Mol. Cells *30*, 127-135, August 31, 2010
DOI/10.1007/s10059-010-0097-9

Molecules and Cells

N-Terminal pI Determines the Solubility of a Recombinant Protein Lacking an Internal Transmembrane-like Domain in *E. coli*

Sang Jun Lee*, Yun Hee Han, Young Ok Kim, Bo Hye Nam, Hee Jeong Kong, and Kyung Kil Kim

We examined whether the isoelectric point (pI) of the N-terminal region of the recombinant protein 7xMefp1 acts as a universal index for expression of the protein in soluble form in *E. coli*. Expression analysis of 7xMefp1 fused to various N-terminal sequences with pI values ranging from 2.73 to 13.35 yielded three pI range-specific curves (acidic, neutral, and alkaline curves at pI 2.73-3.25, 4.61-9.58, and 9.90-13.35, respectively) for soluble expression (by facilitated diffusion) as a proportion of total protein. For neutral N-termini (pI 4.61-9.58), the total amount of rMefp1 expressed was minimally affected by ΔG_{RNA} for unfolding the mRNA secondary structure. The highly hydrophilic nature of longer N-terminal sequences with strongly acidic and alkaline pI values reduced the translation of rMefp1-encoding transcripts, thereby reducing the amount of soluble rMefp1 produced. After characterizing both feedback and non-feedback regulation in the acidic, alkaline, and neutral pI ranges, we suggest that three different pI range-specific soluble expression curves exist for the recombinant protein, each defined by specific ranges of the leader sequence pI values.

INTRODUCTION

Many researchers believe that the expression of recombinant proteins in soluble form in *E. coli* is inherently dependent upon the physical properties (e.g., isoelectric point (pI), hydrophobicity, stability, and molecular weight) of the entire amino acid sequence. Expression may also be affected by the presence of positive charges in the amino (N)-terminal region of the signal peptide (Inouye et al., 1982), by mRNA secondary structure (Mukund et al., 1999; Ramesh et al., 1994), and by codon usage (Ikemura, 1981). Although each of these factors has individually been found to affect the expression of specific proteins, they have not been applicable to the overall problem.

In attempting to develop a more general approach to solving the problem, we previously used hydropathy profile analysis to demonstrate that the presence of an internal, positively charged transmembrane (TM)-like domain in the olive flounder protein hepcidin I inhibits its soluble expression (Lee et al., 2008b). Using a novel secretion enhancer, we were able to overcome the obstacle posed by the internal TM-like domain and successfully expressed hepcidin I in soluble form. We also investigated Mefp1, an adhesive protein of the marine mussel *Mytilus edulis* (Waite, 1983). This large protein consists primarily of repeated decapeptide units lacking a TM-like domain (Lee et al., 2008b). We found that the addition of a Met residue to the N-terminus allowed low-level expression of recombinant 7xMefp1, a polypeptide comprising seven copies of the decapeptide repeat unit, in soluble form (Lee et al., 2008a) in *E. coli*. When we fused 7xMefp1 to a truncated OmpA signal peptide ($OmpASP_{tr}$) in an attempt to increase its expression in soluble form, we found that as the pI value of the N-terminus of the recombinant 7xMefp1 fusion protein (rMefp1) increased from 9.90 to 10.82, periplasmic expression of soluble rMefp1 also increased (Lee et al., 2008a).

However, the importance of the pI value of N-terminal regions in soluble expression of proteins lacking TM-like domains has not yet been properly investigated. Therefore, in the present study, we investigated the influence of the pI, hydrophobicity, charge, ΔG_{RNA} for mRNA unfolding (ΔG_{RNA}), codon usage, codon repetition, and molecular weight of the N-terminus on soluble rMefp1 expression in *E. coli*. We found that the N-terminal pI value can be used, over a wide range (2.73-13.35), as a comprehensive biological index to predict the level of soluble rMefp1 expression.

MATERIALS AND METHODS

Bacterial strains and plasmids

The *E. coli* strains XL-1 blue (Stratagene) and TOP10 (Invitrogen) were used for cloning, and BL21 (DE3) (Novagen) for direct expression of the fusion protein. The plasmid pBluescriptIISK(+) (Stratagene) and the TA cloning vector (Promega) were used for cloning, and pET-22b(+) (Novagen) for protein expression.

Reagents and molecular techniques

Restriction endonucleases (Roche) and a His-tag purification kit (Qiagen) were used. All other chemicals were of analytical grades. All molecular techniques were conducted as described

Biotechnology Research Division, National Fisheries Research and Development Institute, Pusan 619-902, Korea
*Correspondence: sangjl@nfrdi.go.kr

Received February 10, 2010; revised April 16, 2010; accepted April 26, 2010; published online July 23, 2010

Keywords: ΔG_{RNA} value, feedback regulatory mechanism, hydrophilicity, inner membrane channel, pI value trigger (facilitated diffusion)

by Sambrook et al. (1989). Nucleotide sequencing using the dideoxy chain-termination method (Sanger et al., 1977) was performed using the Sequenase 2.0 kit (United States Biochemical). The computer program DNASIS™ (Hitachi, Japan, 1997) was used to analyze the pI characteristics and hydrophobicity of polypeptide sequences.

Construction of expression vectors

To construct vectors for the expression of the 7xMefp1 target protein fused to N-terminal sequences with various pI values, we designed primer pairs specific for the individual N-termini (Supplementary Table 1). These primers were used to amplify DNA cassettes using the control N-terminal clone pET-22b(+)(*ompASP*$_1$(*Met*)-*7xmefp1**) as a template (Lee et al., 2008a). The amplified DNA was cloned into a TA cloning vector, the entire *Nde*I-*Xho*I fragment of which was then subcloned into pET-22b(+) by replacing the *pel* signal sequence and the polylinker as described previously (Lee et al., 2008a). The resulting constructs are listed in Table 1. To facilitate Western blot analysis, a C-terminal His-tag was added to each of the fusion proteins in the pET-22b(+) subcloning step.

Protein expression

E. coli BL21 (DE3) cells were transformed with the plasmid constructs listed in Table 1, and transformants were cultured in LB medium overnight at 30°C in the presence of 100 μg/ml ampicillin. The culture was then diluted 1:100 in LB medium and grown until it reached an optical density of 0.6 at 600 nm. Then, isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM, and the culture was grown for another 3 h to allow expression of the recombinant protein. A 1-ml aliquot was then removed from each culture and centrifuged. Cell pellets were resuspended in 100-200 μl of PBS. Cells were disrupted by sonication, in which 15 pulses at 30% power output were applied in 2-s cycles, and then centrifuged at 16,000 rpm for 30 min at 4°C. (The resulting supernatant fractions contained the soluble protein fraction.) The insoluble pellets were resuspended in an equal volume to that of the supernatants. To prepare periplasmic fractions, IPTG-induced cells were subjected to osmotic shock as described by Nossal and Heppel (1966). Protein fractions were quantified by a Bradford assay (Bradford, 1976), separated by SDS-PAGE on 15% acrylamide gels (Laemmli, 1970), and visualized using Coomassie brilliant blue stain. Band intensities were determined densitometrically using Quantity One 1-D image analysis software (Bio-Rad).

Western blot analysis

After electrophoresis, His-tagged rMefp1 proteins were transferred to a Hybond-P membrane (GE Healthcare) and detected using (in sequence) an anti-His tag (C-term) primary antibody, an alkaline phosphatase-conjugated anti-mouse secondary antibody, and a chromogenic Western blotting kit (Invitrogen), according to the manufacturers' protocols. Molecular weight markers (Benchmark, His-tagged) were used. Band intensities were quantified densitometrically using Quantity One 1-D image analysis software.

Determination of N-terminal ΔG_{RNA} values

The ΔG_{RNA} values for unfolding the RNA secondary structures of the N-termini were calculated using the program MFOLD 3 (www.bioinfo.rpi.edu/applications/mfold) (Zuker, 2003) (Table 1 and Supplementary Table 1). To investigate the effects of ΔG_{RNA} on soluble expression, we compared longer N-termini (which generally had smaller ΔG_{RNA} values) to shorter N-termini (which generally had larger ΔG_{RNA} values). To alter ΔG_{RNA}, we used synonymous codons. For example, in the $(Glu)_n$-containing N-termini ME_8-I and ME_6-I, the Glu^{GAA} and Glu^{GAG} codons are used with equal frequency, resulting in ΔG_{RNA} values of -8.30 and -8.50, respectively. In contrast, in ME_8-II and ME_6-II, the Glu^{GAA} codon is used for all but one Glu codon, reducing the ΔG_{RNA} values to -4.00 and -4.30, respectively. Similarly, in the $(Lys)_n$-containing N-termini, we used Lys^{AAA} codons to reduce ΔG_{RNA} values. In the $(Arg)_n$-containing N-termini, we used the Arg^{AGA} codon to reduce ΔG_{RNA} values, but it reduced translational efficiency when used as the sole Arg codon in MR_2AK (Supplementary Fig. S2) and caused translational arrest when used as the sole Arg codon in MR_4AK, MR_6AK, and MR_8AK (Supplementary Fig. S2). Therefore, we instead used alternating Arg^{CGT} and Arg^{CGC} codons, as detailed in Supplementary Table 1.

RESULTS AND DISCUSSION

Effects of N-termini with high pI values (9.90-13.35) on the expression of soluble rMefp1

We previously showed that changing the pI value of the N-terminus of the 7xMefp1 fusion protein rMefp1 (by altering the number of Lys residues) also modulated levels of periplasmic soluble expression (Lee et al., 2008a). Here, to assess the general relationship between N-terminal pI value and soluble rMefp1 expression, using the control clone pET-22b(+) (*ompASP*$_1$-*7xmefp1**) (Lee et al., 2008a), we constructed a series of rMefp1 fusion protein clones in which various numbers of Lys or Arg residues were inserted between the $OmpASP_1$(Met) truncated signal sequence and the 7xMefp1* sequence. The clones pET-22b(+)(*ompASP*$_1$-(*Lys*)$_n$-*7xmefp1**) (where n = 0 (control), 1, 2, 3, 4, 5, 6, or 8) and pET-22b(+)(*ompASP*$_1$-(*Arg*)$_n$-*7xmefp1**) (where n = 1, 2, 4, 6, or 8) yielded proteins with N-terminal pI values of 9.90, 10.55, 10.82, 10.99, 11.11, 11.21, 11.28, and 11.41 (for $(Lys)_0$ (control) to $(Lys)_8$, respectively) and of 11.52, 12.51, 12.98, 13.20, and 13.35 (for $(Arg)_1$-$(Arg)_8$, respectively). We analyzed the pI, hydrophobicity, charge, and ΔG_{RNA} of these N-termini in the $OmpASP_1$(Met)-$(Lys/Arg)_n$-Ala^1-Lys^2 clones (where Ala^1-Lys^2- represents the beginning of the native Mefp1 sequence) as previously described (Lee et al., 2008a) (Table 1 and Supplementary Table 1).

Soluble rMefp1 was expressed at higher levels with $(Lys)_n$-containing N-termini with pI values ranging from 10.55 to 11.28 (but not the one with a pI value 11.41) than with the control N-terminal sequence $OmpASP_1$(Met)-Ala^1-Lys^2- (where Ala is encoded by GCT; sequence name MAK-I; pI 9.90). The highest level of soluble rMefp1 expression was obtained with the N-termini with pI values of 10.82, 10.99, and 11.21 (expression decreased slightly at pI 11.28 and then substantially at pI 11.41) (Fig. 1A, Table 1, and Supplementary Fig. S1D). Therefore, decreased expression with longer N-termini cannot alone be explained by the high number of positive charges.

In the case of the $(Arg)_n$-containing N-termini, the highest expression of soluble rMefp1 was obtained using an N-terminus containing a single inserted Arg^{AGA} (Met-Arg^{AGA}-Ala^1-Lys^2; sequence name MRAK-I; pI 11.52). As the number of Arg residues continued to increase (up to an N-terminal pI of 13.25), expression gradually decreased (Fig. 1A, Table 1, and Supplementary Fig. S1E). Again, the decreased expression of soluble rMefp1 with longer N-termini could not be readily explained by the high positive charge.

At present, we do not know the basis for these observed changes in soluble rMefp1 expression. The $(Lys)_n$-containing N-termini cover a narrow pI range (9.90-11.41) that encompasses

Table 1. Characteristics of the various N-terminal sequences used to create recombinant 7xMefp1[a] fusion proteins (rMefp1) and relative levels of the soluble and insoluble forms of the corresponding fusion proteins synthesized from derivatives of the expression plasmid pET-22b(+)(*ompASP$_1$(Met)-7xmefp1**)[b].

Seq. No.	N-terminal sequence	pI[c]	Hydrophobicity index[d]	Charge	ΔG_{RNA}[e]	Soluble rMefp1 level[f]	Insoluble rMefp1 level[f]	Figure
1	MD_5AA	2.73	1.09	-5	≤ -9.90	0.50	0.83	1A, S1A
2	MD_3AA	2.87	0.56	-3	≤ -8.80	0.91	1.20	1A, S1A
3	MDA	3.00	N/A[h]	-1	≤ -6.70	1.40	1.59	1A, S1A
4	ME_8-I	2.75	2.08	-8	≤ -8.30	0.49	0.42	1A, 2C, S1A
5	ME_8-II	2.75	2.08	-8	≤ -4.00	N/D[i]	N/D[i]	2C
6	ME_6-I	2.82	1.82	-6	-8.50	0.65	0.62	1A, 2C, S1A
7	ME_6-II	2.82	1.82	-6	≤ -4.30	N/D[i]	N/D[i]	2C
8	ME_4	2.92	N/A[h]	-4	≤ -5.80	0.79	0.66	1A, S1A
9	MEE	3.09	N/A[h]	-2	-5.80	1.42	1.81	1A, S1A
10	MAE	3.25	N/A[h]	-1	≤ -5.60	1.72	1.92	1A, S1A
11	MC_6	4.61	-0.64	-	≤ -6.80	1.65	2.04	1A, S1B
12	MC_3	4.75	N/A[h]	-	≤ -7.10	1.93	2.95	1A, S1B
13	MAC	4.83	N/A[h]	-	≤ -5.20	1.96	1.84	1A, S1B
14	MAY	5.16	N/A[h]	-	-5.20	1.74	1.73	1A, S1B
15	MAA	5.60	N/A[h]	-	-5.60	2.25	2.23	1A, S1B
16	MNN	5.70	N/A[h]	-	≤ -2.50	N/D[i]	N/D[i]	2B
17	MTT	5.70	N/A[h]	-	≤ -3.10	N/D[i]	N/D[i]	2B
18	MWW	5.85	N/A[h]	-	≤ -7.50	N/D[i]	N/D[i]	2B
19	MGG	5.85	N/A[h]	-	≤ -7.80	1.93	2.10	1A, 2B, S1B
20	MAKD	6.59	N/A[h]	0	-7.00	2.30	2.63	1A, S1B
21	MAKE	6.79	N/A[h]	0	≤ -5.80	2.05	2.99	1A, S1B
22	MCH	7.13	N/A[h]	-	-3.60	1.83	3.11	1A, S1C
23	MAH	7.65	N/A[h]	-	-5.20	1.81	2.72	1A, S1C
24	MAH_3	7.89	N/A[h]	-	≤ -5.40	1.54	3.75	1A, S1C
25	MAH_5	8.01	-0.33	-	-6.70	1.37	4.36	1A, S1C
26	MAKC	8.78	N/A[h]	+1	-7.80	1.73	2.98	1A, S1C
27	MKY	9.58	N/A[h]	+1	≤ -2.50	1.51	4.04	1A, 2B, S1C
28	MAKY	9.58	N/A[h]	+1	-5.20	N/D[i]	N/D[i]	2B
29[j]	MAK-I (Ala: GCT) (control)[g]	9.90	N/A[h]	+1	-7.80	1.00	1.00	1A, 2A-C, S1A-E
30	MAK-II (Ala: GCA)	9.90	N/A[h]	+1	-7.80	N/D[i]	N/D[i]	2A
31	MAK-III (Ala: GCC)	9.90	N/A[h]	+1	-7.80	N/D[i]	N/D[i]	2A
32	MAK-IV (Ala: GCG)	9.90	N/A[h]	+1	-7.80	N/D[i]	N/D[i]	2A
33[j]	MKAK	10.55	N/A[h]	+2	≤ -2.40	1.57	3.14	1A, S1D
34	MK_2AK-I	10.82	N/A[h]	+3	≤ -2.10	1.69	3.43	1A, 2C, S1D
35[j]	MK_2AK-II	10.82	N/A[h]	+3	≤ -4.70	N/D[i]	N/D[i]	2C
36	MK_3AK	10.99	1.14	+4	≤ -2.00	1.80	2.96	1A, S1D
37	MK_4AK	11.11	1.32	+5	≤ -1.80	1.72	2.59	1A, S1D
38	MK_5AK	11.21	1.53	+6	≤ -1.70	1.93	2.34	1A, S1D
39	MK_6AK	11.28	1.69	+7	≤ -1.56	1.39	1.67	1A, S1D
40	MK_8AK	11.41	1.93	+9	≤ -1.39	0.44	0.39	1A, S1D
41	MRAK-I	11.52	N/A[h]	+2	≤ -2.90	1.69	2.92	1A, 2C, S1E
42	MRAK-II	11.52	N/A[h]	+2	≤ -5.00	N/D[i]	N/D[i]	2C
43	MR_2AK	12.51	N/A[h]	+3	≤ -5.10	1.26	1.27	1A, S1E
44	MR_4AK	12.98	1.32	+5	-7.70	1.07	1.14	1A, S1E
45	MR_6AK	13.20	1.69	+7	-10.20	0.93	0.88	1A, S1E
46	MR_8AK	13.35	1.93	+9	≤ -11.70	0.55	0.63	1A, S1E

[a]The Mefp1 decapeptide sequence is Ala-Lys-Pro-Ser-Tyr-Pro-Pro-Thr-Tyr-Lys (Waite, 1983).
[b]The plasmid pET-22b(+)(*ompASP$_1$(Met)-7xmefp1**) (Lee et al., 2008a) encodes the control N-terminal sequence (MAK-I) and was used as the basis for cloning all of the N-terminal variants.
[c]pI values were calculated using DNASIS software.
[d]Hydrophobicity index was calculated using DNASIS software by the Hopp-Woods method, with a window size of 6 and a threshold line of 0.00. On the Hopp-Woods scale, hydrophilic regions are given a positive value, and hydrophobic regions a negative one.
[e]The DNA sequences used to calculate the ΔG_{RNA} values for unfolding of the mRNA species encoding the fusion proteins begin near the translation initiation codon region of pET-22b(+) (5′-AAG AAG GAG ATA TA-3′) and include the forward primer sequences listed in Supplementary Table 1. MFOLD 3 software (Zuker, 2003) was used to calculate ΔG_{RNA} values. In cases in which the program yielded multiple ΔG_{RNA} values for a particular sequence (due to the existence of multiple possible folded structures), the higher/highest ΔG_{RNA} value is shown.
[f]Values indicate mean levels of soluble and insoluble fusion proteins, expressed relative to the levels of soluble and insoluble MAK-I control fusion protein, respectively, which were given values of 1.00 (Fig. 1).
[g]The control sequence MAK-I consists of the N-terminus of Met (OmpASP$_1$) + Ala-Lys (the first two amino acids of Mefp1) (Lee et al., 2008a) .
[h]Not applicable because the peptide is shorter than the minimum window size of 6 required for Hopp-Woods hydrophobicity calculations.
[i]Not detected.
[j]Primer constructed previously (Lee et al., 2008a).

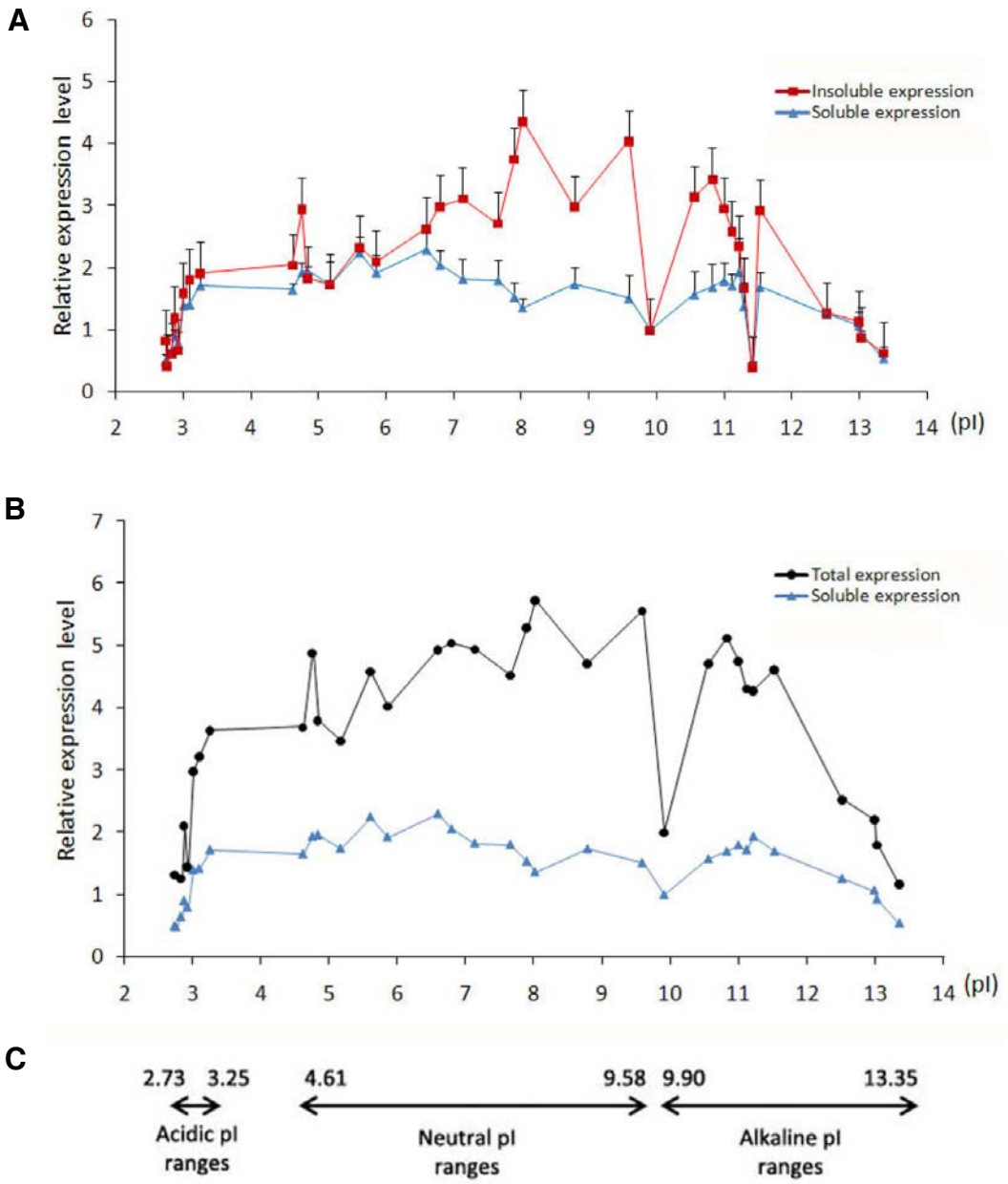

Fig. 1. Western blot analysis of insoluble (■), soluble (▲), and total (●) rMefp1 expression induced by N-terminal peptide sequences with various pI values. (Representative Western blot images showing insoluble and soluble rMefp1 expression induced by N-terminal peptide sequences with various pI values are presented in Supplementary Figs. S1A-S1E.) Gels were loaded with protein (approximately 20 μg per well) from the soluble fraction obtained from each clone and with an equal volume of the insoluble fraction. Anti-His-tag antiserum was used to detect the rMefp1 synthesized from a derivative of the *E. coli* expression vector pET-22b(+), which attaches a His-tag sequence to the C-terminal end of the fusion protein. (A) Expression levels were measured by densitometry relative to the levels of soluble and insoluble MAK-I (control protein; pI 9.90) (Lee et al., 2008a), which were given values of 1.00 (without correcting for the slightly higher amount of insoluble rMefp1). Data represent the mean relative levels of rMefp1 in three different samples analyzed by Western blotting (Supplementary Figs. S1A-S1E), each obtained from a different colony (as shown in Table 1). Satellite bands were excluded from the densitometric analysis. (B) The data in (A) were re-plotted as a Hubbert curve (Hubbert, 1956) to show mean total and soluble rMefp1 expression in presenting pI range-specific total and soluble expression for N-termini with pI values ranging from 9.90 to 13.35. (C) The boundaries of the acidic, neutral, and alkaline pI ranges were deduced from the soluble rMefp1 expression curve (B).

the optimal pI—the expression curve peaks within this range. This hyperbolic change in expression may be related to a pI-dependent inner membrane translocation mechanism that secretes the target precursor protein from the cytoplasm to the periplasm. Soluble expression of rMefp1 fusion proteins with $(Arg)_n$-containing N-termini (pI 11.52-13.35) gradually decreased as the value of *n* increased, suggesting the involvement of a similar inner membrane translocation mechanism.

We next examined hydrophilicity as an alternative index for the charge of the N-termini described above. Those N-termini

with pI values of 11.41 (Met-$(Lys)_8$-AK; charge + 9), 12.98 (Met-$(Arg)_4$-AK; charge + 5), 13.20 (Met-$(Arg)_6$-AK; charge + 7), and 13.35 (Met-$(Arg)_8$-AK; charge + 9), which produced only very weak expression of soluble rMefp1, are highly hydrophilic (Hopp-Woods hydrophobicity indices (*h*) (see "Materials and Methods") 1.93, 1.32, 1.69, and 1.93, respectively). Thus, highly hydrophilic N-termini appeared to reduce soluble rMefp1 expression.

However, the $(Lys)_n$-containing N-termini for which n = 4, 5, or 6 (Met-$(Lys)_4$-AK (charge +5, pI 11.11), Met-$(Lys)_5$-AK (charge + 6, pI 11.21), and Met-$(Arg)_6$-AK (charge + 7, pI 11.28)), which are also highly hydrophilic (h = 1.32, 1.53, and 1.69, respectively), yielded higher levels of soluble rMefp1 expression than Met-$(Arg)_4$-AK (charge +5, pI 12.98), for which n = 4 and h = 1.32 (Fig. 1A and Table 1). The hydrophilicity of each $(Lys)_n$-containing N-terminus is equal to or greater than that of the corresponding $(Arg)_n$-containing N-terminus. Therefore, the $(Lys)_n$-containing N-termini appear to be superior to the $(Arg)_n$-containing N-termini in terms of their ability to promote soluble rMefp1 expression.

The highly hydrophilic $(Lys)_8$-, $(Arg)_4$-, $(Arg)_6$-, and $(Arg)_8$-containing N-termini failed to increase the expression of soluble rMefp1 (Fig. 1A), which lacks a TM-like domain, whereas we previously found that the insertion of $(Lys)_6$, $(Arg)_6$, $(Arg)_8$, or $(Arg)_{10}$ sequences enhanced the secretion of a target protein containing a TM-like domain (Lee et al., 2008b). Therefore, the potential of $(Lys)_n$- and $(Arg)_n$-containing N-termini to enhance secretion (Lee et al., 2008b) or reduce soluble expression, by increasing pI and *h*, appears to depend on whether the target protein has a TM-like domain.

Effects of N-termini with lower pI values (2.73-9.58) on the expression of soluble rMefp1

To assess the influence of N-termini with lower pI values on soluble rMefp1 expression, we created additional N-terminal variants of the $OmpASP_1$(Met)-7xMefp1* clones with pI values ranging from 2.73 to 9.58. In these clones, 2-8 amino acid residues were inserted after the $OmpASP_1$ sequence, and the first two residues of the 7xMefp1 polypeptide (Ala^1-Lys^2-) were replaced. (These clones were of the form pET-22b(+)(*ompASP*$_1$-(*X*)-*mefp1*$^{3\text{-}10}$-*6xmefp1**), where X is the insert.) These N-termini were compared with the control MAK-I sequence (Met-Ala^1-Lys^2; pI 9.90) in terms of their ability to induce soluble expression of rMefp1 fusion proteins. The sequences of the N-termini were Met-$(Asp)_5$-$(Ala)_2$ (pI 2.73); Met-$(Asp)_3$-$(Ala)_2$ (pI 2.87); Met-Asp-Ala (pI 3.00); Met-$(Glu^{GAA}Glu^{GAG})_4$ (sequence name ME_8-I; pI 2.75); Met-$(Glu^{GAA}Glu^{GAG})_3$ (sequence name ME_6-I; pI 2.82); Met-$(Glu)_2$ (pI 3.09); Met-Ala-Glu (pI 3.25); Met-Cys_6 (pI 4.61); Met-$(Cys)_3$ (pI 4.75); Met-Ala-Cys (pI 4.83); Met-$(Ala)_2$ (pI 5.60); Met-$(Gly)_2$ (pI 5.85); Met-Ala-Lys-Asp (pI 6.59); Met-Ala-Lys-Glu (pI 6.79); Met-Cys-His (pI 7.13); Met-Ala-His (pI 7.65); Met-Ala-$(His)_3$ (pI 7.89); Met-$(His)_5$ (pI 8.01); Met-Ala-Lys-Cys (pI 8.78); and Met-Lys-Tyr (pI 9.58) (Table 1 and Supplementary Table 1). Soluble expression generally increased as the N-terminal pI increased from 2.73 to 3.25 and was higher with all N-termini with pI values ranging between 4.61 and 9.58 than with the control N-terminus (Fig. 1A, Table 1, and Supplementary Figs. S1A-S1C).

Analysis of the curves showing soluble expression of rMefp1 fusion proteins with $(Asp)_n$- and $(Glu)_n$-containing N-termini (pI 2.73-3.25) revealed that their slopes decreased as the value of n increased (Fig. 1A and Table 1). Thus, as hydrophilicity increased, the expression of soluble rMefp1 decreased (as was the case with the $(Lys)_n$- and $(Arg)_n$-containing N-termini).

The above N-terminal sequences with pI values in the range 4.75 to 9.58 are shorter than the minimum six residues required for calculation of hydrophobicity indices on the Hopp-Woods scale (the hydrophobicity of the Cys-containing N-terminus Met-$(Cys)_6$ (pI 4.61) is -0.64). They all contain Cys, Ala, Gly, Lys, Asp, Glu, and/or His residues. Only Met-Ala-Lys-Cys and Met-Lys-Tyr are charged (Table 1), so the size of the positive charge cannot be used as a useful index for soluble rMefp1 expression. Therefore, we examined ΔG_{RNA} as a potential determinant of the expression results; mRNA transcripts with strong secondary structures in their 5′ regions are translated less efficiently (Mukund et al., 1999; Ramesh et al., 1994). The calculated ΔG_{RNA} values of the N-termini with pI values in the range 4.61-9.58 varied widely, from -2.50 to -7.80 (Table 1 and Supplementary Table 1), but did not seem to correlate with the level of expression of soluble rMefp1 (Fig. 1A and Table 1).

Overall, analysis of N-termini with pI values ranging from 2.73 to 13.35 yielded at least four different curves describing the relationship between pI and expression: expression increased from pI 2.73 to 3.25, remained relatively high from pI 4.61 to 9.58, increased hyperbolically from pI 9.90 to 11.41, and decreased from pI 11.52 to 13.35 (Fig. 1A). These curves clearly show that the pI value of the N-terminus can be used, over a wide range (2.73-13.35), to predict the level of expression of soluble rMefp1. Both a $(Lys)_n$-specific curve (pI 9.90-11.41) and an $(Arg)_n$-specific curve (pI 11.52-13.35) fell within the alkaline range (pI 9.90-13.35) (Fig. 1A). Because soluble expression depends on membrane transport mechanisms (as discussed above), we surmise that the number of specific soluble expression curves at alkaline pI values directly relates to the number of membrane channels involved in the transport of rMefp1.

To test the presumptive number of curves in the alkaline pI range (pI 9.90-13.35), we next re-plotted the data as a Hubbert curve (Hubbert, 1956), which encompassed the $(Lys)_n$-specific (pI 9.90-11.41) and $(Arg)_n$-specific (pI 11.52-13.35) curves. Removal of the data points relating to Lys-containing N-termini with pI values of 11.28 and 11.41, which yielded especially low expression of soluble rMefp1, produced a smooth curve in the pI range 9.90-13.35 (Fig. 1B). Three soluble rMefp1 expression curves were apparent: an acidic curve from pI 2.73 to 3.25, a neutral curve from pI 4.61 to 9.58, and an alkaline curve from pI 9.90 to 13.35. These observations again suggest the presence of distinct inner membrane channels that transport rMefp1 fusion proteins with acidic, neutral, and alkaline N-termini.

Effect of the N-terminal pI value on the expression of insoluble rMefp1

We next assessed the influence of the N-terminal pI on the expression of insoluble rMefp1 fusion protein using the same N-terminal sequences used to assess soluble expression (Fig. 1A, Table 1, and Supplementary Figs. S1A-S1E). The combined amount of soluble and insoluble rMefp1 was considered to be an indirect index of translation (Fig. 1B). With $(Asp)_n$- and $(Glu)_n$-containing N-termini (pI 2.73-3.25, ΔG_{RNA} -5.60 to -9.90), the amount of insoluble rMefp1 decreased gradually as n increased and pI decreased. With N-termini of pI 4.61-9.58 (ΔG_{RNA} -2.50 to -7.80), more insoluble rMefp1 was produced than MAK-I control (pI 9.90, ΔG_{RNA} -7.80). Insoluble rMefp1 expression induced by the $(Lys)_n$-containing N-termini (ΔG_{RNA} -7.80 to -1.39) increased between pI 9.90 and 10.82 and then decreased markedly at pI 11.41, while that induced by the $(Arg)_n$-containing N-termini (ΔG_{RNA} -2.90 to -11.70) decreased gradually, from a peak at pI 11.52 to a much lower level than that of the control at pI 13.35 (Fig. 1A and Table 1). As with the soluble expression curves (described above), the separately generated insoluble expression curves for $(Lys)_n$- and $(Arg)_n$-

containing N-termini (pI ranges 9.90-11.41 and 11.52-13.35, respectively) were combined to yield a single total expression curve encompassing the pI range 9.90-13.35 (Fig. 1B).

Although ΔG_{RNA} influences translational efficiency, we believe that it is more accurately reflected by the total amount of target protein expressed than by the amount of target protein expressed in soluble form, which is also determined by the efficiency of its secretion into the periplasm through an N-terminal pI range-specific inner membrane channel. Therefore, we also investigated the influence of N-terminal pI, ΔG_{RNA}, and hydrophilicity on total and soluble expression of rMefp1 fusion proteins with N-termini with pI 2.73-3.25 (Fig. 1B and Table 1). With the N-termini Met-$(Asp)_5$-$(Ala)_2$ (sequence name MD_5AA; pI 2.73, ΔG_{RNA} -9.90, *h* 1.09), Met-$(Asp)_3$-$(Ala)_2$ (MD_3AA; pI 2.87, ΔG_{RNA} -8.80, *h* 0.56), and Met-Asp-Ala (MDA; pI 3.00, ΔG_{RNA} -6.70), total rMefp1 expression decreased as ΔG_{RNA} and hydrophilicity increased. Moreover, decreased total rMefp1 expression correlated with decreased soluble rMefp1 expression. Soluble rMefp1 expression changed markedly as a factor of both the N-terminal pI value and overall level of protein expression.

With the $(Glu)_n$-containing N-terminal sequences ME_8-I (pI 2.75, ΔG_{RNA} -8.30, *h* 2.08), ME_6-I (pI 2.82, ΔG_{RNA} -8.50; *h* 1.82), ME_4 (pI 2.92; ΔG_{RNA} -5.80), ME_2 (pI 3.09, ΔG_{RNA} -5.80), and MAE (pI 3.25, ΔG_{RNA} -5.60), total rMefp1 expression decreased as *n* and *h* increased (Fig. 1B and Table 1), consistent with the observed decrease in soluble rMefp1 expression. However, the small difference in the ΔG_{RNA} values of ME_8-I (ΔG_{RNA} -8.30) and ME_6-I (ΔG_{RNA} -8.50) would not be expected to greatly influence overall rMefp1 expression, which was seemingly determined primarily by the hydrophilicity of the N-terminus. With $(Glu)_n$-containing N-termini, pI also greatly influences soluble rMefp1 expression, which (with such N-termini) is thus apparently also determined by the overall level of rMefp1 expression and the activity of a pI range-specific inner membrane channel.

Of the N-termini with pI 2.73-3.25, the majority of rMefp1 was expressed in insoluble form with five, but not with the remaining three, which yielded slightly less insoluble protein than soluble protein (Table 1, seq. nos. 1-4, 6, 8-10), suggesting that no active transport (Overton, 1895) of the soluble form into the periplasm occurred. As all rMefp1 fusion proteins larger than the control Met-7xMefp1* (9.8 kDa) were secreted into the periplasm in soluble form, we suggest that rMefp1 fusion proteins with acidic N-termini are transported via the acid-specific inner membrane channels of *E. coli* by facilitated diffusion.

Based on these results, we postulate that soluble expression of with these N-termini is regulated by the relationship between the N-terminal pI values of all the cytoplasmic rMefp1 fusion proteins and the pI range-specific inner membrane channel. Notably, the pI range-specific inner membrane channel is irreversibly attached to the inner membrane, while the pI values of the N-termini of all of the target proteins are variable, so that the level of soluble expression would be determined by the various pI values of the N-termini of all of the target proteins. Therefore, we introduced the term "pI value trigger" to describe the proportion of total protein expressed in soluble form in the periplasm, as determined by translocational efficiency or the diffusion coefficient for transit through the pI range-specific inner membrane channel, which is in turn controlled by the various pI values of the N-termini of all of the target proteins.

In general, greater total rMefp1 expression was induced by N-termini with pI values in the range 4.61-9.58 (ΔG_{RNA} -2.50 to -7.80) than by the control MAK-I N-terminus. This expression did not correlate closely with ΔG_{RNA} but did, for MAH_5 (pI 8.01, ΔG_{RNA} -6.70; *h* -0.33), appear to be determined by the N-terminal hydrophobicity (MAH_3 (pI 7.89, ΔG_{RNA} -5.40) and MAH (pI 7.65, ΔG_{RNA} -5.20), do not have calculable hydrophobicities) (Table 1). However, it is possible to represent the pI value as an index for the soluble/total rMefp1 ratio in terms of a pI value trigger or diffusion coefficient. This ratio varies substantially: at high pI value triggers (pI 4.83, 5.16, 5.60), it was almost 1:2, while at low pI value triggers (pI 8.01), it was less than 1:4 (Table 1, seq. nos. 11-15, 19-27). However, soluble expression of rMefp1 was higher with all N-termini with pI values in the range 4.61-9.58 than with the control N-terminus (Fig. 1B). The neutral curve to which it relates shows a non-gradual change in slope, in contrast to the other two curves (pI 2.73-3.25 and 9.90-13.35), which have large positive and/or negative slopes, respectively. We surmise that soluble rMefp1 fusion proteins with N-termini pI values in the range 4.61-9.58 pass from the total rMefp1 pool through the neutral inner membrane channel to the periplasm by facilitated diffusion (as suggested above for fusion proteins with N-termini of pI 2.73-3.25).

In examining total rMefp1 expression induced by the $(Lys)_n$-containing N-termini (pI 9.90-11.41, ΔG_{RNA} -7.80 to -1.39), we found that total expression was lower with MAK-I (pI 9.90, ΔG_{RNA} -7.80) than with MK_2AK-I (pI 10.82, ΔG_{RNA} -2.10), consistent with the favorable change in ΔG_{RNA}. However, total expression induced by MK_3AK (pI 10.99, ΔG_{RNA} -2.00, *h* 1.14) was much higher than that induced by MK_8AK (pI 11.41, ΔG_{RNA} -1.39, *h* 1.93), in spite of the slightly more favorable ΔG_{RNA} value of the latter (Fig. 1A and Table 1). With longer $(Lys)_n$-containing N-termini, decreases in total rMefp1 expression were not consistent with the reductions in ΔG_{RNA}, but were instead related to increases in hydrophilicity. Here, to decrease ΔG_{RNA}, we used the Lys^{AAA} codon (Supplementary Table 1). As the number of these Lys^{AAA} codons was increased, total rMefp1 expression rose and then gradually fell (without translational arrest) (Fig. 1A and Table 1), indicating that the use of the Lys^{AAA} codon prevented translational inhibition. The soluble and insoluble rMefp1 expression curves for these $(Lys)_n$-containing N-termini show the typical hyperbolic form (Fig. 1A), leading us to speculate that a feedback mechanism regulates total and soluble rMefp1 protein levels, and that the trigger for this feedback is the increased hydrophilicity of the N-terminus.

Among the $(Lys)_n$-containing N-termini, total rMefp1 expression was highest with MK_2AK-I (pI 10.82, ΔG_{RNA} -2.10), but soluble rMefp1 expression was higher with MK_3AK (pI 10.99, ΔG_{RNA} -2.00), MK_4AK (pI 11.11, ΔG_{RNA} -1.80), and MK_5AK (pI 11.21, ΔG_{RNA} -1.70) than with MK_2AK-I, despite their slightly reduced production of total rMefp1 (Fig. 1B and Table 1). This observation suggests that total protein synthesis and soluble expression are controlled independently. The regulation of total protein synthesis appears to be complicated, and the inner membrane protein very sensitive to the N-terminal pI value.

With $(Lys)_n$-containing N-termini of pI 9.90-11.41, the most rMefp1 was expressed in insoluble form (with the exception of one N-terminus, which yielded slightly less insoluble protein than soluble protein) (Table 1, seq. nos. 29, 33, 34, 36-40), suggesting that fusion proteins with these termini were translocated through the alkaline pI-specific inner membrane channel by facilitated diffusion (similar to the $(Asp)_n$- and $(Glu)_n$-containing N-termini).

Among the $(Arg)_n$-containing N-termini with pI values in the range 11.52-13.35, highest total expression of rMefp1 was obtained with MRAK-I (pI 11.52, ΔG_{RNA} -2.90), which uses the Arg^{AGA} codon. Total expression gradually decreased as the number of Arg-encoding codons was increased: MR_2AK (R_2 encoded by $Arg^{CGT}Arg^{CGC}$, pI 12.51, ΔG_{RNA} -5.10) < MR_4AK (R_4 = $(Arg^{CGT}Arg^{CGC})_2$, pI 12.98, ΔG_{RNA} -7.70, *h* 1.32) < MR_6AK (R_6

= ($Arg^{CGT}Arg^{CGC}$)$_3$, pI 13.20, ΔG_{RNA} -10.20, *h* 1.69) < MR_8AK (R_8 = ($Arg^{CGT}Arg^{CGC}$)$_4$, pI 13.35, ΔG_{RNA} -11.70, *h* 1.93) (Fig. 1B, Table 1, and Supplementary Table 1). With these N-termini, rMefp1 expression decreased as ΔG_{RNA} and *h* increased. While effective in MRAK-I, the Arg^{AGA} codon could not be used in MR_2AK due to a reduction in translational efficiency (Supplementary Fig. S2), or in MR_4AK, MR_6AK, or MR_8AK, due to translational arrest (Supplementary Fig. S2). These results show that repetitive use of a particular codon can slow or arrest translation, regardless of the value of ΔG_{RNA}.

As with the $(Lys)_n$-containing N-termini, the $(Arg)_n$-containing N-termini yielded higher insoluble expression of rMefp1 fusion proteins than soluble expression (with the exception of one N-terminus, which yielded slightly less insoluble protein than soluble protein) (Table 1, seq. nos. 41, 43-46). These differences in the expression of soluble rMefp1 appeared to be caused by different pI triggers for inner membrane secretion of the total rMefp1 (the soluble expression of these fusion rMefp1 occurring as a result of facilitated diffusion through the alkaline pI-specific inner membrane channel). The curves displaying soluble and insoluble expression of these rMefp1 fusion proteins showed similar trends (Fig. 1A).

As discussed above, greater amounts of soluble rMefp1 were expressed with the $(Lys)_n$-containing N-termini with hydrophilicities of 1.32 (MK_4AK; pI 11.11, ΔG_{RNA} -1.80, pI value trigger 0.40), 1.53 (MK_5AK; pI 11.21, ΔG_{RNA} -1.70, pI value trigger 0.45), and 1.69 (MK_6AK; pI 11.28, ΔG_{RNA} -1.56, pI value trigger 0.45) than with the $(Arg)_n$-containing N-terminus with a hydrophilicity of 1.32 (MR_4AK; pI 12.98, ΔG_{RNA} -7.70, pI value trigger 0.48) (Table 1). This phenomenon results from the complicated nature of the influence of ΔG_{RNA} and *h* on total protein expression and of the pI value trigger on soluble expression. The higher soluble expression with $(Lys)_n$-containing sequences than with $(Arg)_n$-containing sequences may be better explained by ΔG_{RNA} than by hydrophilicity or pI trigger values.

The least effective inducers of soluble rMefp1 expression were the N-terminal sequences MD_5AA (pI 2.73, ΔG_{RNA} -9.90, *h* 1.09), MK_8AK (pI 11.41, ΔG_{RNA} -1.39, *h* 1.93), and MR_8AK (pI 13.35, ΔG_{RNA} -11.70, *h* 1.93). However, the fusion proteins containing these N-termini were all expressed in the periplasm, as described in "Materials and Methods" (data not shown). Both soluble and insoluble expression levels were observed (Fig. 1A, Table 1, and Supplementary Figs. S1A, S1D, and S1E), indicating that a proportion of the rMefp1 produced was secreted into the periplasm in soluble form through the acidic or alkaline pI-specific inner membrane channels. The remaining rMefp1 protein in the cytoplasm was collected in insoluble form, as described in "Materials and Methods". Therefore, it seems that a mechanism exists that converts even the smallest amounts of unsecreted protein into an insoluble form in the cytoplasm.

We consider combining the insoluble expression curves for the $(Lys)_n$- and $(Arg)_n$-containing N-termini (pI 9.90-13.35) into a single curve to be logical. Removing the data points relating to N-termini with pI values of 11.28 and 11.41, responsible for a sharp but transitory decrease in insoluble expression, yielded a relatively smooth total expression hyperbolic curve (Fig. 1B). The alkaline curves for total and soluble rMefp1 expression (Fig. 1B) showed similar trends.

We have suggested the existence of a hydrophilicity-based feedback regulatory mechanism controlling the expression of soluble and insoluble rMefp1 fused to $(Lys)_n$-containing N-termini (pI 9.90-11.41). We suggest also that the similar patterns of soluble and insoluble expression obtained with the $(Asp)_n$- and $(Glu)_n$-containing N-termini (pI 2.73-3.25) and from the $(Arg)_n$-containing N-termini (pI 11.52-13.35) (Fig. 1A) are the result of hydrophilicity-based feedback regulation. However, the same is not true of the rMefp1 fusion proteins with $(Asp)_n$- or $(Arg)_n$-containing N-termini because their expression is also influenced by ΔG_{RNA}. Designing longer $(Asp)_n$- or $(Arg)_n$-containing N-termini with lower ΔG_{RNA} values than those of shorter N-termini (to provide a broader range of ΔG_{RNA} values for closer examination of this issue) was problematic. Instead, we decided to examine the relationship between ΔG_{RNA} and hydrophilicity in feedback regulation of soluble and insoluble rMefp1 expression using two pairs of $(Glu)_n$-containing N-termini with differing codon usages. These were ME_8-I (pI 2.75, ΔG_{RNA} -8.30) and ME_8-II (pI 2.75, ΔG_{RNA} -4.00), which both have a hydrophilicity of 2.08; and ME_6-I (pI 2.82, ΔG_{RNA} -8.50) and ME_6-II (pI 2.82, ΔG_{RNA} -4.30), which both have a hydrophilicity of 1.82 (Table 1 and Supplementary Table 1). Analysis of these N-termini showed a modest dependence on ΔG_{RNA}, but a much greater dependence on hydrophilicity (Fig. 2C). Overall, the results from the longer N-termini with pI values in the ranges 2.73-3.25, 9.90-11.41, and 11.52-13.35 show that hydrophilicity is much more important than ΔG_{RNA} in inhibitory feedback regulation of total rMefp1 synthesis and soluble rMefp1 expression. Therefore, this feedback regulation seems to depend primarily on the N-terminal hydrophilicity.

The hydrophilicity of the acid (pI 2.73-3.25) and alkaline (9.901-13.35) N-termini appears to have a dual function: it influences 1) protein synthesis, thus controlling overall rMefp1 expression, and 2) the pI value trigger, thus controlling soluble rMefp1 expression. Therefore, we believe that the reduction in total rMefp1 expression linked to increased N-terminal hydrophilicity results from decreased translational efficiency. This presumptive regulation of the translational efficiency and total and soluble protein expression by N-terminal hydrophilicity suggests that it is important for feedback regulation in *E. coli*. Differences between the hydrophilic amino acids present in different N-termini (Asp, Glu, Lys, and Arg) suggest that hydrophilicity is a more important factor for feedback regulation than is the presence of a specific recognition site.

We also investigated the influence of charge on total rMefp1 expression. Analysis of the $(Asp)_n$-, $(Glu)_n$-, $(Arg)_n$-, and $(Lys)_n$-containing N-termini showed that, as the positive or negative charge increased, total rMefp1 expression level decreased (Fig. 1A and Table 1). This effect correlated more closely with hydrophilicity than with the value of *n*. However, pI values in the ranges 2.73-3.25 and 9.90-13.35 can be used as an index to predict levels of total and soluble rMefp1 expression, regardless of hydrophilicity.

Our analysis of the relationship between N-terminal hydrophilicity and overall rMefp1 expression revealed both feedback and non-feedback regulatory control of soluble rMefp1 expression. In non-feedback regulation, which we observed with N-termini of pI 4.61-9.58 and no calculated hydrophilicity, ΔG_{RNA} had little effect on total rMefp1 expression, while the rate at which the cytoplasmic rMefp1 was secreted into the periplasm in soluble form (by facilitated diffusion through inner membrane channels specific for N-termini with neutral pI values) was controlled by the pI value of the N-terminus.

In feedback regulation, which we observed with the N-termini with values in the ranges 2.73-3.25 and 9.90-13.35, total rMefp1 expression was controlled by ΔG_{RNA} when the N-termini were short and had no calculated hydrophilicity and by hydrophilicity when the N-termini contained multiple hydrophilic amino acids. For both short and long N-termini, the rate at which cytoplasmic rMefp1 was secreted into the periplasm in soluble form was controlled by the specific pI value trigger for facilitated diffusion through inner membrane channels specific

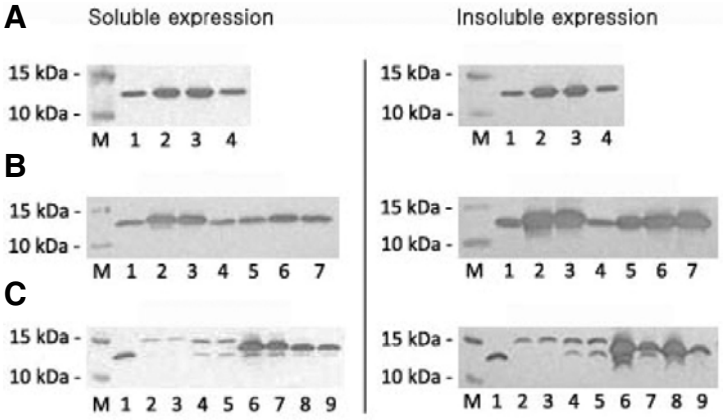

Fig. 2. Comparative Western blot analysis of soluble and insoluble rMefp1 expression directed by N-termini with identical pI values. Western blots analysis was performed as described in Fig. 1. (A) Comparison of four Met-Ala-Lys N-termini with different Ala codons. All have a pI value of 9.90 and a ΔG_{RNA} value of -7.80. MAK-I uses Ala^{GCT}, MAK-II uses Ala^{GCA}, MAK-III uses Ala^{GCC}, and MAK-IV uses Ala^{GCG}. Lanes: M, molecular markers; 1, MAK-I; 2, MAK-II; 3, MAK-III; 4, MAK-IV. (B) N-termini with different amino acid compositions. Lanes: M, molecular markers; 1, MAK-I (pI 9.90); 2, MNN (pI 5.70); 3, MTT (pI 5.70); 4, MWW (pI 5.85); 5, MGG (pI 5.85); 6, MAKY (pI 9.58); 7, MKY (pI 9.58). (C) N-termini with different ΔG_{RNA} values with or without calculated hydrophobicity indices. Lanes: M, molecular marker; 1, MAK-I (pI 9.90; Ala^{GCT}; ΔG_{RNA} -7.80); 2, ME_8-I (pI 2.75, ΔG_{RNA} -8.30); 3, ME_8-II (pI 2.75, ΔG_{RNA} -4.00); 4, ME_6-I (pI 2.82, ΔG_{RNA} -8.50); 5, ME_6-II (pI 2.82, ΔG_{RNA} -4.30); 6, MK_2AK-I (pI 10.82, ΔG_{RNA} -2.10); 7, MK_2AK-II (pI 10.82, ΔG_{RNA} -4.70); 8, MRAK-I (pI 11.52, ΔG_{RNA} -2.90); 9, MRAK-II (pI 11.52, ΔG_{RNA} -7.80).

for N-termini with acidic or alkaline pI values.

N-terminal pI values affect soluble rMefp1 expression levels, even when increased hydrophilicity reduces overall protein synthesis through feedback regulation. Therefore, as an index for soluble rMefp1 expression over a wide pI range, N-terminal pI is a more important than hydrophilicity, charge, or ΔG_{RNA}, irrespective of whether soluble rMefp1 expression is controlled by pI-based feedback, as shown in Figs. 1A, 1B, and Table 1.

Effects of codon selection and duplication, and amino acid sequence, on expression of soluble and insoluble rMefp1 fusion proteins with N-termini with identical pI values

We showed that with N-termini of pI 2.73-3.25 and 9.90-13.35, the N-terminal ΔG_{RNA} and hydrophilicity influence overall rMefp1 expression, and that for all N-termini of pI 2.73-13.35, the pI value controls soluble rMefp1 expression via the pI value trigger. Therefore, we hypothesized that in N-termini with identical pI values, differences in codon usage or amino acid sequence might influence expression of soluble and insoluble rMefp1. We assessed the expression of soluble and insoluble rMefp1 induced by three sets of N-termini with identical pI values. These sets possessed (i) identical amino acid sequences and ΔG_{RNA} values but different codon selections (Met-Ala^{GCT}-Lys (MAK-I), Met-Ala^{GCA}-Lys (MAK-II), Met-Ala^{GCC}-Lys (MAK-III), and Met-Ala^{GCG}-Lys (MAK-IV) (all pI 9.90, ΔG_{RNA} -7.80)); (ii) identical ΔG_{RNA} values but different amino acid sequences (MNN and MTT (both pI 5.70); MWW and MGG (both pI 5.85); and MKY and MAKY (both pI 9.58)); or (iii) identical amino acid sequences but different ΔG_{RNA} values (Met-$(Glu^{GAA}Glu^{GAG})_4$ (ME_8-I; ΔG_{RNA} -8.30) and Met-$(Glu^{GAA})_7Glu^{GAG}$ (ME_8-II; ΔG_{RNA} -4.00); Met-$(Glu^{GAA}Glu^{GAG})_3$ (ME_6-I; ΔG_{RNA} -8.50) and Met-$(Glu^{GAA})_5Glu^{GAG}$ (ME_6-II; ΔG_{RNA} -4.30); Met-$(Lys^{AAA})_2$-Ala-Lys (MK_2AK-I; ΔG_{RNA} -2.10) and Met-Lys^{AAA}-Lys^{AAG}-Ala-Lys (MK_2AK-II; ΔG_{RNA} -4.70); and Met-Arg^{AGA}-Ala-Lys (MRAK-I; ΔG_{RNA} -2.90) and Met-Arg^{CGT}-Ala-Lys (MRAK-II; ΔG_{RNA} -5.00)) (Table 1 and Supplementary Table 1).

Levels of soluble and insoluble rMefp1 obtained using MAK-I (using Ala^{GCT}) were similar to those for MAK-IV (using Ala^{GCG}) but very different from those for MAK-II (Ala^{GCA}) and MAK-III (Ala^{GCC}), in spite of their identical ΔG_{RNA} values (Fig. 2A). MNN and MTT, which have identical pI values, similar molecular weights (377.41 and 351.41, respectively) and slightly different ΔG_{RNA} values (-2.50 and -3.10, respectively), were quite similar in terms of soluble and insoluble rMefp1 expression (Fig. 2B). MWW and MGG, which have identical pI values, very different molecular weights (521.61 and 263.31, respectively) and similar ΔG_{RNA} values (-7.50 and -7.80), differed in terms of insoluble rMefp1 expression (the smaller MGG yielded higher expression) but not soluble rMefp1 expression (Fig. 2B). MKY and MAKY, which have identical pI values and similar amino acid sequences and molecular weights (440.55 and 511.62, respectively), but different ΔG_{RNA} values (-2.50 and -5.20, respectively), yielded similar amounts of soluble and insoluble rMefp1 (Fig. 2B).

The longer ME_8-I/II and ME_6-I/II pairs, which have identical pI values and hydrophilicity, but different codon selections and ΔG_{RNA} values (ME_8-I and -II: ΔG_{RNA} -8.30 and -4.00, respectively; ME_6-I and -II: ΔG_{RNA} -8.50 and -4.30, respectively), yielded similar levels of soluble and insoluble rMefp1 (Fig. 2C). The shorter MK_2AK-I/II and MRAK-I/II pairs, which also have identical pI values, but different codon selections and ΔG_{RNA} values and no calculated hydrophobicity indices (MK_2AK-I and -II: ΔG_{RNA} -2.10 and -4.70, respectively; MRAK-I and -II: ΔG_{RNA} -2.90 and -5.00, respectively), yielded very different levels of insoluble rMefp1 according to their ΔG_{RNA} values, but similar amounts of soluble rMefp1 (Fig. 2C). Of course, higher overall rMefp1 expression might be expected to translate into increased levels of soluble rMefp1.

In the case of the MAK N-termini, all of which have the same ΔG_{RNA}, single-base differences at the third "wobble" position of the Ala-encoding codons resulted in large changes in the expression of soluble and insoluble rMefp1. The Ala^{GCC} codon yielded higher soluble and insoluble rMefp1 expression than Ala^{GCT}, even though both codons use the same anticodon (CGG). Similarly, Ala^{GCA} yielded more soluble and insoluble rMefp1 than Ala^{GCG}, even though both codons use a CGU anticodon. Thus, MAK-I and MAK-IV (which contain Ala^{GCT} and Ala^{GCG}, respectively), which use different anticodons (CGG and CGU, respectively), yielded similar amounts of soluble and insoluble rMefp1. This observation suggests that a single base change can alter total rMefp1 expression by affecting the efficiency of translation.

Small differences in N-terminal molecular weight between MNN and MTT, and MKY and MAKY, had little effect on soluble and insoluble rMefp1 expression. The large difference in the molecular weights of the MWW and MGG N-termini had a significant effect on the amount of insoluble protein produced, but not on soluble rMefp1 expression.

To examine the effects of N-terminal ΔG_{RNA} and hydrophilicity on soluble and insoluble rMefp1 expression, we compared N-termini with identical pI values but different ΔG_{RNA} values and/or

hydrophilicity. To reduce the ΔG_{RNA} values of the ME_8-I (Met-$(Glu^{GAA}Glu^{GAG})_4$) and ME_6-I (Met-$(Glu^{GAA}Glu^{GAG})_3$) N-termini (which have ΔG_{RNA} values of -8.30 and -8.50, respectively), we altered Glu codon selection and thus created ME_8-II (Met-$(Glu^{GAA})_7Glu^{GAG}$) and ME_6-II (Met-$(Glu^{GAA})_5Glu^{GAG}$), which have ΔG_{RNA} values of -4.00 and -4.30, respectively (Table 1 and Supplementary Table 1). However, these changes in ΔG_{RNA} did not affect the amount of soluble and insoluble rMefp1 produced (Fig. 2C) and there was no translational arrest or reduction in translational efficiency such as we observed with the $(Arg)_n$-containing N-termini with repetitive Arg codon usage. Thus, altering the ΔG_{RNA} values of longer N-termini did not affect soluble or insoluble rMefp1 expression. Increased hydrophilicity, meanwhile, greatly reduces the production of soluble and insoluble rMefp1 fusion protein.

By contrast, with shorter N-termini lacking calculated hydrophobicity indices, a change in ΔG_{RNA} had a substantial effect on soluble and insoluble rMefp1 expression. (Compare MK_2AK-I and -II (pI 10.82; ΔG_{RNA} -2.10 and -4.70, respectively) and MRAK-I and -II (pI 11.52; ΔG_{RNA} -2.90 and -5.00, respectively) in Fig. 2C.) Therefore, ΔG_{RNA} appeared to control the synthesis of rMefp1 fusion protein with shorter N-termini. The reduction in total rMefp1 expression seen with longer N-termini was caused by their increased hydrophilicity.

These attempts to increase soluble expression of a target protein lacking a TM-like domain showed that N-termini with small ΔG_{RNA} values, low hydrophilicity, and codon usage that prevents translational arrest or a reduction in translational efficiency are needed to maximize the rate of translation, and thus increase total protein synthesis. Furthermore, the amount of soluble product obtained increases as the amount of total protein increases, and when the N-terminus has a pI value that can trigger translocation through inner membrane channels.

CONCLUSION

In this study, we investigated the effect of the pI (in the range 2.73-13.35) of the N-terminal region of 7xMefp1, a protein lacking a TM-like domain (Lee et al., 2008b), on soluble and insoluble expression. We also analyzed the relationships between N-terminal pI value and hydrophilicity, charge, ΔG_{RNA} value, molecular weight, amino acid sequence, and codon selection and repetition. We found that the N-terminal pI value can be used as a comprehensive biological index to represent the level of soluble rMefp1 expression over a wide range because it takes account of all of the factors affecting overall and soluble expression.

In plotting soluble expression against N-terminal pI, we identified three curves, one for each pI range studied (acidic, neutral, and alkaline). We surmise that each such curve is derived from a different periplasmic secretion pathway involving an inner membrane channel that is specific range of N-terminal pI values (acidic, neutral, or alkaline). Each pI range-specific soluble expression curve has a clear form and boundary and does not interfere with the other curves (Fig. 1B). N-terminal pI values may influence membrane channel permeability by guiding appropriate channel selection in a wide pI range, and by modulating the kinetics of channel translocation in the narrow pI ranges that allow soluble expression. Therefore, we suggest that there exist in the *E. coli* inner membrane three types of membrane channels, each specific for target protein N-termini whose pI values fall in certain ranges (acidic, neutral, and alkaline). Collectively, they may constitute a useful index for defining the periplasmic secretion pathways that underpin soluble expression.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENT

This work was supported by a grant from the National Fisheries Research and Development Institute (RD-10-BT-007).

REFERENCES

Bradford, M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. *72*, 248-254.

Hubbert, M.K. (1956). Nuclear Energy and the Fossil Fuels. Drilling and Production Practice. (American Petroleum Institute & Shell Development Co. Publication No. 95), pp. 9-11 and 21-22.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. *151*, 389-409.

Inouye, S., Soberon, X., Franceschini, T., Nakamura, K., Itakura, K., and Inouye, M. (1982). Role of positive charge on the amino-terminal region of the signal peptide in protein secretion across the membrane. Proc. Natl. Acad. Sci. USA *79*, 3438-3441.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.

Lee, S.J., Han, Y.H., Nam, B.H., Kim, Y.O., and Reeves, P.R. (2008a). A novel expression system for recombinant marine mussel adhesive protein Mefp1 using a truncated OmpA signal peptide. Mol. Cells *26*, 34-40.

Lee, S.J., Park, I.S., Han, Y.H., Kim, Y.O., and Reeves, P.R. (2008b). Soluble expression of recombinant olive flounder hepcidin I with a novel secretion enhancer. Mol. Cells *26*, 140-145.

Mukund, M.A., Bannerjee, T., Ghosh, I., and Datta, S. (1999). Effect of mRNA secondary structure in the regulation of gene expression: unfolding of stable loop causes the expression of Taq polymerase in *E. coli*. Curr. Sci. *76*, 1486-1490.

Nossal, N.G., and Heppel, L.A. (1966). The release of enzymes by osmotic shock from *Escherichia coli* in exponential phase. J. Biol. Chem. *241*, 3055-3062.

Overton, E. (1895). Uber die osmotischen Eigenschaften der lebenden Pflanzen und Tierzelle. Vierteljahreschr. Naturforsch. Ges. Zürich *40*, 159.

Ramesh, V., De, A., and Nagaraja, V. (1994). Engineering hyperexpression of bacteriophage μC protein by removal of secondary structure at the translation initiation region. Protein Eng. *7*, 1053-1057.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, 2nd eds. (Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press).

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463-5467.

Waite, J.H. (1983). Evidence for a repeating 3,4-dihydroxyphenylalanine- and hydroxyproline-containing decapeptide in the adhesive protein of the mussel, *Mytilus edulis* L. J. Biol. Chem. *258*, 2911-2915.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406-3415.