research articles Journal of proteome research

Improving Solubility of *Shewanella oneidensis* MR-1 and *Clostridium thermocellum* JW-20 Proteins Expressed into *Escherichia coli*

Irina Kataeva,*,† Jessie Chang,† Hao Xu,† Chi-Hao Luan,‡ Jizhong Zhou,§ Vladimir N. Uversky,∥,#,⊥ Dawei Lin,† Peter Horanyi,† Z. J. Liu,† Lars G. Ljungdahl,† John Rose,† Ming Luo,‡ and Bi-Cheng Wang†

Southeast Collaboratory for Structural Genomics, Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia 30602, Center for Biophysical Sciences and Engineering, Southeast Collaboratory for Structural Genomics, University of Alabama at Birmingham, Alabama 35294, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6038, Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 635 Narnhill Dr. MS 4021, Indianapolis, Indiana 46202-3763, Institute for Biological Instrumentation, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia, and Molecular Kinetics, Inc. 6201 La Pas Trail, Suite 160, Indianapolis, Indiana 46268

Received April 19, 2005

Low solubility of proteins overexpressed in *E. coli* is a frequent problem in high-throughput structural genomics. To improve solubility of proteins from mesophilic *Shewanella oneidensis* MR-1 and thermophilic *Clostridium thermocellum* JW20, an approach was attempted that included a fusion of the target protein to a maltose-binding protein (MBP) and a decrease of induction temperature. The MBP was selected as the most efficient solubilizing carrier when compared to a glutathione S-transferase and a Nus A protein. A tobacco etch virus (TEV) protease recognition site was introduced between fused proteins using a double polymerase-chain reaction and four primers. In this way, 79 *S. oneidensis* proteins have been expressed in one case with an N-terminal 30-residue tag and in another case as a fusion protein with MBP. A foreign tag might significantly affect the properties of the target polypeptide. At 37 °C and 18 °C induction temperatures, only 5 and 17 tagged proteins were soluble, respectively. In fusion with MBP 4, 34, and 38 proteins were soluble upon induction at 37°, 28°, and 18 °C, respectively. The MBP is assumed to increase stability and solubility of a target protein by changing both the mechanism and the cooperativity of folding/unfolding. The 66 *C. thermocellum* proteins were expressed as fusion proteins with MBP. Induction at 37°, 28°, and 18 °C produced 34, 57, and 60 soluble proteins, respectively. The higher solubility of *C. thermocellum* proteins in comparison with the *S. oneidensis* proteins under similar conditions of induction correlates with the thermophilicity of the host. The two-factor Wilkinson−Harrison statistical model was used to identify soluble and insoluble proteins. Theoretical and experimental data showed good agreement for *S. oneidensis* proteins; however, the model failed to identify soluble/insoluble *Clostridium* proteins. A suggestion has been made that the Wilkinson−Harrison model is not applicable to *C. thermocellum* proteins because it did not account for the peculiarities of protein sequences from thermophiles.

Keywords: gene cloning and expression • improvement of protein solubility • role of fusions on protein solubility

Introduction

A common problem in the high throughput expression of heterologous proteins in *Escherichia coli* is their low solubility. Overexpression in the cytoplasm of *E. coli* is often accompanied by misfolding and aggregation of target polypeptides. The protein cannot reach a native conformation and is partially or completely segregated into inclusion bodies.[1] Although the formation of inclusion bodies greatly increases protein stability[2] and could simplify protein purification, the recovery of insoluble protein results, in many cases, in an improperly folded polypeptide lacking biological activity. There are different approaches to improve solubility of heterologous protein. They include, but are not limited to (a) the use of promoters other than the T7 promoter, in particular, promoters activated by a temperature downshift;[3] (b) coexpression with molecular chap-

* To whom correspondence should be addressed. Department of Biochemistry and Molecular Biology, A216 Fred C. Davison Life Sciences Complex, The University of Georgia, Athens, GA 30602; E-mail: kataeva@uga.edu.

† University of Georgia.

‡ University of Alabama at Birmingham.

§ Oak Ridge National Laboratory.

∥ Indiana University School of Medicine.

Russian Academy of Sciences.

⊥ Molecular Kinetics.

10.1021/pr050108j CCC: $30.25

Published on Web 11/05/2005

erones;[4,5] (c) fusion protein technology; (d) decrease of the growth temperature and/or varied induction conditions.[6] It has been noticed that, in some cases, the fusion of a target protein to the C-terminus of a "carrier" protein can help to produce a soluble polypeptide.[1,7] Such fusion partners, as glutathione-S-transferase (GST), thioredoxin, NusA protein, maltose-binding protein (MBP), elastin-like polypeptide,[8] and others, can work very well as solubilizing agents.[7,9] One reasonable explanation of this phenomenon is that the carrier protein, with its tight and rapidly foldable structure, can possess a "priming" effect on the folding of the C-terminally fused target protein. In other words, the carrier protein folds first and promotes the adoption of the correct structure in the downstream-folding polypeptide.[10] It has also been suggested that "solubilizing" proteins may directly interact with the target protein acting as "intramolecular" chaperones.[11] Decreasing the growth or induction temperature slows down production of the target protein and therefore decreases the chance of aggregation and favors correct folding.

In the present publication, the cloning and expression of genes from two different bacteria are reported, the genomes of which have recently been sequenced. One organism is a Gram-negative facultatively anaerobic proteobacterium *Shewanella oneidensis* MR-1 associated mainly with aquatic habitats.[12] The ability to effectively reduce polyvalent metals and radionuclides, including solid-phase Fe and Mn oxides, has generated considerable interest in this organism via its potential role in the biogeochemical cycling and the bioremediation of contaminant metals and radionuclides. The other organism is an obligatory anaerobic thermophilic bacterium *Clostridium* thermocellum JW20 known for its versatile ability to decompose plant biopolymers.[13] Selected proteins from both organisms were induced at different temperatures and expressed with or without fusion with MBP. The two-factor Wilkinson−Harrison statistical model[7] was used to identify soluble and insoluble proteins in these sets. Predicted and observed solubilities of proteins from both sources were compared. The applicability of the Wilkinson−Harrison statistical model to select soluble proteins from mesophiles and thermophiles, the correlation between protein stability and solubility, and the "solubilizing" mechanism of MBP as a carrier protein are discussed.

Materials and Methods

Target Selection. From a total of 4869 *S. oneidensis* MR-1 genes and 3805 *C. thermocellum* JW20 genes, 66 and 86 protein targets, respectively, were selected for this study. The criteria for protein selection were as follows: (1) those for which no previous protein production work has been done; (2) those that had no significant similarity (BLAST, *E*-value $\leq 1 \times 10^{-4}$)[14] with any sequence in the PDB;[15] (3) those having a length between 70 and 700 residues; (4) those that have been assigned to at least one Pfam family[16] that do not yet have any structural representative; (5) those that have no more than two predicted trans-membrane fragments; and (6) those that have less than three predicted coiled- coil fragments. These proteins were further prioritized by functional annotations and protein family coverage. If there are more than three proteins represented by same Pfam family, then the proteins with the shortest length, predicted to be enzymes, and those not in a complex were selected.

Statistical Modeling. A modified Wilkinson−Harrison statistical model was used for distinguishing soluble and insoluble protein expression.[7] In contrast to the original model using five parameters to correlate with the insolubility of proteins, the revised version is based on only two critical parameters which are strongly correlated with inclusion bodies formation.[17] One parameter is an average charge which accounts for differences in the number of Asp plus Glu vs Lys plus Arg; a second parameter is the total content of the turn-forming residues which accounts for the number of Asn, Gly, Pro, and Ser. Thus, the revised solubility model involves calculation of a canonical variable (CV) or composite parameter for the protein for which the solubility is being predicted. The CV in the model is defined as follows:

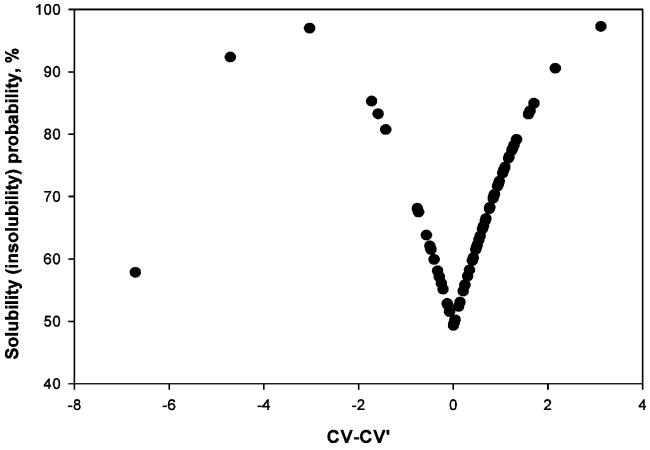

Figure 1. Relationship between protein solubility probability and the parameter CV−CV′ of the *S. oneidensis* proteins. If the CV−CV′ is negative, the protein is predicted to be soluble; if the CV−CV′ is positive, the protein is predicted to be insoluble.

$$\mathrm{CV} = \lambda_1\left(\frac{N+G+P+S}{n}\right) + \lambda_2\left|\frac{(R+K)-(D+E)}{n} - 0.03\right|$$

where n is number of residues in protein;

N, G, P, and S are the numbers of Asn, Gly, Pro, or Ser, respectively;

R, K, D, and E are the numbers of Arg, Lys, Asp, or Glu, respectively;

λ_1 and λ_2 are fixed constants (15.43 and −29.56, respectively).

The solubility probability (SP) is based on the parameter CV−CV′, where CV′ is a discriminant equal to 1.71. If CV−CV′ is positive, the protein is predicted to be insoluble; in case when CV−CV′ is negative, the protein is predicted to be soluble. The dependence of solubility/insolubility probability on CV−CV′ calculated for a group of *S. oneidensis* proteins is shown in Figure 1. The solubility probability (SP) is predicted using the following equation:[7]

$$\mathrm{SP} = 0.4934 + 0.276|\mathrm{CV} - \mathrm{CV}'| - 0.0392\ (\mathrm{CV} - \mathrm{CV}')^2$$

Cloning, Vectors and Cells Used. Multiple DNA sequences were cloned using the Gateway Cloning Technology based on specific recombination between homologous DNA sequences (Invitrogen).[18] The genomic DNA of *C. thermocellum* JW20 and entry clones of *S. oneidensis* MR-1 were used as templates for gene amplification by the polymerase chain reaction (PCR). For PCR accuracy, the high fidelity and specificity AccuPrime *Pfx* DNA polymerase (Invitrogen) was used.[19] Primers were designed using XPression Primer 3.0 software. To generate entry clones, pDONR221 (Invitrogen) was used as an entry vector. For the creation of expression clones, the five following

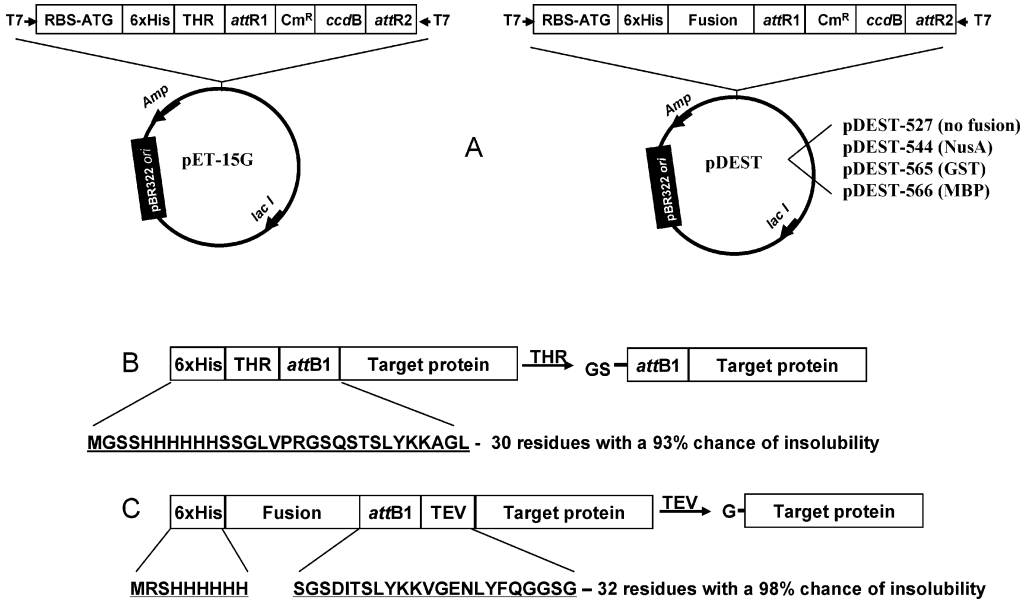

Figure 2. A, Recombination regions of the expression vectors used; B, protein expressed from pET-15G vector and cut with thrombin; C, protein expressed from pDEST-527, pDEST-544, pDEST-565, or pDEST-566 vectors and cut with TEV protease. Abbreviations: RBS, a ribosome-binding site; *att*R1, *att*R2, *att*B1, and *att*B2 are recombination sites *att*R1, *att*R2, *att*B1, and *att*B2, respectively; Cm^R, a chloramphenicol resistance gene; *ccd*B, a cytotoxicity gene B; THR, a thrombin cleavage site; MBP, a maltose-binding protein; TEV, a TEV protease cleavage site; GS and G, amino acid residues originated from THR and TEV protease sites, respectively.

expression vectors were used: pET-15G, pDEST-527, pDEST-544, pDEST-565, and pDEST-566 (Figure 2). The pDEST-527, pDEST-544, pDEST-565, and pDEST-566 expression vectors were received from Dominic Esposito (National Cancer Institute at Frederick, Maryland). The pET-15G vector was created by conversion of pET15b vector (Invitrogen) into a gateway-compatible vector using the Gateway Cloning Cassette.[20] This vector encoded, starting from a 5′-terminus, a 6xHis tag, a thrombin cleavage site and an *att*R1 recombination site. The pDEST-527 vector encoded the 6xHis tag and the *att*R1 recombination site, also starting from the 5′ terminus. Three other vectors encoded the 6xHis tag, a carrier protein, and the *att*R1 site. The carrier proteins encoded by pDEST-544, −565, and −566 were Nus A protein, GST and MBP, respectively. Recombination regions of destination vectors and the structure of the expressed proteins are shown in Figure 2A−C. For cloning of each gene into pET-15G vector, two primers were designed (Figure 3A). A forward primer contained an *att*B1 site following the 18−21 bases of gene specific sequence (GSS), *att*B1-GSS. A reversed primer was *att*B2-GSS. For cloning into other destination vectors, the principle of the adapter PCR was applied to shorten primer length and to introduce a protease cleavage site. This method utilizes four primers instead of two in two different PCRs (Figure 3B).[21] For the 1st PCR forward, tobacco etch virus(TEV)-GSS primer and reverse 1/2 *att*B2-GSS primer were used. For the 2nd PCR forward, *att*B1-TEV and reversed *att*B2 primers were used. The product of 1st PCR served as a template in the 2nd PCR.

One Shot TOP10 and BL21 Star (DE3) One Shot competent cells (Invitrogen) were used to transform BP and LR reactions, respectively.

Cell Growth and Lysis. Cultures were grown in 96-deep well plates (Qiagen) in 0.5 mL LB medium supplemented with 100 μg/mL ampicillin. For aeration, the plates were sealed with an AirPore tape sheets (Qiagen). Night cultures were started using frozen stock cultures and a 96-pin applicator (Nunc) and grown at 37 °C. In the morning, three identical 96-well plates were inoculated with 10 μL of night cultures. The plates were incubated with shaking at 37 °C for 4 h. Then the IPTG was added to a final concentration of 1 mM in each well and the plates were shaken for additional 4 h at three different temperatures: one palte at 18 °C, another plate at 28 °C, and

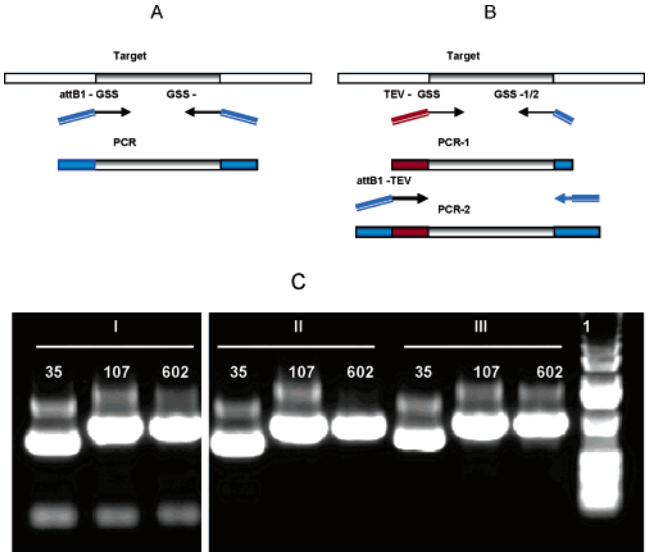

Figure 3. A, Gene amplification by one PCR using two primers; B, Gene amplification using two PCRs and four primers. Abbreviations: *att*B1 and *att*B2, recombination sites *att*B1 and *att*B2, respectively; GSS, gene-specific sequence; TEV, a TEV protease recognition site. C, Two variants of amplification of three C. *thermocellum* genes using the modified adapter PCR method. I, 1st PCR was run for 5 cycles, then another pair of primers was added and the 2nd PCR was run for 20 cycles. Rest of first pair of primers is seen on the bottom of gel. II, 1st PCR was run for 20 cycles; III, amplification product from the 1st PCR was used as a template for the 2nd PCR run for 20 cycles. The target genes used are Cth35 (474 bp), Cth107 (834 bp), and Cth602 (924 bp).

the third plate at 37 °C. To collect cells, the plates were centrifuged and dried on filter paper. For chemical lysis of a small amount of bacterial cells, a combination of mild nonionic detergent and a lysozyme was used. 200 μL of a freshly made Sigma CelLytic B Plus reagent[22,23] was added to each well. The plates were stirred for 30 min at room temperature. The lysed solution was designated as a "whole fraction". The "soluble fraction" was obtained by removal of insoluble fraction by filtration through Empore small volume filter plates (available from Fisher). The whole and the soluble fractions were used for the detection of protein expression and solubility, respectively.

Expression and Solubility Tests. To evaluate protein expression, the proteins of the whole fraction (10 μL) were separated by SDS-PAGE using Criterion 4−20% gradient gels (Bio-Rad). Proteins were stained with Coomassie brilliant blue. The soluble fraction was used to detect soluble proteins either by SDS-PAGE or by the enzyme-linked immunosorbent assay (ELISA) with penta anti-His antibodies (Qiagen) and rabbit anti-mouse IgG-alkaline phosphatase conjugate (Pierce)[24] following recommendations of the supplier.

Results

Cloning into Different Destination Vectors. Cloning into the pET-15G vector inserted a target gene between the *att*B1 and *att*B2 sites, so that the expressed protein had a 30-residue N-terminal tag (four residues from the vector, a 6×His tag, a thrombin cleavage site, and an *att*B1 site). Cleavage with thrombin resulted in a final sample of target protein still containing thirteen amino acid residues at its N-terminus (two residues of the thrombin site plus eleven residues of *att*B1 site, see Figure 2B). Proteins expressed from pET-15G were found mostly in inclusion bodies. Soluble proteins could not be crystallized or produced crystals of low diffraction quality, possibly due to the presence of the above-mentioned N-terminal tag.

To avoid the inclusion of an N-terminal tag, target genes were re-cloned, introducing the TEV protease cleavage site between the *att*B1 site and the GSS. The only way to place the cleavage site in the above position was to use a forward primer containing sequence encoding the TEV site (Figure 3B). To shorten the length of primers and, correspondingly, to decrease mistakes in their synthesis and PCR gene amplification, two pairs of primers and two PCRs were used. In the 1st PCR, gene-specific primers were used. The forward and reverse primers contained the TEV site followed by 18−21 bases of gene-specific sequence (GSS), and 1/2 *att*B1 site followed by GSS, respectively. The product of the 1st PCR was used as a template in the 2nd PCR. In this PCR, the universal primers were used; these primers did not contain any GSS and could be used with any gene. The forward universal primer contained the *att*B1 sequence followed by the TEV site sequence. The reversed universal primer contained the *att*B2 site.

Two sub-variants of the double PCR were evaluated and both found to work very well (Figure 3C). In one variant (Figure 3CI), the adapter PCR was run 5 cycles, then the universal primers were added to the PCR mixture and the 2nd PCR was run 20 more cycles. Figure 3CI shows that after addition of the secondary primers, a new template generated in the 1st PCR was used. The rest of the gene-specific primers is seen on the bottom of gel. In another variant, the 1st PCR was run 20 cycles (Figure 3CII). Then an aliquot of this PCR was added to a new 2nd PCR mixture and the PCR was run 20 cycles (Figure 3CIII). Both variants finally resulted in the same product, but the first variant saves time and reagents. The product of the 2nd PCR flanked by the *att*B1-TEV protease site and by the *att*B2 site was used in LR reaction to generate expression clones. A cut of the expressed proteins with TEV protease resulted in the presence of only one glycine residue originating from the TEV protease site at the N-terminus of the protein (Figure 2C).

Expression and Solubility of the 20 Small *C. thermocellum* JW20 Proteins. Twenty genes from *C. thermocellum* encoding small polypeptides with molecular mass range of 5.6−13.8 kDa were used to evaluate the effect of the N-terminal 29-residue tag and the three commonly used carrier proteins, NusA, GST, and MBP on solubility of the target proteins (Table 1). These carrier proteins were selected because they are significantly different in size and predicted solubility and because they were relatively large in comparison to the targets. The NusA is a 55-kDa protein with a high chance of solublility (95%); the 38.5-kDa MBP had a 55% chance to be insoluble; the 26-kDa GST had 56% chance of solubility.[7] Using a limited number of small proteins of human origin, a positive correlation has been found between the relative size of the carrier protein and the solubility of fusion protein.[7] The twenty *C. thermocellum* genes were cloned and expressed and the level of expression was relatively high. The proteins expressed with the tag were totally insoluble at 37 °C. Reduction of the induction temperature to 28° and 18 °C resulted in the appearance of two and four partially soluble proteins, respectively. The effect of GST, MBP, and NusA on the solubility of the *C. thermocellum* proteins was different. In fusion with GST, five proteins were soluble at 37 °C, five at 28 °C, and six at 18 °C. Fusion with NusA resulted in the expression of eight (at 37 °C), eleven (at 28 °C) and eleven (at 18 °C) soluble proteins. MBP solubilized twelve proteins at 37 °C, and thirteen proteins at either 28 °C or 18 °C. The total number of recovered soluble proteins was fifteen. Thus, the order of the efficiency of the carrier proteins to solubilize *C. thermocellum* passenger proteins induced at 37 °C was MBP ≥ NusA > GST. Because fusion with MBP was the most efficient in the expression of soluble polypeptides, especially in induction at 37 °C, this carrier protein was chosen for larger-scale cloning and expressions.

Expression and Solubility of the *S. oneidensis* MR-1 Proteins. The 86 genes from *S. oneidensis* encoding proteins with a molecular mass range of 6.5 to 71.4 kDa were selected (Table 2). The 79 genes were amplified and cloned using, in one case, pET-15G and in the other case, pDEST-566 expression vectors. The majority of recombinant proteins were expressed at a relatively high level. The solubility probabilities (SP) of the recombinant proteins were calculated using the modified Wilkinson−Harrison statistical model[7] and compared to the experimental data. The solubility probabilities of each (1) target sequence, (2) target sequence plus N-terminal 30-residue tag originating from pET-15G vector and (3) target sequence plus a 32-residue sequence plus maltose-binding protein originating from pDEST-566 vector (Figure 2B and C, Table 2) were calculated. On the basis of the SP values of target sequences alone, 38 potentially soluble and 41 potentially insoluble proteins have been selected. The 30-amino acid residue N-terminal tag of pET-15G vector had a 93% chance to be insoluble. Correspondingly, the SP values of all target proteins having this tag shifted to a more insoluble region. Only six proteins with this tag were predicted to be soluble and 73 proteins predicted to be insoluble. Expression of the target proteins using pDEST-566 vector resulted in the production

Table 1. Expression and Solubility[a] of Twenty *Clostridium thermocellum* JW20 Proteins without Fusion and Fused to Different Carrier Proteins and Induced at Three Different Temperatures

protein			solubility at 37 °C				solubility at 28 °C					solubility at 18 °C		
ID	size, Da	expression	no fusion	NusA	GST	MBP	no fusion	NusA	GST	MBP	no fusion	NusA	GST	MBP
80	12306.67	good[b]	no[c]	no	no	no	no	no	no	no	no	no	no	no
111	10055.52	good	no	no	no	no	no	no	no	no	no	no	no	no
171	11334.05	good	no	no	no	no	no	no	no	no	no	no	no	no
350	10022.77	good	no	yes[d]	no	no	no	yes	no	partial	no	yes	no	partial
372	13230.39	good	no	no	yes	yes	no	yes	yes	yes	partial	yes	yes	yes
483	13014.12	good	no	no	no	no	no	no	no	no	no	no	no	no
486	9523.69	good	no	yes	partial	yes	partial	yes	partial	yes	partial	yes	partial	yes
568	9349.46	good	no	no	no	partial	no	no	no	yes	no	no	no	yes
758	12500.82	good	no	no	no	partial	no	yes	no	partial	no	yes	yes	yes
1012	7812.03	good	no	yes	no	yes	no	yes	no	yes	partial	yes	no	yes
1308	8541.67	good	no	partial[e]	no	yes	no	yes	no	yes	no	yes	no	yes
1354	7662.57	good	no	yes	yes	yes	no	yes	yes	yes	no	yes	yes	yes
1583	12440.23	good	no	no	no	no	no	no	no	no	no	no	no	no
1809	5607.67	good	no	yes	partial	yes	partial	yes	partial	yes	partial	yes	yes	yes
1855	9448.1	good	no	no	no	no	no	no	no	no	no	no	no	no
2047	10454.85	good	no	no	yes	yes	no	yes	yes	yes	no	yes	yes	yes
2167	10385.21	good	no	no	no	no	no	no	no	no	no	no	no	no
2409	13827.63	good	no	no	no	partial	no	no	no	yes	no	no	no	partial
2505	8162.68	good	no	partial	no	yes	no	yes	no	yes	no	yes	no	yes
2917	12890.65	good	no	yes	no	yes	no	partial	no	yes	no	yes	no	yes
total:		20	0	8	5	12	2	11	5	13	4	11	6	13

[a] Expression and solubility of proteins was monitored by SDS-gel electrophoresis. [b] The presence of an abundant protein band of an appropriate molecular mass in the whole protein fraction was designated as "good expression" (good). A protein was designated as "insoluble" (no),[c] as "soluble" (yes),[d] and partially soluble (partial)[e] when the corresponding protein band was totally associated with insoluble protein fraction, with soluble protein fraction, or was distributed between the soluble and the insoluble protein fractions, respectively.

of polypeptides, which, in addition to the target protein, contained an MBP sequence (with a 55% chance of insolubility) and a 29-residue sequence having a 95% chance to be insoluble. MBP plus the 29-residue tag had a 60% chance to be insoluble (Table 2). Nine fusion proteins were predicted to be soluble and the rest was predicted to be insoluble. In general, the predicted solubility of the fused proteins was lower than that of the target sequences alone, but somewhat higher than that predicted for target sequences plus an insoluble 30-residue tag.

Comparison of the predicted and real solubility of proteins expressed from pET-15G vector revealed a very good correlation in a case when proteins were induced under standard conditions (at 37 °C). The insoluble proteins were predicted with relatively high accuracy (74 predicted vs 73 experimentally insoluble proteins). The prediction of solubility, however, was not as precise. In one experiment, four proteins were soluble and one was partially soluble, but the solubility of only one protein (SO0898) coincided with the prediction. The expression from the pDEST-566 vector in fusion with MBP plus a 29-residue sequence gave four soluble proteins. All four were predicted to be insoluble.

Reduction of the induction temperature increased the number of soluble and partially soluble proteins. In particular, the number of soluble proteins expressed from pET-15G vector increased to 17 when induction was performed at 18 °C. A similar tendency was observed for the expression from the pDEST-566 vector with 34 proteins being soluble already at 28 °C. Further decrease in the induction temperature to 18 °C did not significantly affect the solubility, producing only 4 more soluble proteins. In contrast to these data, the solubility of several small human proteins expressed in fusions with different carrier proteins was not affected by the induction temperature.[6]

Table 2 also illustrates the effect of molecular mass of the expressed proteins on their solubility. The values of molecular masses in Table 2 are given for target proteins only without tags and fusions. The use of the pET-15G expression vector resulted in the production of proteins with molecular mass enlarged by 3.2 kDa (the 30-residue tag value) so the molecular mass of the resulting product did not exceed 74 kDa. It is difficult to find a relationship between size and solubility of proteins expressed from pET-15G because the average solubility is low. However, when target proteins were fused with a 38 kDa MBP and a 29-residue sequence (3.2 kDa), the solubility of the expressed construct (including target sequence plus tag plus MBP sequence) was noticeably decreased when its molecular mass exceeded 60 kDa.

Expression and Solubility of the *C. thermocellum* JW20 Proteins. The 66 *C. thermocellum* genes were selected for cloning. These genes encoded proteins with a wide range of molecular masses from 8.4 to 99.5 kDa (Table 3). Preliminary prediction of solubility probability of the target sequences themselves selected 25 potentially soluble and 41 potentially insoluble proteins. The addition of the N-terminal 29-residue tag resulted in a decrease in the number of soluble proteins to 13, increasing the number of insoluble proteins to 53. To overcome the insolubility problem, and based on the data from Table 1, the target proteins were originally expressed in fusion with MBP, although this fusion did not favor solubility based on the SP values (6 potentially soluble and 60 potentially insoluble proteins, Table 3). The 64 proteins were expressed at a relatively high level. Unexpectedly, the solubility of the fused proteins was significantly higher than that predicted by the statistical model. For example, 34 fused proteins were soluble and 32 were insoluble (at 37 °C). The decrease in the induction temperature to 28 °C produced 57 soluble and 9 insoluble proteins; 60 proteins were soluble and 6 proteins were insoluble when induced at 18 °C. These data are summarized in Table 3, which shows that the statistical model used to predict target protein solubility did not work for the *C. thermocellum* proteins.

Comparison of solubility profiles for *S. oneidensis* and *C. thermocellum* overexpressed proteins (Figure 4A) indicates that the *Shewanella* proteins are mostly insoluble at 37 °C even

Table 2. Expression, Solubility Probability (SP) and Solubility of the *Shewanella oneidensis* MR-1 Proteins[a]

target ID	MW Da	SP (T)[b]	SP (T−T)[c]	expression	solubility at 37 °C	solubility at 18 °C	SP (T−T-MBP)[d]	expression	solubility at 37 °C	solubility at 28 °C	solubility at 18 °C
SO2800	6503	62% soluble	84% insoluble	good	no	no	65% insoluble	good	no	no	no
SO2358	7557	80% soluble	75% insoluble	good	no	no	53% insoluble	good	no	good	good
SO1110	8237	68% soluble	54% insoluble	good	no	good	64% insoluble	good	no	good	good
SO0335	8401	95% soluble	49% insoluble	good	no	no	52% soluble	good	no	good	good
SO0044	8720	52% insoluble	86% insoluble	good	no	partial	58% insoluble	good	no	good	good
SO1783	8776	88% insoluble	90% insoluble	good	no	no	68% insoluble	good	no	good	good
SO2092	9097	86% soluble	61% insoluble	good	no	no	50% soluble	good	no	good	good
SO1951	9641	50% insoluble	61% insoluble	good	no	no	68% insoluble	good	no	good	partial
SO2575	10452	56% insoluble	83% insoluble	good	no	no	59% insoluble	good	no	good	good
SO3439	11202	52% insoluble	83% insoluble	good	no	no	58% insoluble	good	no	no	no
SO1170	12176	83% soluble	55% insoluble	good	no	no	50% insoluble	good	no	good	good
SO1475	12363	80% insoluble	86% insoluble	good	no	no	65% insoluble	good	no	good	good
SO0299	12467	52% insoluble	82% insoluble	good	no	no	58% insoluble	good	no	no	no
SO2089	13216	94% soluble	73% insoluble	good	no	no	59% soluble	good	no	good	good
SO3429	14326	51% insoluble	78% insoluble	good	partial	good	58% insoluble	good	no	no	no
SO0923	14932	66% insoluble	85% insoluble	good	no	no	61% insoluble	good	no	good	good
SO2750	15614	60% soluble	69% imsoluble	good	no	good	52% insoluble	good	no	no	no
SO0883	15677	51% soluble	74% insoluble	good	no	no	57% insoluble	good	good	good	good
SO2263	16529	81% insoluble	92% insoluble	good	no	no	66% insoluble	good	no	no	no
SO2573	16735	83% soluble	55% insoluble	good	no	no	53% insoluble	good	no	partial	good
SO2201	16781	78% soluble	49% insoluble	good	no	no	51% insoluble	good	good	good	partial
SO3490	17637	71% soluble	57% insoluble	good	no	no	50% insoluble	good	no	good	good
SO0898	19568	92% soluble	73% solubility	good	good	good	52% soluble	good	no	good	good
SO2667	19797	62% insoluble	68% insoluble	good	no	no	71% insoluble	no	no	no	no
SO1830	19970	60% soluble	66% insoluble	good	no	no	54% insoluble	good	no	no	no
SO1518	20071	58% soluble	65% insoluble	good	no	good	54% insoluble	good	no	good	good
SO2840	20134	78% insoluble	81% insolublle	good	no	good	70% insoluble	good	no	no	partial
SO3668	20917	60% insoluble	75% insoluble	good	no	no	59% insoluble	good	good	good	partial
SO0036	20936	84% insoluble	92% insoluble	good	no	no	69% insoluble	partial	no	partial	partial
SO1672	22319	50% insoluble	71% insoluble	good	no	no	57% insoluble	good	no	partial	good
SO2326	22653	53% insoluble	73% insoluble	good	no	good	58% insoluble	good	no	no	partial
SO2512	23413	61% soluble	60% insoluble	good	no	no	52% insoluble	good	no	good	good
SO2050	23890	54% insoluble	72% insoluble	good	no	no	59% insoluble	good	no	good	good
SO1867	24784	53% soluble	66% insoluble	good	no	no	55% insoluble	good	no	partial	good
SO2751	25103	73% insoluble	84% insoluble	good	no	no	65% insoluble	good	no	no	no
SO3770	25519	81% soluble	62% insoluble	good	no	no	57% insoluble	no	no	no	no
SO1650	26038	50% soluble	68% insoluble	good	no	no	56% insoluble	good	no	no	no
SO1350	26041	83% insoluble	90% insoluble	good	no	no	70% insoluble	good	no	no	no
SO0875	26042	68% insoluble	81% insoluble	good	no	no	63% insoluble	partial	partial	partial	good
SO2039	26147	67% soluble	53% insoluble	good	no	no	50% soluble	good	no	partial	no
SO2948	27128	52% soluble	66% insoluble	good	no	good	55% insoluble	good	no	no	no
SO2352	28616	51% soluble	65% insoluble	partial	no	good	55% insoluble	good	no	good	good
SO3540	28798	60% insoluble	74% insoluble	good	no	good	60% insoluble	good	no	good	good
SO1090	29995	83% insoluble	83% insoluble	good	partial	partial	71% insoluble	good	no	no	no
SO1788	30344	75% soluble	57% soluble	good	no	no	55% soluble	good	no	no	no
SO1165	31350	92% insoluble	92% insoluble	good	no	no	85% insoluble	good	no	no	no
SO1248	32122	61% insoluble	74% insoluble	good	no	good	60% insoluble	good	no	good	good
SO1091	32179	83% insoluble	84% insoluble	no	no	no	71% insoluble	no	no	no	no
SO2351	33443	75% insoluble	83% insoluble	good	no	no	67% insoluble	good	no	no	no
SO1556	33649	67% soluble	51% soluble	good	no	no	60% insoluble	good	no	partial	good
SO3701	33799	77% insoluble	85% insoluble	good	no	good	68% insoluble	good	no	partial	no
SO0602	34646	61% soluble	54% insoluble	good	no	no	50% insoluble	no	no	no	no
SO3529	34718	53% soluble	61% insoluble	good	no	no	54% insoluble	good	no	good	partial
SO1344	35022	54% soluble	61% insoluble	good	no	no	53% insoluble	good	no	no	no
SO2049	35743	69% soluble	54% soluble	good	no	no	51% insoluble	good	no	no	no
SO0304	36350	50% soluble	62% insoluble	good	no	no	55% insoluble	good	no	no	no
SO0569	36734	61% insoluble	73% insoluble	good	no	no	60% insoluble	good	no	no	no
SO1583	36857	81% insoluble	87% insoluble	good	no	no	71% insoluble	good	no	no	no
SO1249	37007	65% insoluble	75% insoluble	good	no	no	62% insoluble	good	no	no	no
SO0471	37989	50% insoluble	63% insoluble	good	no	no	55% insoluble	good	no	no	partial
SO2177	38148	69% soluble	55% soluble	no	no	no	51% soluble	good	no	no	no
SO2342	38708	55% soluble	58% insoluble	good	no	partial	53% insoluble	good	no	partial	partial
SO2338	38783	55% insoluble	70% insoluble	good	no	no	63% insoluble	no	no	no	no
SO0332	39744	82% soluble	70% soluble	good	no	partial	62% soluble	good	no	no	partial
SO1313	39810	62% insoluble	72% insoluble	good	no	no	61% insoluble	no	no	no	no
SO3756	40289	62% insoluble	72% insoluble	good	no	no	61% insoluble	good	no	no	no
SO0054	43159	70% insoluble	78% insoluble	good	no	no	65% insoluble	good	no	no	no
SO1403	44258	63% insoluble	72% insoluble	partial	no	no	61% insoluble	good	no	no	no
SO1811	44546	58% insoluble	66% insoluble	good	no	no	58% insoluble	good	no	no	no
SO1774	45441	65% insoluble	74% insoluble	good	partial	no	63% insoluble	good	no	no	no
SO1252	50888	56% insoluble	65% insoluble	good	no	no	57% insoluble	good	no	no	no
SO2693	53609	58% insoluble	66% insoluble	good	no	no	59% insoluble	good	no	no	no
SO0330	54450	57% soluble	52% insoluble	good	no	no	50% insoluble	good	no	partial	partial
SO1810	54486	62% insoluble	70% insoluble	good	no	no	61% insoluble	good	no	no	no
SO0506	54965	57% soluble	66% insoluble	partial	no	no	58% insoluble	good	no	no	no
SO2680	57721	71% soluble	62% soluble	good	no	no	58% soluble	good	no	no	partial
SO2487	64608	69% soluble	75% insoluble	good	good	good	67% insoluble	good	no	no	partial
SO2420	67004	55% soluble	63% insoluble	good	no	good	57% insoluble	good	no	no	no
SO2507	71407	56% soluble	63% insoluble	good	no	no	58% insoluble	no	no	no	no

[a] Expression and solubility of proteins were monitored by SDS-gel electrophoresis and ELISA, respectively. [b] SP of target sequence. [c] SP of tag-target sequence. [d] SP of tag-MBP-target sequence.

Table 3. Expression and Solubility of the *Clostridium thermocellum* JW20 Proteins[a]

target ID	MW Da	SP (T)[b]	SP (T−T)[c]	SP (T−T-MBP)[d]	expression	solubility at 37 °C	solubility at 28 °C	solubility at 18 °C
2690	8381	85% soluble	52% insoluble	75% insoluble	partial	partial	partial	good
3052	9729	97% soluble	97% soluble	71% soluble	good	no	partial	no
3543	10353	91% soluble	73% soluble	52% soluble	good	good	good	good
2112	10369	62% soluble	57% insoluble	56% insoluble	good	good	good	good
3414	10468	64% soluble	55% insoluble	55% insoluble	good	partial	partial	partial
3433	10965	62% soluble	62% inslouble	56% insoluble	good	no	partial	partial
1963	11338	60% soluble	67% insoluble	68% insoluble	good	no	good	partial
1952	11502	58% soluble	95% insoluble	56% insoluble	good	good	good	good
2241	13209	87% soluble	71% soluble	52% soluble	good	partial	good	partial
2878	13342	97% soluble	87% soluble	59% soluble	good	no	good	good
3220	16382	57% soluble	56% insoluble	55% insoluble	good	no	no	no
1613	16493	68% insoluble	75% insoluble	62% insoluble	no	no	no	partial
246	16937	72% insoluble	77% insoluble	63% insoluble	good	partial	good	good
2430	17048	63% insoluble	71% insoluble	61% insoluble	good	no	no	partial
523	17629	64% insoluble	71% insoluble	61% insoluble	good	partial	partial	partial
1898	18058	68% soluble	55% soluble	51% insoluble	good	partial	partial	good
1125	18068	53% insoluble	63% insoluble	58% insoluble	good	good	good	good
88	18189	62% insoluble	73% insoluble	62% insoluble	good	no	good	good
367	18615	53% soluble	64% insoluble	76% insoluble	good	partial	good	good
464	19037	83% soluble	72% soluble	55% soluble	good	good	good	good
72	19828	68% soluble	56% soluble	51% insoluble	good	no	partial	partial
3274	20048	68% soluble	56% soluble	51% insoluble	good	partial	partial	partial
837	20496	84% insoluble	86% insoluble	69% insoluble	good	partial	partial	partial
3617	21950	84% insoluble	86% insoluble	69% insoluble	good	no	partial	good
403	22462	76% insoluble	80% insoluble	66% insoluble	good	no	no	good
3456	23122	65% insoluble	75% insoluble	72% insoluble	good	good	good	good
1286	23757	55% insoluble	62% insloube	58% insoluble	good	no	no	no
532	24160	52% soluble	56% insoluble	56% insoluble	good	good	good	good
3040	24548	49% soluble	58% insoluble	56% insoluble	good	no	partial	no
2659	24852	60% soluble	50% soluble	52% insoluble	partial	no	partial	partial
1271	25423	60% insoluble	66% insoluble	60% insoluble	good	partial	partial	partial
2743	26296	84% insoluble	86% insoluble	60% insoluble	good	no	good	good
3014	26522	49% insoluble	57% insoluble	56% insoluble	good	good	good	good
2476	26709	65% insoluble	70% insoluble	62% insoluble	good	good	good	good
3031	26718	59% insoluble	65% insoluble	59% insoluble	good	partial	partial	partial
2847	26817	62% insoluble	69% insoluble	61% insoluble	good	no	partial	good
2412	27514	62% insoluble	68% insoluble	61% insoluble	good	good	good	good
294	28079	74% insoluble	77% insoluble	66% insoluble	good	no	good	good
3373	29341	54% insoluble	61% insoluble	56% insoluble	good	good	good	good
1625	30731	74% insoluble	77% insoluble	66% insoluble	no	no	partial	good
1887	30881	70% insoluble	74% insoluble	64% insoluble	good	partial	good	good
1272	30896	97% insoluble	97% insoluble	83% insoluble	good	partial	partial	good
1649	31177	81% soluble	74% soluble	59% soluble	no	no	no	no
1005	31701	75% insoluble	80% insoluble	66% insoluble	good	partial	partial	good
1746	32621	77% insoluble	80% insoluble	68% insoluble	good	good	good	good
2642	33543	91% insoluble	91% insoluble	76% insoluble	good	no	partial	good
3542	34067	81% soluble	74% soluble	59% soluble	good	partial	good	good
214	34268	57% soluble	50% soluble	52% insoluble	good	good	good	good
1960	35004	56% soluble	50% soluble	52% insoluble	good	no	good	good
3164	35705	78% insoluble	82% insoluble	60% insoluble	good	no	no	good
1965	36855	55% soluble	50% insoluble	55% insoluble	good	no	no	good
1509	37910	60% insoluble	65% insoluble	60% insoluble	good	no	no	no
1266	38182	82% insoluble	83% insoluble	71% insoluble	good	no	good	good
756	43528	62% insoluble	66% insoluble	61% insoluble	good	no	partial	partial
1794	43866	85% insoluble	86% insoluble	74% insoluble	good	no	partial	good
1767	48150	62% insoluble	66% insoluble	61% insoluble	good	no	good	good
2327	53578	76% insoluble	78% insoluble	69% insoluble	good	good	partial	partial
1918	55411	73% insoluble	95% insoluble	67% insoluble	partial	no	partial	partial
894	55838	62% insoluble	64% insoluble	61% insoluble	good	no	partial	good
147	58880	78% insoluble	80% insoluble	71% insoluble	good	good	good	good
635	59609	65% insoluble	68% insoluble	63% insoluble	good	no	partial	good
2021	62406	72% insoluble	74% insoluble	62% insoluble	good	good	good	good
2682	62771	70% insoluble	72% insoluble	66% insoluble	good	partial	good	good
1938	70677	56% insoluble	59% insoluble	57% insoluble	good	no	good	good
2694	70782	67% insoluble	68% insoluble	64% insoluble	good	partial	good	good
2905	74274	53% soluble	49% insoluble	51% insoluble	good	good	partial	good

[a] Expression and solubility of proteins were monitored by SDS-gel electrophoresis and ELISA, respectively. [b] SP of target sequence. [c] SP of tag-target sequence. [d] SP of tag-MBP-target sequence.

being fused with MBP whereas the solubility of MBP-fused *Clostridium* proteins is much higher at this temperature (8% versus 65%). Reducing the induction temperature positively affected the solubility of proteins from both species. However, even when induced at 28 °C and 18 °C, the solubility of the *S. oneidensis* proteins was still lower (47% and 51%, respectively)

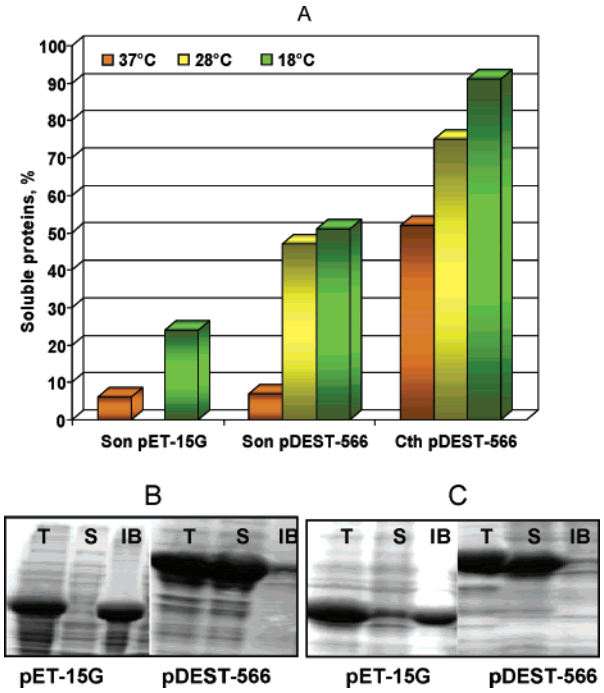

Figure 4. A, Solubility of *S. oneidensis* and *C. thermocellum* proteins expressed at different temperatures. Abbreviations: Son, *Shewanella oneidensis*; Cth, *Clostridium thermocellum*. B, different solubility of two *S. oneidensis* proteins Son107 (B) and Son602 (C) expressed into *E. coli* from pET-15G vector (no fusion), and from pDEST-566 vector (fusion with MBP). Abbreviations: T, total cell fraction; S, soluble cell fraction; IB, inclusion bodies fraction.

than that of the *C. thermocellum* proteins (88% and 92%, respectively). A reasonable explanation for this difference can be proposed based on the difference in the habitat of the two bacteria. In fact, *S. oneidensis* with its optimum growth temperature (T_{opt}) of 30 °C is located between psychrotrophes (T_{opt} 25 °C) and mesophiles (T_{opt} 37 °C), whereas *C. thermocellum* is a typical thermophilic bacterium with T_{opt} 65 °C. Because proteins from thermophiles are better adapted to the extreme conditions, they evolved to keep native structures at higher environmental temperature than proteins from mesophilic organisms. A strong positive correlation between thermostability and solubility of overexpressed proteins has been reported.[25]

Discussion

Several factors at the sequence and structure levels have been proposed to contribute to the formation of inclusion bodies upon protein overexpression in *E. coli*.[25,26] The expression of prokaryotic targets seems to be easier because it is not complicated by such factors as a codon bias and post-translational modification. The local concentration of protein, increased aggregation due to the limited solubility, existence and half-life of partially folded intermediates, interaction with molecular chaperones and the presence of N- or C-terminal tags/fusions are the most probable factors affecting the expression of prokaryotic proteins in soluble form.[25,27,28] The low solubility of overexpressed proteins is a serious problem of the high-throughput protein production. Thus, the design of a proper model to predict soluble vs insoluble proteins is one of the most important needs of structural genomics. It is obvious that, due to the high protein diversity, the above-mentioned factors will contribute to the solubility of target polypeptides to different extents; i.e., under such conditions, any model can give only a relative probability of the solubility of a given list of proteins.[28] In the present paper, a two-parameter model has been used to predict the propensity of proteins to be soluble or insoluble,[7] mainly because of its simplicity. The model takes into account the relative content of positively and negatively charged residues and the mole fraction of turn-forming residues. Both increments have been shown to play an important role in protein folding and solubility.[29] This model has been applied only to a limited number of mainly human proteins.[7,17] In the present study, we evaluated the relevance of the Wilkinson-Harrison model to predict solubility of recombinant proteins containing long tags, fusions, and originating from meso- and thermophilic bacteria.

Recombinant polypeptides usually contain affinity tag(s) for easy detection and purification. The Gateway cloning technology (Invitrogen) based on a specific recombination between homologous DNA fragments has many advantages for use in structural genomics projects. The disadvantage of this system is the presence of an additional eleven-residue recombination site between the target sequence and the affinity tag either at the N- or C-terminus. To avoid a negative effect of the tag on protein crystallization, a proteolytic site is often introduced between the target protein sequence and the recombination site making the tag even longer. The presence of the tag confers new properties on the polypeptide interfering with its folding and solubility. The negative effect of a 6xHis tag on the solubility of expressed proteins has already been reported.[21] It has also been shown using chimeric polypeptides that the artificial N-terminal sequences are very important for solubility of polypeptides.[30,31] A similar effect has been observed upon expression of polypeptides with and without hydrophobic signal peptides.[32]

We used two similar N-terminal tags upon cloning target genes into pET-15G or pDEST-527 expression vectors. The pET-15G encoded a 30-residues tag composed starting from an N-terminus of a 6×His, a thrombin cleavage site, and an attB1 recombination site (Figure 2B). The pDEST-527 encoded a 32-residues tag contained starting also from the N-terminus, the 6×His followed by the attB1 and the TEV protease cleavage site (Figure 2C). Both tags contained many turn-forming residues and were probably composed of a random coil. However, besides the 6×His, the TEV protease (or thrombin) cleavage sites, and the attB1 site are natural protein fragments, i.e., it should not be excluded that the tags possess some simple structural elements. Random coil motif is known to be present in many proteins,[33] and in some cases the random coil is associated with protein terminus.[34] According to recent studies, the random coil state is less "random" than it was suggested before, being characterized by the presence of larger or smaller amount of residual or partially folded structure (http://www.sanger.ac.uk/Users/sgj/thesis/html/node5.html). There are many indications that random coil can spontaneously form β-sheets and provoke aggregation.[35−37] Elimination of partially folded state might increase protein stability.[38] Thus, random coil and/or partially folded conformations might be responsible for protein instability and aggregation.

In the fusion expression vectors (pDEST-544, -565, and -566), a carrier protein sequence was present between the 6×His and the attB1 (Figure 2A); i.e., the tag was split into two fragments, the external N-terminal 6×His tag and 2/3 rest of the tag

located internally. In other words, the artificial tags presumably composed of random coil might be involved in protein folding similarly to native protein termini. An artificial internal sequence composed of random coil is comparable to native partially ordered regions (e.g., loops). Correspondingly, the propensity of the tagged protein to be soluble might be predicted by the model based on the content of turn-forming residues developed for regular globular proteins. SP estimations for the above two tags showed that both of them had a very high chance to be insoluble. The 30-residue tag had 93% chance of insolubility, and the 32-resides tag had 98% propensity to be insoluble. Correspondingly, the solubility probabilities of target proteins with one of these tags were lower in comparison with that of target sequences alone. This observation implies that upon selection of putative soluble polypeptides, it is important to consider all foreign residues accompanying target sequences.

We also applied this model to evaluate the solubility of fusions. The whole polypeptide sequence including the carrier, the passenger and the tag was analyzed. We used proteins with molecular masses in most cases exceeding 10 kDa. Polypeptide chains beyond 50−100 amino acid residues tend to form domains with relatively independent fold.[39] The original Wilkinson−Harrison model[17] was designed using 81 proteins only five of which contained less than 100 residues, and 69 proteins contained more than 150 residues; i.e., had a high chance to contain domains. Fusion of the carrier and the passenger is very similar to the combination of the domains in one polypeptide.

As pointed above, the Wilkinson−Harrison model was used to analyze mostly human proteins.[7,17] In the present manuscript, we expanded usage of the model applying it to predict solubility of prokaryotic proteins of both mesophilic (*S. oneidensis*) and thermophilic (*C. thermocellum*) bacteria.

On the basis of the results of such analysis, the model seemed to work better in prediction of insoluble proteins as mentioned before.[6,7] Analysis of the data presented here showed that the model worked very well to predict the propensity of *S. oneidensis* proteins to be insoluble. Unexpectedly, the model was not useful for the selection of the *C. thermocellum* proteins, which were much more soluble than predicted.

To explain the observed difference in the reliability of the statistical model to predict soluble and insoluble proteins from *S. oneidensis* and *C. thermocellum*, the sequence peculiarities of mesophilic and thermostable proteins have to be compared. It is known that the thermophilic polypeptides in comparison with their mesophilic homologues possess increased amount of salt bridges and side chain-side chain hydrogen bonds. Furthermore, Arg and Tyr are significantly more abundant, whereas Cys and Ser are less frequent in thermophilic proteins.[26,40,41] Proteins from meso- and thermophiles also significantly differ by their aliphatic index, which is directly related to the content of aliphatic residues Ala, Ile, Leu, and Val.[42,43] At the level of secondary structure, the thermophilic proteins are characterized by a greater amount of α-helical structure, avoiding Pro in their α-helices to a greater extent than mesophilic proteins. Overall, evolutionary pressure led to the increased internal hydrophobicity and increased external polarity of thermophilic proteins.[44] The modified Wilkinson−Harrison statistical model used in the present study to distinguish soluble and insoluble proteins only takes into account average charge and content of turn-forming residues. Difference in the hydrophobic residue content is not considered in this model, which might be a reason for poor prediction of solubility of *C. thermocellum* (and probably other thermostable) proteins.

What is the role of MBP as a carrier protein? It has been suggested that it can serve as an artificial chaperone but the mechanism of this effect has not been discussed.[45,46] The MBP is the most popular carrier to improve stability and solubility of a passenger protein although, according to the two-factor Wilkinson-Harrison statistical model, it has a 55% chance to be insoluble.[7] Fused with the *S. oneidensis* proteins, the MBP significantly increased their solubility at lower induction temperature in comparison to unfused variants. This effect cannot be explained only by a decrease in the local protein concentration at lower induction temperature. Compared to such carrier proteins as GST, NusA, and thioredoxin, MBP, although usually being more efficient in the solubilization of targets, has less propensity to be soluble and has a moderate molecular mass. This protein is composed of two domains and possesses a two-state reversible thermal or guanidine chloride unfolding.[45,47] In other words, the unfolding of MBP is cooperative and does not involve formation of long-living intermediate(s). It has been found that a lifetime of partially unfolded intermediates strongly influences the propensity of proteins to aggregate, probably by exhausting molecular chaperones.[25,28] The fusion with MBP is assumed to change the folding mechanism of the passenger protein so that it achieves its native conformation in a limited time and by a simpler mechanism. The decrease in the stability of MBP-protein fusions upon mutation of residues located in the interface supports this point.[46] Selection of an efficient stabilizing carrier protein based on such factor as and mechanism of folding/unfolding might be a promising approach to the expression of target proteins in soluble form.

Conclusions

Expression of proteins in fusion with carrier proteins at low induction temperature is an effective approach to increase solubility of recombinant proteins.

The revised Wilkinson−Harrison statistical model used to select soluble vs insoluble proteins is more efficient in distinguishing insoluble proteins. The use of this model might be limited by thermophilicity of the host organism.

The presence of tags and fusions significantly affect structure and solubility of target proteins and must be taken into account upon prediction of protein solubility.

Thermostable proteins from *C. thermocellum* are significantly more soluble at each induction temperature than proteins from mesophilic *S. oneidensis*, indicating a positive correlation between protein thermostability and solubility.

Acknowledgment. This work was supported in part with funds from the National Institute of Health (GM62407), The Georgia Research Alliance, and The University of Georgia Research Foundation. We thank Dominic Esposito (National Cancer Institute, Frederick, Maryland) for the expression vectors used in this study and for constant interest to our work.

References

(1) Baneyx, F.; Mujacic, M. Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.* **2004**, *22*, 1399−1408.
(2) Wei, G.; Tang, J. C. Formation of inclusion bodies may be the key factor for the stability of expressed products in *E. coli*. *Biochem. Mol. Biol.* **1995**, *37*, 895−911.

(3) Qing, G.; Ma, L.-C.; Khorchid, A.; Swapna, G. V. T.; Mal, T. K.; Takayama, M. M.; Xia, B.; Phadtare, S.; Ke, H.; Acton, T.; Montelione, G. T.; Ikura, M.; Inouye, M. Cold-shock induced high-yield protein production in *Escherichia coli*. *Nat. Biotechnol.* **2004**, *22*, 877−882.
(4) Thomas, J. G.; Ayling, A.; Baneyx, F. Molecular Chaperones, Folding Catalysts, and the Recovery of Active Recombinant Proteins from *E.coli*. *Appl. Biochem. Biotech.* **1997**, *66*, 197−238.
(5) Nishihara, K. Chaperone Coexpression Plasmids: Differential and Synergistic Roles of DnaK-DnaJ- GroE and GroEL-GroES in Assisting Folding of an Allergen of Japanese Cedar Pollen, Cryj2, in *Eschericia coli*. *Appl. Environ. Microbiol.* **1998**, *64*, 1694−1699.
(6) Hammarström, M.; Hellgren, N.; van Den Berg, S.; Berglund, H.; Hard, T. Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* **2002**, *11*, 313−321.
(7) Davis, G. D.; Elisee, C.; Newham, D. M.; Harrison, R. G. New Fusion Protein Systems Designed to Give Soluble Expression in *Escherichia coli*. *Biotechnol. Bioeng.* **1999**, *65*, 382−388.
(8) Trabbic-Carlson, K.; Liu, L.; Kim, B.; Chilkoti, A. Expression and purification of recombinant proteins from *Escherichia coli*: Comparison of an elastin-like polypeptide fusion with an oligohistidine fusion. *Protein Sci.* **2004**, *13*, 3274−3284.
(9) Terpe, K. Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **2003**, *60*, 523−533.
(10) Malakhov, M. P.; Mattern, M. R.; Malakhiva, O. A.; Drinker, M.; Weeks, S. D.; Butt, T. R. 2004 SUMO fusions and SUMO-specific protease fpr efficient expression and purification of proteins. *J. Struct. Function. Genomics* **2004**, *5*, 75−86.
(11) Shinde, U.; Inouye, M. Propeptide-mediated folding in subtilisin: the intramolecular chaperone concept. *Adv. Exp. Med. Biol.* **1996**, *379*, 147−154.
(12) Venkateswaran, K.; Moser, D. P.; Dollhopf, M. E.; Lies, D. P.; Saffarini, D. A.; MacGregor, B. J.; Ringelberg, D. B.; White, D. C.; Nishijima, M.; Sano, H.; Burghardt, J.; Stackebrandt, E.; Nealson, K. H. Polyphasic Taxonomy of the Genus *Shewanella* and Description of *Shewanella oneidensis* sp. nov. *Int. J. Syst. Bacteriol.* **1999**, *49*, 705−724.
(13) Wiegel, J.; Dykstra, M. *Clostridium thermocellum*: adhesion and sporulation while adhered to cellulose and hemicellulose. *Appl. Microbiol. Biotechnol.* **1984**, *20*, 59−65.
(14) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI−BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235- 242.
(16) Sonnhammer, E. L.; Eddy, S. R.; Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **1997**, *28*, 405−420.
(17) Wilkinson, D. L.; Harrison, R. G. Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology* **1991**, *9*, 443−448.
(18) Hartley, J. L.; Temple, G. F.; Brasch, M. A. DNA Cloning Using *in vitro* Site-Specific Recombination. *Genome Res.* **2000**, *10*, 1788−1795.
(19) Takagi, M.; Nishioka, M.; Kakihara, H.; Kitabayashi, M.; Inoue, H.; Kawakami, B.; Oka, M.; Imanaka, T. Characterization of DNA polymerase from *Pyrococcus* sp. strain KOD1 and its application to PCR. *Appl. Environ. Microbiol.* **1997**, *63*, 4504−4510.
(20) Luan, C. H.; Qiu, S.; Finley, J. B.; Carson, M.; Gray, R. J.; Huang, W.; Johnson, D.; Tsao, J.; Reboul, J.; Vaglio, P.; Hill, D. E.; Vidal, M.; Delucas, L. J.; Luo, M. High-throughput expression of *C. elegans* proteins. *Genome Res.* **2004**, *14*, 2102−1210.
(21) Woestenenk, E. A.; Hammarström, M.; van den Berg, S.; Härd, T.; Berglund, H. His tag effect on solubility of human proteins produced in *Escherichia coli*: a comparison between four expression vectors. J. *Struct. Funct. Genomics* **2004**, *5*, 217−229.
(22) Expert-Bezancon, N.; Rabilloud, T.; Vuillard, L.; Goldberg, M. E. Physical-chemical features of nondetergent sulfobetaines active as protein-folding helpers. *Biophys. Chem.* **2003**, *100*, 469−479.
(23) Vonrhein, C.; Schmidt, U.; Ziegler, G. A.; Schweiger, S.; Hanukoglu, I.; Schulz, G. E. Chaperone-assisted expression of authentic bovine adrenodoxin reductase in *Escherichia coli*. *FEBS Lett.* **1999**, *443*, 167−169.
(24) Steinert, K.; Wulbeck, M.; Ribbe, J. Immunoprecipitation with Penta-His antibody. *QIAGEN News* **1998**, *2*, 20.
(25) Idicula-Thomas, S.; Balaji, P. V. Understanding the relationship between the primary structure of proteins and the propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci.* **2005**, *14*, 582−592.
(26) Kumar, S.; Tsai, C.-J.; Russinov, R. Factors enhancing protein thermostability. *Prot. Eng.* **2000**, *13*, 179−191.
(27) McHugh, C. A.; Tammariello, R. F.; Millard, C. B.; Carra, J. H. Improved stability of a protein vaccine through elimination of a partially unfolded state. *Protein Sci.* **2004**, *13*, 2736−2743.
(28) Zbilut, J. P.; Giuliani, A.; Colosimo, A.; Mitchell, J. C.; Colafranceschi, M.; Marwan, N.; Charles, L.; Webber, C. L., Jr.; Uversky, V. N. Charge and Hydrophobicity Patterning along the Sequence Predicts the Folding Mechanism and Aggregation of Proteins: A Computational Approach. *J. Proteome Res.* **2004**, *3*, 1243−1253.
(29) Kumar, S.; Nussinov, R. How do thermophilic proteins deal with heat? *Cell. Mol. Life Sci.* **2001**, *8*, 1216−1233.
(30) Ni, L.; Zhou, J.; Hurley, T. D.; Weiner, H. Human liver mitochondrial aldehyde dehydrogenase: Three-dimensional structure and the restoration of solubility and activity of chimeric forms. *Protein Sci.* **1999**, *8*, 2784−2790.
(31) Takano, K.; Tsuchimori, K.; Yamagata, Y.: Yutani, K. Effect of foreign N-terminal residues on the conformational stability of human lysozyme. *Eur. J. Biochem.* **1999**, *266*, 675−682.
(32) Beena, K.; Udgaonkar, J. B.; Varadarajan, R. Effect of signal peptide on the stability and folding kinetics of maltose-binding protein. *Biochemistry* **2004**, *43*, 3608−3619.
(33) Odgren, P. R.; Harvie, L. W., Jr.; Fey, E. G. Phylogenetic occurrence of coiled coil proteins: Implications for tissue structure in metazoa via a coiled coil tissue matrix. *Proteins: Struct., Funct., Genet.* **1998**, *24*, 467−484.
(34) Wallace, B. A.; Kohl, N. The C-terminus of bacteriorodopsin is a random coil. *BBA* **1984**, *17*, 93−98.
(35) von Bergen, M.; Barghorn, S.; Biernat, J.; Mandelkow, E.-M.; Mandelkow, E. Tau aggregation is driven by a transition from random coil to beta sheet structure. *Biochim. Biophys. Acta* **2005**, *1739*, 158−166.
(36) Nguyen, H. D.; Hall, C. K. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *PNAS* **2004**, *101*, 16180−16185.
(37) Uversky, V. N.; Fink, A. L. Conformational constraints for the amyloid fibrillation: The importance of being unfolded. *Biochim. Biophys. Acta* **2004**, *1698*, 131−153.
(38) McHugh, C. A.; Tammariello, R. F.; Millard, C. B.; Carra, J. H. Improved stability of a protein vaccine through elimination of a partially unfolded state. *Protein Science* **2004**, *13*, 2736−2743.
(39) Janin, J.; Wodak, S. J. Structural domains in protein and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol.* **1983**, *42*, 21−78.
(40) Baker, P. J. From hyperthermophiles to psychrophiles: the structural basis of temperature stability of the amino acid dehydrogenases. *Biochem. Soc. Trans.* **2004**, *32*, 264−268.
(41) Scandurra, R.; Cousalvi, V.; Chiaraluce, R.; Politi, L.; Engel, P. C. Protein stability in extremophilic archaea. *Front. Biosci.* **2000**, *5*, 787−795.
(42) Ikai, A. Thermostability and aliphatic index of globular proteins. *J. Biochem.* **1980**, *88*, 1895−1898.
(43) Lu, B.; Wang, G.; Huang, P. A comparison of amino acid composition of proteins from thermophiles and mesophiles. *Wei Sheng Wu Xue Bao* **1998**, *38*, 20−25.
(44) Merkler, D. J.; Farrington, G. K.; Wedler, F. C. Protein thermostability. Correlations between calculated microscopic parameters and growth temperatures for closely related thermophilic and mesophilic bacilli. *Int. J. Pept. Res.* **1981**, *18*, 430−442.
(45) Bach, H.; Mazor, Y.; Shaky, S.; Berdichevsky, A. S.-L. Y.; Gutnick, D. L.; Benhar, I. *Escherichia coli* maltose-binding protein as a molecular chaperone for recombinant intracellular cytoplasmic single-chain antibodies. *J. Mol. Biol.* **2001**, *312*, 79−93.
(46) Fox, J. D.; Kapust, R. B.; Waugh, D. S. Single amino acid substitutions on the surface of *Escherichia coli* maltose-binding protein can have a profound impact on the solubility of fusion proteins. *Protein Sci.* **2001**, *10*, 622−530.
(47) Ganesh, C.; Shah, A. N.; Swaminathan, C. P.; Surolia, A.; Varadarajan, R. Thermodynamic characterization of the reversible, two-state unfolding of maltose-binding protein, a large two-domain protein. *Biochemistry* **1997**, *36*, 5020−5028.

PR050108J