do not observe any decrease in the efficiency of the proteolysis step when comparing the GST fusion proteins with the GB1 and thioredoxin constructs. Although the yield from our GST vector in the present study is lower, GST manages to solubilise a number of proteins and has proven successful in some studies. Since GST additionally functions as an affinity purification handle it is a fusion protein well worth to include, although an optimisation of the gene sequence for improved translational efficiency in *E. coli* may be in place to improve yields. The solubilising effect of thioredoxin has also been questioned [4, 38] but appears to be a good choice for smaller target proteins [2, 5]. However, in our hands the GB1 domain is an even better choice. The GB1 domain can in addition be used for affinity purification whereas thioredoxin cannot unless the engineered His-patch variant is used [39].

Mass spectrometry

Mass spectrometry analysis was performed to investigate if *E. coli* processed the initial methionine in our His-tagged proteins. Nine proteins produced from the His tag (pTH27) vector and eight proteins produced from the GB1 + His tag (pTH34) vector were analysed (Table 2). All proteins produced from the His tag vector have a major peak corresponding to the expected mass provided that the initial methionine is still present, but several also have smaller peaks corresponding to the mass with the methionine removed and also with the following glycine removed (Figure 7). This is in line with the observations of Hirel and co-workers that a proline in the third position can inhibit removal of initial methionine even if the amino acid in the second position favours methionine removal. They also observed that the second amino acid can be removed to some extent in this case [40]. None of the proteins produced as fusion proteins and subjected to proteolysis before purification has these kinds of heterogeneities.

The HMG14 protein

Protein 203 is an example of atypical behaviour that is frequently encountered when working with large sets of proteins. It is expressed and soluble in the His tag vector but expressed to low or undetectable levels in the other fusion vectors. This results in higher yields of purified protein

Table 2. Mass spectrometry analysis of purified proteins.

Protein	TH27 (His tag)		TH34 (GB1 + His tag)	
	Calculated[a]	Measured	Calculated[b]	Measured
13	10,786	10,785	10,655	10,654
16	10,230	10,230	10,100	10,098
66	9288	9282[c]	9158	9150[c]
67	10,584	10,584	10,454	10,453
74	14,339	14,340	14,210	14,209
77	15,147	15,147	15,017	15,016
93	22,236	22,237	22,107	22,106
214	22,839	22,839	22,707	22,708
203	12,529	12,529	–	–

[a] With initial methionine remaining.

[b] After removal of fusion protein by proteolysis.

[c] The mass difference is probably due to the six cysteines of this protein being in an oxidised state.

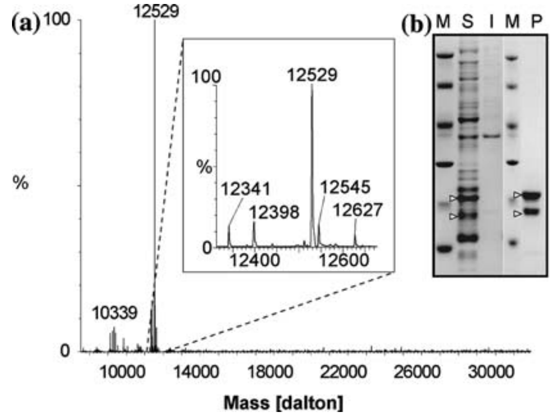

Figure 7. Zero charge spectrum (a) and SDS-PAGE gels (b) for protein 203 when produced from the His tag vector pTH27. The mass spectrum has the expected peak (12,529) as well as a peak corresponding to the truncated protein (10,339) that lacks 21 amino acids at the C-terminus. The inset shows the region around 12 kDa with smaller peaks corresponding to full-length protein with the initial methionine removed (12,398), the following glycine removed (12,341), and oxidation of methionine (12,545). The lanes in (b) are designated (M) protein size marker with the protein masses 97, 66, 45, 30, 20, and 14 kDa respectively, (S) and (I) soluble and insoluble fractions following lysis, and (P) purified protein. Arrows indicate the positions of the full-length and the truncated proteins.

for the His-tagged protein as compared to the fusion proteins, not because the solubility is higher but because the expression level is higher. The mechanism for this effect is not clear. In addition to the atypical expression pattern, the expressed product is seen as two distinct bands with apparent molecular masses of 18 and 21 kDa on a SDS-PAGE gel (Figure 7). Both bands are also present after IMAC purification. A mass spectrum confirmed the presence of a shorter product in addition to the expected protein and from the mass of the shorter product it is possible to predict that 21 amino acids are lacking at the C-terminus (Figure 7). Tryptic digests and peptide mapping of the two different proteins cut out from SDS-PAGE gel confirmed this. The HMG14 example illustrates the difficulties in finding a uniform approach for all proteins, as in this case the addition of an N-terminal fusion partner decreases the purification yield.

Notes on methodology

In the present study, 43 of the 45 selected target genes could be expressed to detectable levels and 38 are at least in part soluble when produced with one or more of the tested expression vectors. After removal of the fusion partner and IMAC purification, 15 of the proteins are recovered with high yield (> 8 mg/l) and a total of 29 of the proteins could be detected on silver stained SDS-PAGE protein gels. Although many of the proteins in this study are purified to moderate or low levels, it is likely that the purification yield of most of these can be improved by using more gentle production conditions such as lower cultivation temperature, more rapid lysis at a lower temperature, and addition of protease inhibitors to avoid protein degradation. Here we have chosen one set of conditions and worked in a small scale high-throughput fashion. As the present study has been performed in triplicates we do observe, in some cases large, variations in expression levels and solubility. The differences can in part be due to experimental variations associated with the small scale of the screening experiments. However, the overall conclusions are not affected by these variations.

Finally, we recognise that the Bradford assay as well as the qualitative SDS-PAGE method that we use for determination of concentrations both are protein dependent and that accurate measurements require calibrations for each protein [41]. Hence the concentrations are not quantitatively comparable between the different proteins. However, the protein dependence of the concentration measurements as performed here is not expected to interfere with the focus of this study, which is to compare the yields of a protein when it is produced using different expression vectors. A comparison of the Bradford and SDS-PAGE analyses reveals differences that can be attributed to the calibration issue and the fact that degradation products and other impurities affect the Bradford assay. Still, there is good agreement between the two data sets with regard to which proteins that can be purified and individual differences between the fusion proteins.

Summary

In this study, we have compared the purification yields for a set of 45 human proteins produced with four different N-termini; three with a fusion protein followed by a protease site and a His tag whereas the fourth is only the His tag. The fusion partners are removed by proteolysis before purification. In general, we observe higher yields when working with a GB1 domain + His fusion as compared to using only the His tag or the thioredoxin + His fusion, and even lower yields are seen when using a GST + His fusion. The combined effects on expression levels and solubility of the fusion protein can account for the differences in yield of purified protein. Specifically, the GB1 + His and thioredoxin + His fusion proteins have high average expression levels and the GB1 + His fusion proteins also have significantly higher average soluble expression. The influence of the fusion partners on amounts of soluble protein varies with the size of the target proteins. We also find correlations between both target protein size as well as sequence characteristics and production of soluble protein in *E. coli*. For some proteins there is a high increase of the solubility, derived from the fusion partner, but many of these target proteins fail to be purified to similar levels after removal of the fusion protein. From our set of 45 human genes and on a scale of 5 ml cultures, we were able to purify 15 proteins (33%) to yields above 40 μg and in total 29 proteins (64%) to levels detectable on silver stained gels. From our data we conclude that the solubilising fusion protein GB1 generally increases the final yield of purified protein, but that not all proteins benefit from the increase in solubility.

Acknowledgements

We would like to acknowledge the contributions of John Löfblom and Kristina Bergström for initial cloning of genes 202 to 215, Harry Brumer and Hongbin Henriksson for the mass spectrometry analysis and Susanne van den Berg for construction of the vectors pTH8 and pTH18. This work was supported by the European Commission Integrated Project SPINE (QLG2-CT-2002-00988) as part of the Framework 5 Quality of Life and Management of Living Resources Program, by the Wallenberg Consortium North (WCN) and by the Swedish foundation for strategic research (SSF).

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s10969-005-9003-7 and is accessible for authorized users.

References

1. Waugh, D.S. (2005) *Trends Biotechnol.* **23**, 316–320.
2. Hammarström, M., Hellgren, N., van den Berg, S., Berglund, H. and Härd, T. (2002) *Protein Sci.* **11**, 313–321.
3. Braun, P., Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E. and LaBaer, J. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 2654–2659.
4. Shih, Y.P., Kung, W.M., Chen, J.C., Yeh, C.H., Wang, A.H. and Wang, T.F. (2002) *Protein Sci.* **11**, 1714–1719.
5. Dyson, M.R., Shadbolt, S.P., Vincent, K.J., Perera, R.L. and McCafferty, J. (2004) *BMC Biotechnol.* **4**, 32.
6. Woestenenk, E.A., Hammarström, M., van den Berg, S., Härd, T. and Berglund, H. (2004) *J. Struct. Funct. Genom.* **5**, 217–229.
7. Smith, D.B. and Johnson, K.S. (1988) *Gene* **67**, 31–40.
8. LaVallie, E.R., DiBlasio, E.A., Kovacic, S., Grant, K.L., Schendel, P.F. and McCoy, J.M. (1993) *Biotechnology (N Y)* **11**, 187–193.
9. Huth, J.R., Bewley, C.A., Jackson, B.M., Hinnebusch, A.G., Clore, G.M. and Gronenborn, A.M. (1997) *Protein Sci.* **6**, 2359–2364.
10. Bedouelle, H. and Duplay, P. (1988) *Eur. J. Biochem.* **171**, 541–549.
11. di Guan, C., Li, P., Riggs, P.D. and Inouye, H. (1988) *Gene* **67**, 21–30.
12. Davis, G.D., Elisee, C., Newham, D.M. and Harrison, R.G. (1999) *Biotechnol. Bioeng.* **65**, 382–388.
13. Sachdev, D. and Chirgwin, J.M. (1999) *J. Protein Chem.* **18**, 127–136.
14. Nominé, Y., Ristriani, T., Laurent, C., Lefèvre, J.F., Weiss, E. and Travé, G. (2001) *Protein Exp. Purif.* **23**, 22–32.
15. Ashraf, S.S., Benson, R.E., Payne, E.S., Halbleib, C.M. and Grøn, H. (2004) *Protein Exp. Purif.* **33**, 238–245.
16. Lechner, M.S. and Laimins, L.A. (1994) *J. Virol.* **68**, 4262–4273.
17. Kapust, R.B. and Waugh, D.S. (2000) *Protein Exp. Purif.* **19**, 312–318.
18. Woestenenk, E.A., Hammarström, M., Härd, T. and Berglund, H. (2003) *Anal. Biochem.* **318**, 71–79.
19. Braun, P. and LaBaer, J. (2003) *Trends Biotechnol.* **21**, 383–388.
20. Busso, D., Kim, R. and Kim, S.H. (2004) *J. Struct. Funct. Genom.* **5**, 69–74.
21. Chambers, S.P., Austen, D.A., Fulghum, J.R. and Kim, W.M. (2004) *Protein Exp. Purif.* **36**, 40–47.
22. Savchenko, A., Yee, A., Khachatryan, A., Skarina, T., Evdokimova, E., Pavlova, M., Semesi, A., Northey, J., Beasley, S., Lan, N., Das, R., Gerstein, M., Arrowmith, C.H. and Edwards, A.M. (2003) *Proteins* **50**, 392–399.

23. Scheich, C., Sievert, V. and Büssow, K. (2003) *BMC Biotechnol*. **3**, 12.

24. Trésaugues, L., Collinet, B., Minard, P., Henckes, G., Aufrère, R., Blondeau, K., Liger, D., Zhou, C.Z., Janin, J., Van Tilbeurgh, H. and Quevillon-Cheruel, S. (2004) *J. Struct. Funct. Genom*. **5**, 195–204.

25. Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H. and Gerstein, M. (2004) *J. Mol. Biol*. **336**, 115–130.

26. Wang, W. and Malcolm, B.A. (1999) *Biotechniques* **26**, 680–682.

27. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) *Nucleic Acids Res*. **32 Database issue**, D138–D141.

28. Zhang, G., Gurtu, V. and Kain, S.R. (1996) *Biochem. Biophys. Res. Commun*. **227**, 707–711.

29. Bradford, M.M. (1976) *Anal. Biochem*. **72**, 248–254.

30. Etchegaray, J.P. and Inouye, M. (1999) *J. Bacteriol*. **181**, 5852–5854.

31. Stenström, C.M., Jin, H., Major, L.L., Tate, W.P. and Isaksson, L.A. (2001) *Gene* **263**, 273–284.

32. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M. and Arrowsmith, C.H. (2000) *Nat. Struct. Biol*. **7**, 903–909.

33. Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L.S., Miller, M.D., McPhillips, T.M., Miller, M.A., Scheibe, D., Canaves, J.M., Guda, C., Jaroszewski, L., Selby, T.L., Elsliger, M.A., Wooley, J., Taylor, S.S., Hodgson, K.O., Wilson, I.A., Schultz, P.G. and Stevens, R.C. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 11664–11669.

34. Yokoyama, S. (2003) *Curr. Opin. Chem. Biol*. **7**, 39–43.

35. Scheich, C., Leitner, D., Sievert, V., Leidert, M., Schlegel, B., Simon, B., Letunic, I., Büssow, K. and Diehl, A. (2004) *BMC Struct. Biol*. **4**, 4.

36. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) *Structure* **11**, 1453–1459.

37. Uversky, V.N., Gillespie, J.R. and Fink, A.L. (2000) *Proteins* **41**, 415–427.

38. Kapust, R.B. and Waugh, D.S. (1999) *Protein Sci*. **8**, 1668–1674.

39. Lu, Z., DiBlasio-Smith, E.A., Grant, K.L., Warne, N.W., LaVallie, E.R., Collins-Racie, L.A., Follettie, M.T., Williamson, M.J. and McCoy, J.M. (1996) *J. Biol. Chem*. **271**, 5059–5065.

40. Hirel, P.H., Schmitter, M.J., Dessen, P., Fayat, G. and Blanquet, S. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 8247–8251.

41. Read, S.M. and Northcote, D.H. (1981) *Anal. Biochem*. **116**, 53–64.