BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Sweeping Away Protein Aggregation with Entropic Bristles: Intrinsically Disordered Protein Fusions Enhance Soluble Expression

Aaron A. Santner,[†,∥] Carrie H. Croy,[†] Farha H. Vasanwala,[†] Vladimir N. Uversky,[†,‡,§,⊥] Ya-Yue J. Van,[†] and A. Keith Dunker*[,†,‡]

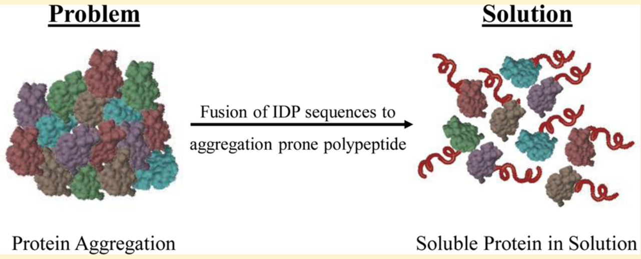
[†]Molecular Kinetics Inc., Indianapolis, Indiana 46268, United States

[‡]Department of Biochemistry and Molecular Biology, Indiana University, Indianapolis, Indiana 46202, United States

[§]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

S *Supporting Information*

ABSTRACT: Intrinsically disordered, highly charged protein sequences act as entropic bristles (EBs), which, when translationally fused to partner proteins, serve as effective solubilizers by creating both a large favorable surface area for water interactions and large excluded volumes around the partner. By extending away from the partner and sweeping out large molecules, EBs can allow the target protein to fold free from interference. Using both naturally occurring and artificial polypeptides, we demonstrate the successful implementation of intrinsically disordered fusions as protein solubilizers. The artificial fusions discussed herein have a low level of sequence complexity and a high net charge but are diversified by means of distinctive amino acid compositions and lengths. Using 6xHis fusions as controls, soluble protein expression enhancements from 65% (EB60A) to 100% (EB250) were observed for a 20-protein portfolio. Additionally, these EBs were able to more effectively solubilize targets compared to frequently used fusions such as maltose-binding protein, glutathione *S*-transferase, thioredoxin, and N utilization substance A. Finally, although these EBs possess very distinct physiochemical properties, they did not perturb the structure, conformational stability, or function of the green fluorescent protein or the glutathione *S*-transferase protein. This work thus illustrates the successful de novo design of intrinsically disordered fusions and presents a promising technology and complementary resource for researchers attempting to solubilize recalcitrant proteins.


Problem
Solution
Fusion of IDP sequences to
aggregation prone polypeptide
Protein Aggregation
Soluble Protein in Solution

The inability to obtain large quantities of functional protein remains a critical limitation for many fields, including structure determination initiatives and modern drug discovery. Proteome-wide structure determination efforts highlight problems with recombinant expression in the preferred *Escherichia coli* host system, with significant problems arising from proteolytic degradation, protein misfolding, and poor solubility. For example, in a study of 424 nonmembrane proteins from the *Methanobacterium thermoautotrophicum* genome, only 50% of the proteins taken through cloning and expression could be purified to a state suitable for structural studies, with ~60% of failures due to poor protein expression levels or insolubility.[1,2] As for the eukaryotic human proteome project, failure rates were 50% for cytoplasmic proteins, 70% for extracellular proteins, and >80% for membrane proteins.[3] Many approaches have been tried for improving soluble expression, but none are generally effective.[4−6]

One of the more effective approaches for improving the solubility, stability, and folding of recombinant polypeptides and/or proteins produced in *E. coli* is to use translational fusion partners.[7−22] The most commonly used fusion proteins include glutathione *S*-transferase (GST),[19] thioredoxin (TRX),[15] N utilization substance A (NusA),[8] and maltose-binding protein (MBP).[9] These four fusions vary in size, structure, and ability to solubilize a given target, but all share the characteristics of being a well-expressed, structured domain or protein. More recently, an elastin-like peptide fusion has been developed for *E. coli* expression.[23,24]

The structured fusions described above likely use three main interrelated mechanisms for enhancing the solubility of the linked target proteins. First, protein solubility can be predicted from amino acid sequence with fairly good reliability.[8,25−27] Thus, linking a soluble protein to an insoluble protein would tend to increase the solubility of the latter by increasing the proportion of solubility-enhancing amino acids. Indeed, NusA was discovered as a solubility-enhancing tag because of its high solubility scores using the Wilkinson−Harrison solubility predictor.[8,27] Second, aggregation requires productive collisions between the proteins. Thus, the soluble fusion partner could help prevent aggregation by simple steric hindrance of the productive collisions. Third, aggregation is thought to be enhanced by segmental interactions between unfolded or partially folded chains.[28,29] Such interactions would be

Received: May 20, 2012
Revised: August 24, 2012
Published: August 27, 2012

weakened if the fusion tag were to stimulate chaperone recruitment or if the fusion protein itself were to act as a molecular chaperone that either slows or reverses segmental aggregation and thereby promotes correct folding.

In this paper, we present our results for a new class of soluble expression enhancing fusions based on intrinsically disordered proteins (IDPs). First, IDP segments are rich in solubility-enhancing polar amino acids, so such segments would be expected to increase the solubility of the protein fusion simply because of their shifts in the overall amino acid composition toward a higher proportion of soluble amino acids. Second, by random movements about its point of attachment, an IDP segment would sweep out a significant region in space and entropically exclude large particles without excluding small molecules such as water, salts, metals, or cofactors.[30] Segments with this property were named "entropic bristles"[30] (EBs). Finally, from studies of a group of intrinsically disordered proteins known as dehydrins, there is substantial evidence that at least some disordered proteins can exhibit chaperone function.[31] Indeed, disordered segments have been suggested to play a role in the structurally characterized chaperone Hsp90.[32] Interestingly, local regions of sequences in several dehydrins show a strong resemblance to local sequences in Hsp90.[31]

Here we test our hypothesis that IDPs can lead to solubility enhancement when they are fused with a collection of insoluble partner proteins. First, we show that the naturally occurring dehydrin IDPs enhance the soluble expression of different partner proteins in *E. coli*. We then describe several artificial polypeptide IDPs that provide solubility-enhancing capacities comparable to or even greater than the capacities of the dehydrin fusions. Comparison of our EB fusions with the commonly used structured fusion proteins demonstrates that IDPs generally outperform several structured solubility enhancer sequences. Finally, we demonstrate the maintenance of biological function and stability for a few of the EB−target hybrids. On the basis of these studies, the resultant expression hybrids provide a promising new approach for preparing soluble proteins that maintain their biological function.

■ EXPERIMENTAL PROCEDURES

Compositional Profiling. Compositional profiling of the dehydrin fusions was conducted using an approach developed for intrinsically disordered proteins.[33,34] Specifically, the fractional difference calculated as $(C_X - C_{\text{reference}})/C_{\text{reference}}$, where C_X is the content of a given amino acid in a disordered protein set and $C_{\text{reference}}$ is the corresponding content in a set of ordered "reference" proteins, was plotted for each amino acid. In Figure 1A, the amino acids are arranged from the most order-promoting to the most disorder-promoting.

Predictions of Intrinsic Disorder. Disorder predictions for different fusions were made using both PONDR VLXT[35,36] and PONDR VSL2 algorithms.[37,38] The PONDR VLXT predictor is a nonlinear neural network classifier and is a result of the merger of three predictors; the PONDR VSL2 algorithm combines two predictors using weights generated by a third meta-predictor. In one recent experiment, PONDR VLXT gave order/disorder prediction accuracies of 67 and 70% on two different data sets containing both structured and disordered proteins, while PONDR VSL2 gave accuracies of 74 and 78% on the same two data sets.[39] Additionally, charge−hydropathy distributions (CH plots) were also analyzed for these proteins using methods described by Uversky et al.[40] The CH plot in Figure 1C is a two-dimensional graph plotting the Kyte−Doolittle hydropathy value[41] of a protein as its *x*-axis coordinate and the mean net charge of the same protein as its *y*-axis coordinate. In these plots, a boundary line demarcates where compact proteins (below) and fully disordered extended proteins (above) cluster.

Prediction of Protein Solubility. The protein solubility predictors used to evaluate the fusions included the sequence-based feature model developed by Wilkinson and Harrison (WH).[27] Critical sequence features with strong correlation of solubility included average charge and turn-forming residue fractions. This WH model design was initially evaluated on a set of 81 proteins and reported an accuracy of 88%.[27] Two newer machine learning predictors SolPro[25] and PROSO[26] were also run on the various fusion sequences. PROSO (PROtein SOlubility predictor) is a machine learning approach trained on a 14200-protein data set and was originally reported with a prediction accuracy of 72%.[26] The SOLpro predictor used a two-tiered SVM strategy trained on 17408 proteins and was originally reported with a prediction accuracy of 74%.[25] Magnan et al. re-evaluated the three predictors side by side with his database and found the accuracy of the WH, ProSo, and SOLpro predictors to be 54, 59, and 74%, respectively.[25]

Design of the Expression Vector. All solubility data presented on artificial EB−target fusions were derived from recalcitrant protein expression cassettes cloned into our pAquoProt expression vector (Molecular Kinetics Inc.) (Figure S2 of the Supporting Information). This vector utilizes an N-terminal hexahistidine and C-terminal HA sequences to allow both purification on affinity resins and immunoassay detection. The sequences encoding the various EB fusions were placed immediately downstream and in frame with the ATG and 6xHis sequences provided in Figure S2 of the Supporting Information. Finally, an enterokinase recognition sequence was encoded between the 6xHis-EB purification domain and the target polypeptide sequence, to allow the user a means for purifying the desired polypeptide from the EB fusion sequence after translation.

Cloning. The coding region for each target protein was amplified by polymerase chain reaction (PCR) with the high-fidelity AccuPrime *Pfx* DNA polymerase (Invitrogen) from their respective cDNA clones using primers designed for use with the In-Fusion Advantage PCR cloning kit (Clontech). The various EB-harboring expression plasmids were digested with the restriction enzyme BamHI (New England Biolabs) and gel purified. The target gene PCR products were then cloned into the BamHI restriction site using a ligation-independent cloning (LIC) method (In-Fusion Advantage PCR, Clontech). Following the cloning reactions, chemically competent Acella cells (EdgeBio) were used for transformation.

Cell Growth and Lysis. Cultures were grown overnight in LB medium supplemented with 100 μg/mL ampicillin at 37 °C. The next morning a 150 μL aliquot of culture was spun down and resuspended in LB medium containing 0.5 M sorbitol and 1 mM betaine for the purpose of inducing expression of endogenous *E. coli* chaperone proteins. The fresh cultures were incubated at 37 °C until they reached an OD_{600} of 0.4, and then expression was induced by addition of 0.2 mM IPTG. Induction of protein expression was conducted for 6 h at the reduced temperature of 25 °C. After this induction period, cells were pelleted by centrifugation and frozen at −20 °C until the expression was analyzed. For soluble protein expression analysis, cell pellets were permeabilized, following the

Table 1. Characteristics, Disorder, and Solubility Predictors for Dehydrin Proteins Originating from *A. thaliana*

					percent[a] disorder					
protein	accession number	MW (kDa)	net charge	length (no. of amino acids)	VLXT	VSL2	mean[b] hydropathy	WH[c] solubility predictor	ProSolI[d] solubility predictor	SolPro[e] solubility predictor
ERD10	NP_564114	29.4	−15	259	64	100	0.3529	75% soluble	72% soluble	81% soluble
ERD14	NP_177745	20.8	−9	185	64	100	0.3595	65% soluble	32% insoluble	82% soluble
COR47	NP_195554	18.0	−5	163	60	100	0.319	52% insoluble	83% soluble	92% soluble
Rab18	CAA48178	18.5	0	186	80	100	0.3686	97% insoluble	55% insoluble	93% soluble
XERO1	NP_190667	13.4	+3	128	60	100	0.5939	97% insoluble	86% soluble	93% soluble
LTI30	NP_190666	20.9	+6	193	21	100	0.3697	94% insoluble	86% soluble	92% soluble

[a] Percentage of amino acids found disordered using the PONDR predictors VLXT[35,36] and VSL2.[37,38] [b] A measure that distinguishes ordered from disordered proteins.[40] [c] The revised Wilkinson−Harrison solubility predictor.[8] [d] The soluble probability value ranges from 0−0.6 (insoluble) to ≥0.6−1.0 (soluble).[26] [e] The probability values for the soluble and insoluble designation range from 0.5 to 1.0.[25]

manufacturer's suggested conditions, under isotonic conditions using a solution containing both a mild nonionic detergent (B-PER Reagent, Pierce) and DNaseI (Sigma-Aldrich). Cell disruption was promoted with vortexing. The resultant cell "lysis" solution was designated as the "total cell extract". The "soluble fractions" and "pellet fractions" were then separated by a moderate-speed centrifugation (10000*g* for 5 min) capable of pelleting large cellular debris and subcellular structures, e.g., mitochondria. The total cell extracts, soluble fractions, and pellet fractions were used for the detection of protein expression and solubility.

Expression and Solubility Test. To evaluate protein expression and solubility, the total cell extract (T), soluble fraction (S), and pellet fraction (P) were separated by sodium dodecyl sulfate−polyacrylamide gel electrophoresis (SDS−PAGE) using the NuPAGE Bis-Tris gradient gel system (Invitrogen). The proteins were transferred to PVDF membranes (Invitrogen) and probed with an anti-His antibody (Santa Cruz Biotechnology, G-18) following a standard Western blotting protocol. Following development, the protein gel blots were scanned, and the pixel density between the soluble and pellet fractions was quantitated using ImageJ (National Institutes of Health).

GST Purification and Activity Assay. Following the cell growth and lysis procedure described above, the GST fusions, His-GST, MBP-GST, EB60A-GST, EB60B-GST, EB144-GST, or EB250-GST, were enriched from the soluble protein fraction using a glutathione column. Specifically, 1 mL of the soluble lysate containing the GST fusion protein was incubated with 0.25 mL of glutathione resin for 1 h at 4 °C while being mixed. Resin was washed with 6 volumes of Dulbecco's phosphate buffer (D-PBS) and then eluted with 3 volumes of D-PBS containing 50 mM glutathione. The eluate concentrations were determined with a Bradford assay (Coomassie protein assay kit, Thermo Scientific).

The GST transferase activity was determined by measuring the coupling of reduced glutathione to a 1-choloro-2,4-dinitrobenzene (CDNB) (Sigma) substrate by observing the increasing absorbance at 340 nm. Specifically, the reaction was initiated when ~20 pmol of the enriched GST fusion protein was added to a 1 mL quartz cuvette containing 2 mM glutathione and 1 mM CDNB in Dulbecco's PBS. Using a kinetics program, the change in the 340 nm reading was measured every 30 s for 5 min on a Varian Cary Eclipse fluorescence spectrophotometer. The GST specific activity was then calculated as the micromoles of CDNB converted per minute per picomole of GST enzyme; all activities were compared to a commercially available active GST standard (Biovision).

GFP Fluorescence and GndHCl-Induced Unfolding. Following the cell growth and lysis procedure described above, the various GFP fusions were partially purified using TALON Superflow Metal Affinity Resin (Clontech). Specifically, 1 mL of the soluble lysate containing the GFP fusion protein was incubated with 0.1 mL of TALON resin for 20 min at 25 °C with rotation. The resin was recovered in a column, washed with 10 volumes of wash buffer [50 mM Tris (pH 8.0), 150 mM NaCl, and 5 mM imidazole], and then eluted with 3 volumes of buffer containing 150 mM imidazole. The eluate concentrations were determined by a Bradford assay (Coomassie protein assay kit, Thermo Scientific).

Protein samples were incubated in the presence of various concentrations of GndHCl at room temperature for 17−72 h. GFP unfolding was monitored using fluorescence spectrophotometry (Cary Eclipse, Varian); the excitation wavelength was 395 nm, and emission was detected at 510 nm.

Enterokinase Cleavage. To demonstrate that enterokinase can cleave the DDDDKS consensus sequence located between the EB and target sequences, 5 μg of purified EB−GFP fusion protein was incubated with 1.5 units of recombinant enterokinase protease (Novagen) over 24 h. The efficiency of the digestion was monitored at discrete time points of 0, 2, 4, 8, and 24 h using a Coomassie-stained SDS−PAGE gel.

■ RESULTS

Characterization of the Intrinsically Disordered Dehydrin Family of Proteins. Evidence that intrinsically disordered proteins (IDPs) may function as molecular chaperones led us to design a recombinant IDP fusion system and test whether this system enhances protein recovery for targets recalcitrant to soluble expression from recombinant bacterial systems. Toward this end, we first analyzed the primary sequence characteristics of a family of plant proteins known as dehydrins, because they have been shown to be intrinsically disordered, to have potential chaperone activity,[42−45] and to function as both an antiaggregant and an enzyme preservation agent.[31,32,46−54] Moreover, two family members have been shown to solubilize membrane proteins identified as recalcitrant to overexpression.[55] Table 1 shows the compilation of characteristics of both the disorder and solubility predictions for the six known *Arabidopsis thaliana* dehydrin proteins. If we focus on the disorder predictors values presented in Table 1 first, PONDR predictors VLXT and VSL2 verify high "percent disorder" values for *A. thaliana*

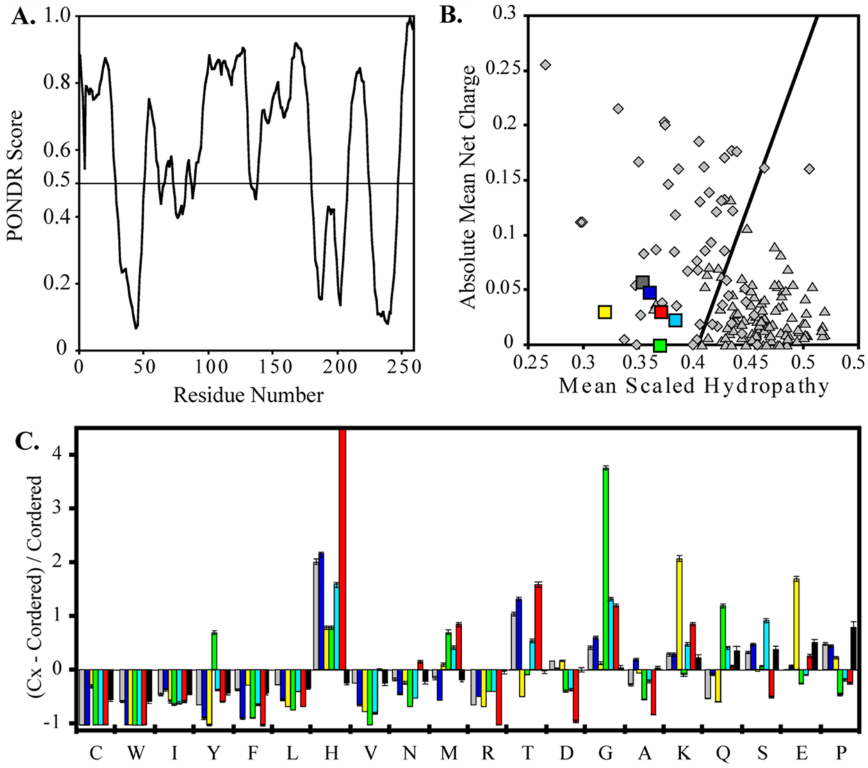


Figure 1. Evaluation of intrinsic disorder in the members of the *A. thaliana* dehydrin family. (A) PONDR VLXT analysis of ERD10. (B) CH plot analysis of dehydrins (squares), ERD10 (gray), ERD14 (blue), COR47 (yellow), Rab18 (green), Xero1 (cyan), LTI30 (red), ordered proteins from DISPROT (light gray triangles), and disordered proteins from DISPROT (light gray diamonds). The black line delineates the boundary above which the most extended, disordered proteins are located. (C) Compositional profiling of dehydrins: Disprot (black), ERD10 (gray), ERD14 (blue), COR47 (yellow), Rab18 (green), Xero1 (cyan), and LTI30 (red).

sequences ERD10, ERD14, COR47, Rab18, and Xero1.[35,36] Figure 1A shows the 64% sequence disorder predicted by VLXT for ERD10 can be attributed mainly to a 90-residue stretch in the central part of the protein (residues 90−132 and 138−179). Further examination of dehydrins by measuring the mean hydropathy indicates that all six dehydrins lie on the disordered side of the boundary and five of the dehydrins have significant net charge and so very likely have extended disordered structures under physiological conditions (Figure 1B). The remaining dehydrin has zero net charge and so might be a collapsed but disordered protein. Finally, the residue abundance plot shown in Figure 1C visually demonstrates how the primary sequences of the *A. thaliana* dehydrin proteins are consistent with disordered polypeptides. Briefly, disordered polypeptides are significantly depleted of bulky hydrophobic and aromatic residues that would normally form the hydrophobic core of a folded globular protein and also possess a low content of Cys and Asn residues.[56,57] Hence, these residues, W, Y, F, I, L, V, C, and N, were proposed to be called order-promoting amino acids while polar amino acids such as A, R, G, Q, S, E, K, and structure-breaking P were called disorder-promoting amino acids.[33,34,36,58,59] As the residue abundance plot presented in Figure 1C progresses from the order-promoting to disorder-promoting residues, we can observe the six dehydrin sequences are both deficient in several order-promoting residues (W, C, and F) and enriched in several disorder-promoting residues (M, E, and K), indicating that they are consistent with the composition of a disordered protein. Additionally, Figure 1C also reveals a consistent deviation of these dehydrin sequences from proteins in general, namely an enrichment of compositionally rare His residues (typically around 2%[60]). This higher His proportion could be functionally important as members of the family have been shown to bind several divalent metals using a conserved His-containing sequence, e.g., the HKGEHHSGDHH core sequence for Cu^{2+} binding by citrus dehydrin CuCOR15.[53,60,61]

Table 1 also demonstrates that dehydrins have favorable soluble expression characteristics by multiple predictor algorithms. The early Wilkinson−Harrison solubility model found only the COR47, ERD10, and ERD14 dehydrin sequences favor the soluble protein when expressed as an independent polypeptide.[8,27] This result can be explained by the fact that this model favors sequences with (a) the presence of "turn-forming" residues N, G, P, and S and (b) a mean net negative. Though five of six dehydrins have favorable enrichments in turn-forming residues for Gly for LTI30, Rab18, and Xero1 and Pro for ERD10 and ERD14, only when coupled with the higher net negative charges of COR47, ERD10, and ERD14 does the calculation tip the probability toward soluble protein expression. Application of the two more recently developed solubility predictors, SOLpro and PROSO,[25,26] both of which use a machine learning approach, broadly shows more favorable solubility scores for the dehydrin proteins than the Wilkinson−Harrison model. Overall, only ERD10 showed a favorable consensus among all three predictors.

Intrinsically Disordered Dehydrin Proteins Enhance Soluble Protein Expression When Used as Fusion Partners to Recalcitrant Proteins. To empirically evaluate the validity of the computational implications, we tested the

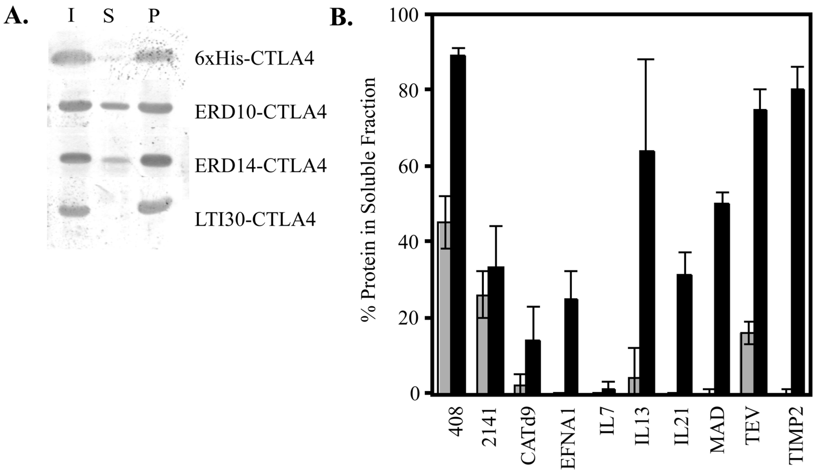


Figure 2. Analyzing the solubilization capabilities of dehydrins. (A) SDS−PAGE analysis of CTLA4 solubility alone or fused to ERD10, ERD14, and LTI30 dehydrins. (B) Ability of the ERD10 fusion (black bars) to enhance the percentage of soluble protein yield compared to that of the 6xHis fusion control (gray bars) for 10 insoluble proteins.

potential solubility enhancing properties of dehydrins for several known insoluble protein targets in a dehydrin fusion system, including CTLA4 shown in Figure 2. Figure 2A shows that, as predicted, CTLA4 when fused to ERD10 and ERD14 sequences yields higher levels of soluble protein than the disordered LTI30 fusion. After minimal success with the neutral Rab18 and positive LTI30 dehydrins, we proceeded to evaluate the ability of the negatively charged dehydrins to aid soluble protein expression (data not shown). Overall, ERD10 and ERD14 showed 74 and 54% rates of success, respectively, for solubilizing the target protein (data not shown). Figure 2B summarizes the ability of ERD10 to enhance the percentage of soluble protein yield for a subset of 10 proteins where multiple experimental sets were collected to allow statistical evaluations. In six of the 10 cases, ERD10 significantly ($p < 0.05$) enhanced the soluble expression of protein targets previously cited in the literature to be recalcitrant to soluble expression in an *E. coli* system.[7,10,13,62−64] These data illustrate that IDP-based fusions successfully enhance soluble protein expression at least for some proteins.

Given the positive results discussed above, we set out to design completely artificial disordered sequences, namely de novo EBs, to serve as solubility enhancers. The rationale for developing artificial disordered sequences is to allow a wide-ranging design of the potentially solubilizing sequences.

Design of Artificial EB Fusion Polypeptides. Designing artificial EB sequences that retain the desirable solubility properties described above for the natural ERD bristles allowed us to uncouple the importance of the disordered nature of the dehydrins from their in vivo biological functions. Additionally, it gave us the ability to minimize the negative cytotoxic effects observed for several natural IDP sequences when the method was attempted in recombinant expression systems.[65] All polypeptides were designed to have low complexity, have net negative charge, and be composed primarily of disorder-promoting residues (Table 2). Table 3 verifies that these sequences are disordered and predicted to be highly soluble. Additionally, the CH plot shown in Figure S1 of the Supporting Information verifies their localization within the trapezium of extended IDP proteins. Finally, pilot expression studies verified that our artificial EBs displayed no obvious cellular toxicity.

Table 2. Characteristics of Artificial Entropic Bristles (EB)

EB fusion	amino acid composition	EB length	MW (kDa)	net charge	pI
EB60A	E-P-Q-S	60	6.8	−24	3.08
EB60B	E-P-Q-G	60	6.7	−25	2.97
EB144	D-E-P-Q-S-G	144	15	−41	2.69
EB250	D-E-P-Q-S-G-I-L-M-F-V	250	26.1	−65	2.48

To explain our design rationale, we will briefly describe how the de novo EB templates listed in Table 2 were created. The residues at the far-right side of Figure 1C represent the most disorder-promoting residues (Q, S, E, and P) and were chosen as constituents of EB60A, in a 2:2:1:1 E:P:Q:S proportion. The rationale for the proportions follows: a high Glu proportion was used because proteins with high net charge densities were found to function as effective intramolecular chaperones;[43,44,66,67] a high Pro content would disrupt secondary structure (except for the polyproline II helix) and contain hydrophobic surfaces for weak binding to possible aggregation patches; Gln was chosen because it is a strongly disorder-promoting residue but was kept at a low proportion (1:6) to avoid the aggregation propensity of polyQ sequences; and Ser was chosen because it not only is hydrophilic but also exhibits one of the largest conformational variabilities of the 20 amino acids.[68] On the basis of such considerations, 360-nucleotide sequences were randomly generated with codon optimization for expression in *E. coli*. This synthetic gene encodes the 120-residue polypeptide that serves as the basis for our de novo EB sequence fusion. Because serines are common sites for posttranslational modification, we also designed an EB60B series of de novo EBs, in which serines were replaced with the disorder-neutral residue G (EB60B, 2:2:1:1 E:P:Q:G). Additionally, as we generated EBs with higher negative charge densities and increased lengths, to avoid potential issues with expression problems associated with high levels of sequence redundancy, we added a larger subset of disordered residues; for example, EB144 uses a 1:2:2:1:2:1 D:E:P:Q:S:G composition. Finally, EB250 was designed with the 1:2:2:1:2:1 D:E:P:Q:S:G template but additionally had several hydrophobic patches to mimic those found in the dehydrin proteins. In all, eight negatively charged EB templates were created. Advantageously, these eight sequences could be developed into

Table 3. Disorder and Solubility Predictions for Various Protein Fusions

	percent[a] disordered					
protein	VLXT	VSL2	mean[b] hydropathy	WH[c] solubility predictor	ProSolI[d] solubility predictor	SolPro[e] solubility predictor
EB60A	100	100	0.2281	97% soluble	88% soluble	97% soluble
EB60B	100	100	0.2133	97% soluble	89% soluble	100% soluble
EB144	100	100	0.2640	86% insoluble	88% soluble	97% soluble
EB250	100	100	0.3424	92% insoluble	87% soluble	97% soluble
GST	12	14	0.4572	58% soluble	58% insoluble	78% insoluble
MBP	11	17	0.4640	52% insoluble	77% soluble	94% soluble
NusA	47	20	0.4423	95% soluble	66% soluble	58% soluble
Trx	6	12	0.028	72.6% soluble	36% insoluble	89% soluble

[a]Percentage of amino acids found disordered using the PONDR predictors VLXT[35,36] and VSL2.[37,38] [b]A measure that distinguishes ordered from disordered proteins.[40] [c]The revised Wilkinson−Harrison solubility predictor.[8] [d]The soluble probability value ranges from 0−0.6 (insoluble) to ≥0.6−1.0 (soluble).[26] [e]The probability value for the soluble and insoluble designation ranges from 0.5 to 1.0.[25]

Table 4. Summary of Solubility Data (percent) for 20 Test Protein Candidates

	6xHis	MBP	EB60A	EB60B	EB144	EB250
408	55 ± 2	66 ± 6	68 ± 5	79 ± 7	71 ± 15	89 ± 5
2141	18 ± 9	17 ± 18	39 ± 2	33 ± 7	56 ± 17	59 ± 9
CATΔ9	3 ± 2	0 ± 0	1 ± 1	0 ± 1	58 ± 6	65 ± 5
EFNA1	4 ± 3	13 ± 4	13 ± 4	4 ± 1	51 ± 7	54 ± 7
FLCN	2 ± 3	0 ± 0	0 ± 0	7 ± 8	94 ± 7	88 ± 3
GADD45	45 ± 8	55 ± 2	57 ± 7	62 ± 6	92 ± 4	94 ± 3
GFP	56 ± 12	35 ± 1	52 ± 8	45 ± 2	84 ± 14	87 ± 15
ID2	42 ± 6	91 ± 8	78 ± 2	90 ± 11	97 ± 4	83 ± 4
IL-7	0 ± 0	10 ± 1	22 ± 2	16 ± 2	25 ± 6	43 ± 2
IL-13	18 ± 1	48 ± 3	81 ± 10	97 ± 3	84 ± 7	84 ± 11
IL-21	0 ± 0	30 ± 5	46 ± 6	57 ± 6	70 ± 8	79 ± 9
MAD	4 ± 2	21 ± 2	41 ± 2	43 ± 3	86 ± 2	92 ± 2
MTHFS	17 ± 3	10 ± 2	15 ± 3	31 ± 9	75 ± 7	36 ± 9
MSTN	0 ± 0	25 ± 3	27 ± 1	45 ± 2	55 ± 16	74 ± 2
PHB	0 ± 0	27 ± 11	26 ± 8	6 ± 2	39 ± 5	43 ± 4
SNW1	7 ± 4	4 ± 5	38 ± 2	58 ± 10	82 ± 3	85 ± 2
TEV	2 ± 3	66 ± 3	79 ± 2	89 ± 11	89 ± 1	57 ± 27
TIMP2	2 ± 4	30 ± 4	45 ± 2	37 ± 4	41 ± 3	66 ± 3
TNSF13b	2 ± 3.5	1 ± 1.2	26 ± 4	30 ± 5	30 ± 5	31 ± 2
WAG2	0 ± 0	9 ± 3	6 ± 4	7 ± 6	46 ± 1	89 ± 8
% *p* value vs His of <0.05	na	65	75	75	95	100
% *p* value vs MBP of <0.05	20	na	50	50	85	85

a very large number of de novo EB fusions. Table 2 summarizes the amino acid compositions, ratios of amino acids, and lengths of the four EB fusions used to present the soluble expression efficiency and activity studies below (Figure S2 of the Supporting Information gives the exact polypeptide sequence for each EB).

The same analyses conducted on the six dehydrins (Table 1) were repeated on the four artificial EBs and four structured solubility-enhancing proteins (Table 3). As expected from the criteria implemented in the design of these intrinsic disorder-based solubilizers, these EBs have very high solubility scores, higher than those of the six *A. thaliana* dehydrins (Table 1) and higher than those of the structured solubility enhancers MBP, GST, NusA, and Trx.

Evaluation of Artificial EB Fusions for Enhancing Soluble Protein Expression. The extent of soluble expression enhancement provided by the four compositionally unique EB fusion sequences (Figure S2 of the Supporting Information) was determined using the widely used *E. coli* recombinant system. The level of soluble expression of each fusion protein was calculated as the percentage of the hybrid polypeptide in the soluble and insoluble cellular fractions by image density analysis. Note that the permeabilization and sample preparation conditions used do not remove proteins in a soluble aggregated form. However, we evaluated protein size by native gels for various GST-EB constructs and observed a discrete, single species (data not shown). Table 4 summarizes the percentage of soluble expression for target fusions previously reported in the literature to be insoluble when expressed in *E. coli*, with the green fluorescent protein (GFP) being an exception and being used as a control.[7,10,13,14,62−64,69−71] Each percentage shown represents the values determined for three independent growths of different expression clones. Using a 6xHis fusion as a control population, we were able to show successful enhancement of soluble protein expression when the target was fused to EB60, EB144, and EB250 in 75, 95, and 100% ($p \leq 0.05$) of the test candidates, respectively. Interestingly, we found that length was a more important determinant than composition. Specifically, the two EBs that were 60 amino acids in length performed similarly but were less successful than the longer 144- and 250-amino acid fusions. Follow-up studies in which EBs of identical

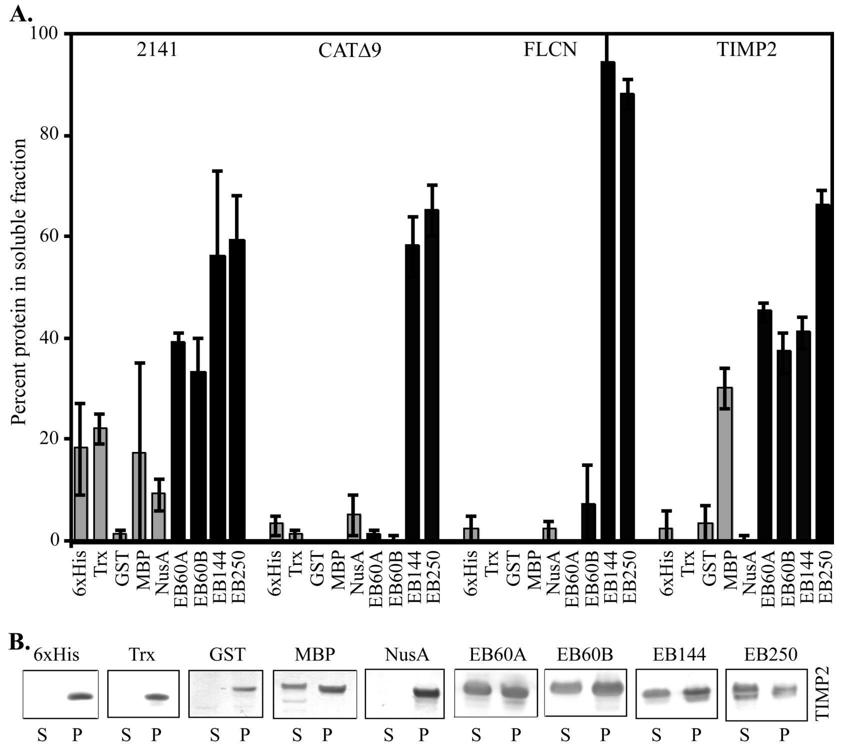


Figure 3. Soluble protein expression comparison of EB sequences with commonly used fusions. (A) Bar graph showing the soluble expression performance of EB60A, EB60B, EB144, and EB250 (black bars) with Trx, GST, NusA, and MBP fusions (gray bars) marketed to enhance soluble protein expression. (B) Western blot analysis of TIMP2 hybridized to the various fusion partners (anti-6xHis blot, G-18 from Santa Cruz Biotechnology).

composition but varying sequence order and length will be conducted to further compare the effects of composition, primary sequence, and length on solubility. Although Table 4 supports the idea that soluble expression levels vary for a given target, the widespread success of EB144 and EB250 supports the notion that a universal IDP-based solubilizer could be a reasonable goal. However, considering certain target protein fusions or downstream applications, development of multiple EB fusions, e.g., serine-free tags like EB60B, is still warranted. The strength of an artificial scaffold is that we maintain almost infinite flexibility to vary composition, length, and physiochemical characteristics as needed.

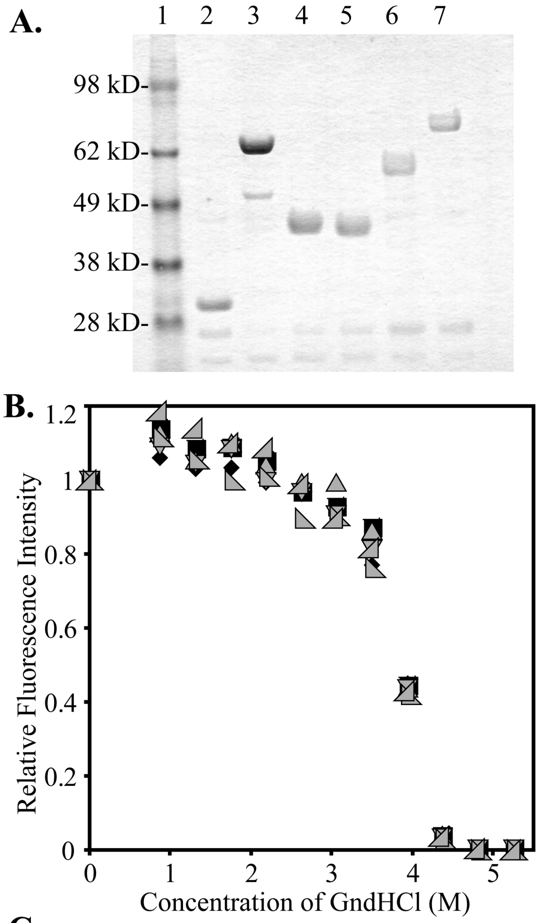
Comparison between the Artificial EB Fusions and Various Fusion Tags Frequently Used To Enhance Soluble Protein Expression. As indicated above, several commonly used solubility-enhancing fusion tags include Trx, GST, NusA, and MBP. All of these fusion peptides are highly soluble proteins, which for the most part agree with the data in Table 3. Table 4 summarizes how our target EB fusion portfolio performed in comparison with the MBP fusion. EB144 and EB250 showed the greatest enhancements with 17 of 20 targets expressing more soluble protein than the MBP hybrid. Next, we selected four of our translatable gene targets (2141, CATΔ9, FLCN, and TIMP2) to perform a side-by-side comparison of four EB fusions with all four of the commonly used structured fusion tags mentioned above. Figure 3 demonstrates that, for these four insoluble proteins, the four artificial EB fusion tags significantly outperform the four commonly used soluble structured protein fusion tags. The Western blots shown in Figure 3B for TIMP2 demonstrate that the expression levels of our fusions are also comparable to those of other fusion systems.

The EB fusions have very distinct physiochemical properties compared to the commonly used solubility-enhancing structured protein. Thus, the question of whether the unusual properties relating to the charge and intrinsic disorder of the EB domains would affect the stability and biological function of the fused partner arises. To test the effects of EB domain fusions on protein folding and stability, fusions with green fluorescent protein (GFP) were studied, and to test the effects of EB domain fusions on function, fusions with the GST enzyme were studied.

Conformational Stability of Recombinant GFP Fusion Proteins. GFP is a member of the fluorescent protein family, members of which harbor a unique chromophore, *p*-hydroxybenzylideneimidazolidone, near the center of a *β*-can that comprises the majority of the folded protein.[72,73] The spectroscopic characteristics of GFP are determined by the local environment of the chromophore.[74] Incubation of GFP in the presence of concentrated solutions of GndHCl can cause the protein to unfold, leading to a decrease in fluorescence intensity that can be easily measured. This trait was exploited to compare the relative stability of various GFP fusion proteins.

The first observation was that GFP with an added 6xHis tag and the various EB fusions all yielded a fluorescent protein with spectral properties similar to those of the original nontagged proteins (data not shown). Because correct folding is required for chromophore formation, these observations show that neither the MBP nor the EB fusion sequences significantly inhibited GFP folding.[75,76]

GFP has unusually slow unfolding kinetics in GndHCl, taking ~3 days to reach quasi-equilibrium after being transferred to unfolding conditions.[74] Six recombinant GFP fusions were partially purified utilizing the 6xHis tag that is present in each recombinant protein (Figure 4A). The fusion proteins were then incubated in the presence of increasing GndHCl concentrations for up to several days at room temperature. Consistent with previous findings, 6xHis-GFP fluorescence was found to increase slightly in the presence of low GndHCl concentrations (<2 M) and then decrease in a GndHCl concentration-dependent manner (Figure 4B[74]). The decrease in fluorescence at each GndHCl concentration was dependent on the incubation time and closely resembled GndHCl-induced unfolding curves that have been reported previously for eGFP.[74] The GFP proteins fused with either MBP or an EB domain in addition to the 6xHis moiety all had GndHCl-induced unfolding curves that were nearly identical to that of the 6xHis-GFP control (Figure 4B). Thus, these data suggest that the translationally fused entropic bristles do not disrupt folding of GFP and normal formation of its unique chromophore, nor do the various fusion tags alter the stability of the folded GFP structure.



C.

GST fusion	MW* (kD)	Transferase # Activity
GST-standard	52.0	440 ± 20
6xHis-GST	56.3	440 ± 10
MBP-GST	136.7	440 ± 10
EB60A-GST	70.0	370 ± 20
EB60B-GST	69.8	390 ± 40
EB144-GST	86.5	420 ± 10
EB250-GST	108.9	410 ± 20

Figure 4. Effect of fusion of various EBs on the conformational stability of GFP and transferase activity of GST. (A) Coomassie gel image showing the enrichment of GFP after purification on a NiNTA column: lane 1, molecular weight standard (Invitrogen); lane 2, 6xHis-GFP (29.3 kDa); lane 3, 6xHis-MBP-GFP (69.5 kDa); lane 4, 6xHis-EB60A-GFP (36.1 kDa); lane 5, 6xHis-EB60B-GFP (36.0 kDa); lane 6, 6xHis-EB144-GFP (44.4 kDa); lane 7, 6xHis-EB250-GFP (55.5 kDa). (B) Fluorescence unfolding curves of GFP constructs fused to the 6xHis tag (◆), MBP (■), EB60A (▲), EB60B (▼), EB144 (right-facing triangle), and EB250 (left-facing triangle). (C) Transferase activity of purified GST fusions toward the synthetic CDNB substrate. Note that the molecular weight is based on the dimer weight of GST and the GST transferase activity assay was measured in units of micromoles of CDNB per minute per picomole of GST to correct for differences in the molecular weight of the various fusions.

Transferase Activity of Recombinant Glutahtione *S*-Transferase (GST) Fusion Proteins. To determine whether the EB fusions interfere with the biological function or protein dimerization of the translational fusion partner, we measured the transferase activity of GST toward the synthetic substrate 1-chloro-2,4-dinitrobenzene (CDNB), a dimerization-dependent activity. The GST-catalyzed rate of conjugation of CDNB with reduced glutathione was determined by standard methods.[77,78] The values reported in Figure 4C are triplicate data points assayed from 6xHis-EB-GST fusion samples enriched after purification on a metal affinity column. Values were normalized to the GST-active standard purchased from Biovision (catalog no. 1243-1) and are reported on a per mole basis because of the size variation among different EB fusions. Lysates purified from different clones and grown on different days yielded results similar to the data shown in Figure 4C. Retention of at least 84% activity, EB60A-GST compared to a standard, indicated that the EB fusions do not interfere with the biological function of GST to conjugate the glutathione onto the synthetic CDNB substrate.

EBs Can Be Removed by Proteolytic Cleavage. To accommodate a potential need to remove fused EBDs for subsequent functional and structural studies of target proteins, a specific enterokinase cleavage site was introduced between the EBD and the target protein. Because EBDs act via constant random motion about their attachment points and therefore sweep out a region of three-dimensional space and sterically exclude other large molecules from that area, it seemed possible that the highly mobile sweeping tail might prevent or slow the rate of proteolytic digestion. However, as Figure 5 illustrates, the EB sequence can be efficiently removed post-translation by proteolytic cleavage by enterokinase in a sequence-specific manner, specifically at the enterokinase cleavage sequence inserted between the EB and the target protein. It is important to remember, however, that the removal of the EB sequence post-translationally may result in aggregation or precipitation of the fusion partner, if the EB sequence is indeed preventing the self-association of the target protein.

■ DISCUSSION

Fusing a collection of aggregation-prone proteins to a highly soluble protein partner is shown to improve the solubility of some aggregation-prone proteins but not others.[7−11,13−22,63] In contrast, fusing long, intrinsically disordered polypeptides called entropic bristles onto an aggregation-prone protein leads to the almost universal improvement of protein solubility, with longer and more negatively charged EBs showing slight enhancements compared to shorter and less charged EBs. More

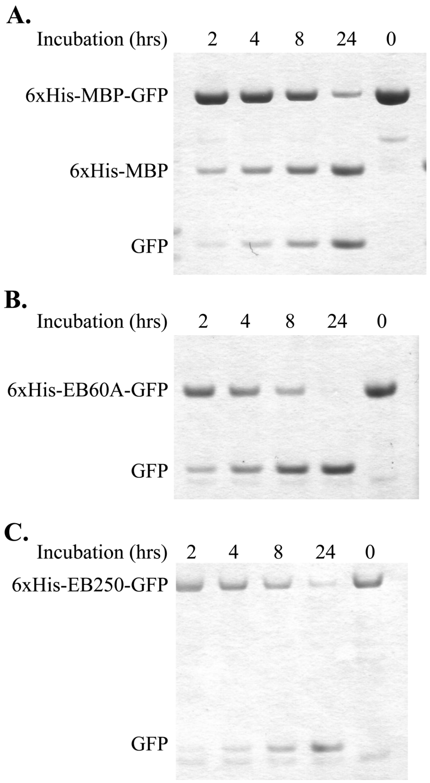

A.
Incubation (hrs) 2 4 8 24 0
6xHis-MBP-GFP
6xHis-MBP
GFP
B.
Incubation (hrs) 2 4 8 24 0
6xHis-EB60A-GFP
GFP
C.
Incubation (hrs) 2 4 8 24 0
6xHis-EB250-GFP
GFP

Figure 5. Removal of the EB fusions via proteolytic cleavage using enterokinase (EK). (A) Coomassie gel image visualizing the cleavage of the 6xHis-MBP fusion from GFP at 0, 2, 4, 8, and 24 h. (B) Coomassie gel image visualizing the cleavage of the 6xHis-EB60A fusion from GFP at 0, 2, 4, 8, and 24 h. (C) Coomassie gel image visualizing the cleavage of the 6xHis-EB250 fusion from GFP at 0, 2, 4, 8, and 24 h.

work is needed to understand secondary effects that likely arise from differences in the details of the EB amino acid sequences and from the interplay between the EB sequences and the sequence and structure of each given recalcitrant protein. The general success of a variety of EBs with significant sequence differences suggests that, compared to water-soluble, structured proteins, EBs have particular properties that significantly improve the solubility of the fused construct.

Computer algorithms not only identify proteins that are likely to be problematic from a solubility perspective[1,8,25−27,79−82] but also have been used to discover a novel solubility-enhancing fusion, e.g., NusA.[8] Solubility-promoting sequence characteristics include high charge and turn-forming residue content and distribution,[27] lower hydrophobic and aromatic residue content,[1,2,79,80] and overall length.[80] Table 1 shows that the predictors identify the commonly used fusions GST, MBP, NusA, and Trx as likely to be highly soluble with some exceptions. When they are applied to the EB fusions, there is a complete agreement among predictors that all these fusions would be soluble. For the untagged targets predicted to be insoluble, the addition of the EB partner was sufficient to shift the sequence composition to a soluble probability in all cases except for one prediction using the WH predictor for the WAG2 protein. Cloning and soluble expression analyses verified the validity of these predictions and showed that the designed IDP fusions were indeed good solubility enhancer fusions. These results support the utility of using these computer algorithms for assessing whether a give EB will likely solubilize a given protein, but more work is needed to determine the reliability of using these algorithms for this purpose.

Solubility enhancement probably involves the following three factors. (1) To a first approximation, the free energies for solubilization are additive over the protein surface,[25,26,83] so adding soluble surface should increase the overall solubility. (2) The highly soluble partner restricts the opportunities for intermolecular interactions between molecules of the aggregation-prone protein. (3) The highly soluble partner provides chaperone activity. Determining the relative contributions from each of these mechanisms would be very difficult, but it is possible to compare structured and disordered proteins with respect to their expected relative contribution to each of these three factors.

Compared to a structured protein with the same number of amino acids, an IDP would contribute a much larger surface for favorable interaction with water. Indeed, an IDP would resemble to some degree the chemical attachment of polyethylene glycol (PEG), which markedly increases protein solubility,[84] likely because of its favorable interactions with water. Likewise, the disordered dehydrin proteins coordinate larger amounts of water per solvent-exposed residue than do folded proteins,[85,86] perhaps because of the presence of polyproline II-type helices, which have larger solvent-accessible surface areas compared to other types of secondary structural elements.[31,87]

Compared to a structured protein with the same number of amino acids, an IDP would provide a much larger excluded volume. In fact, collections of unstructured polymers that enhance solubility have been given the special name of entropic brush.[88] Indeed, entropic brushes based on a variety of polymers have been used to reduce the level of aggregation of particles such as latex particles in paints and to stabilize a wide variety of other colloidal products.[88] From this background on entropic brushes, highly mobile single chains were called “entropic bristles”[30] or EBs. Polypeptide EB domains and other EBs such as PEG probably employ both their large favorable surface area and excluded volume effects to enhance solubility. Indeed, for such molecules, these two factors are highly interrelated. It is because of their affinity for the solvent that such polymers can adopt random-walk configurations in solution,[89] leading to both large favorable interaction surfaces and large excluded volumes.

As for chaperone activity, evidence that TRX, MBP, and NusA fusions utilize chaperone activity has been reported.[78,90−92] In the MBP example, MBP utilizes a hydrophobic surface to bind to the misfolded region and promote refolding.[11] In the NusA example, the protein itself is not the chaperone but instead may help direct recombinant proteins to the endogenous *E. coli* GroEL/GroES chaperone pathway, thereby indirectly improving the native folding characteristics of the fusion partner.[93] Disordered dehydrin proteins, ERD10 and ERD14, are also effective chaperones, preventing aggregation and inactivation of several globular proteins in vitro.[42] Furthermore, there is a growing body of evidence that the disordered regions within chaperones are responsible for their support of protein folding (reviewed in ref 45).

Overall, IDPs likely outperform structured proteins with regard to all three of the factors given above and suggested to promote protein solubility, thus providing a rationale for the better performance exhibited by the IDPs in comparison to that of the structured protein fusion tags.

This work reveals that dehydrin protein family members, ERD10 and ERD14, are effective fusion partners for improving the solubility of aggregation-prone proteins expressed in *E. coli* (Figure 2). Because it is impossible to uncouple the importance of the disordered nature (entropic bristle-like features) of the dehydrins from their in vivo biological functions,[94−96] we next designed a library of low-complexity synthetic polypeptides that are disordered and sample a variety of net charges, charge densities, and lengths for use as de novo entropic bristle fusions.

When these novel polypeptides were assessed, it became clear that the EB polypeptides with net positive charges were not only ineffectual as solubility enhancers but also in some cases detrimental to the overall solubility of the fusion proteins (data not shown). Perhaps their association with negatively charged phospholipid headgroups in membranes and phosphate groups in nucleic acid backbones leads to an apparent lack of solubility when the cells are lysed. Regardless, we did find that net negatively charged EBs significantly improve the solubility of aggregation-prone proteins when compared with that of either a 6xHis or MBP fusion (Table 4). This is consistent with previous studies showing that increased negative charge enhances solubility.[97]

Perhaps a fourth mechanism of enhancing protein solubility is the ability of the highly charged tail to shift the overall isoelectric point of the target protein. That is, amphoteric molecules, including proteins, have long been known to show significantly reduced solubility and even precipitation at or near their isoelectric points.[98−100] A highly charged tail would shift the overall isoelectric point to values outside the range of pH values typically used for protein studies and would thus minimize solubility decreases that could arise from being close to the isoelectric point.

In summary, we have developed a novel set of artificial EB fusions designed on the principles of intrinsic disorder phenomena that are highly effective at improving the soluble expression of heterologous proteins in *E. coli*. These fusions are highly flexible, highly charged polypeptides that will maintain a random coil conformation in solution. When attached to an aggregation-prone protein, the fusion tag extends from the target protein to sweep out or repulse other large molecules so that the target protein can fold without interference. This mechanism of improving solubility is distinct from that of commonly utilized solubility-enhancing fusion proteins that are currently in use. Even with this proposed function, the EBs do not overtly interfere with the stability of GFP or enzymatic activity of GST (Figure 4). If functional maintenance becomes a problem for any particular fusion partner, we have shown that the fusion tags can be removed by incubation with enterokinase (Figure 5).

Using artificial rather than natural sequences as the basis for EB domains provides the researcher with the opportunity to try a variety of sequences that differ in length, net charge, detailed amino acid sequence, etc. An interesting finding herein is that a variety of rather different sequences demonstrated rather similar abilities to increase the solubility of a variety of recalcitrant proteins, suggesting that general disorder properties rather than particular sequences are important for the effects being reported here.

Some regions of disorder contain evolutionarily conserved sequences, exhibit functional conservation, carry out their functions even when isolated from the rest of the protein or even when fused with a different sequence, and have conserved, albeit disordered, structures. Except for the disorder associated with the last characteristic, these features match those of structured protein domains, and even the last one matches if disorder is allowed to be considered a type of "structure". Thus, we previously proposed that such regions should be considered to be "disordered domains".[101] Here we have conducted the first de novo design of disordered domains that carry out a prespecified, biologically useful function, namely solubility enhancement. Feats of protein engineering much more sophisticated than those described here have allowed scientists to manipulate preexisting sequences for the purposes of altering both structure and function of known proteins (examples in refs 102−104), but rather than modifying known templates, our work starts from first principles to develop sets of sequences specifying disordered domains, all of which possess the same preidentified biological function.

Overall, the EB technology described here is a promising new tool that can help us overcome problems associated with protein overexpression using a unique mechanism. EB technology provides a complementary resource for scientists whose research is hindered by poor protein solubility, and we anticipate that many useful modifications of this basic platform will be developed in the coming years.

■ ASSOCIATED CONTENT

S Supporting Information

Computational estimates of the order and disorder status for the various solubility-enhancing fusion proteins and entropic bristles used in this study, a discussion of composition versus sequence as an indicator of structure or disorder, a list of the amino acid sequences of the EBs used here, and the nucleotide sequence of the cloning vector developed for this study. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Molecular Kinetics Inc., 6201 La Pas Trail, Indianapolis, IN 46268. E-mail: akdunker@molecularkinetics.com. Telephone: (317) 278-9220 (Indiana University) or (317) 280-8737 (Molecular Kinetics Inc.).

Present Addresses

[‖]Cook Medical, Bloomington, IN 47402.
[⊥]Department of Molecular Medicine, University of South Florida, Tampa, FL 33612.

Author Contributions

A.A.S. and C.H.C. contributed equally to this work.

Funding

This project has been supported by the National Cancer Institute, National Institutes of Health, under SBIR Grant R44CA110548.

Notes

The authors declare the following competing financial interest(s): Several of the auhors (A.A.S., C.H.C., F.H.V., and Y.-Y.J.V.) are employees of Molecular Kinetics, Inc., and the subject matter of this article may become a future commercial product.

■ ACKNOWLEDGMENTS

We thank Dr. Stephen Randall of the Department of Biology at Indiana University-Purdue University Indianapolis (Indianapolis, IN) for providing the *A. thaliana* dehydrin clones.

■ REFERENCES

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

(2) Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C. H., and Edwards, A. M. (2000) Structural proteomics: Prospects for high throughput sample preparation. *Prog. Biophys. Mol. Biol. 73*, 339−345.

(3) Braun, P., Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E., and LaBaer, J. (2002) Proteome-scale purification of human proteins from bacteria. *Proc. Natl. Acad. Sci. U.S.A. 99*, 2654−2659.

(4) Abrahmsen, L., Moks, T., Nilsson, B., and Uhlen, M. (1986) Secretion of heterologous gene products to the culture medium of *Escherichia coli*. *Nucleic Acids Res. 14*, 7487−7500.

(5) Gottesman, S., and Zipser, D. (1978) Deg phenotype of *Escherichia coli* lon mutants. *J. Bacteriol. 133*, 844−851.

(6) Nilsson, B., Abrahmsen, L., and Uhlen, M. (1985) Immobilization and purification of enzymes with staphylococcal protein A gene fusion vectors. *EMBO J. 4*, 1075−1080.

(7) Chatterjee, D. K., and Esposito, D. (2006) Enhanced soluble protein expression using two new fusion tags. *Protein Expression Purif. 46*, 122−129.

(8) Davis, G. D., Elisee, C., Newham, D. M., and Harrison, R. G. (1999) New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng. 65*, 382−388.

(9) di Guan, C., Li, P., Riggs, P. D., and Inouye, H. (1988) Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene 67*, 21−30.

(10) Dyson, M. R., Shadbolt, S. P., Vincent, K. J., Perera, R. L., and McCafferty, J. (2004) Production of soluble mammalian proteins in *Escherichia coli*: Identification of protein features that correlate with successful expression. *BMC Biotechnol. 4*, 32.

(11) Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F., and Boyer, H. W. (1977) Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin. *Science 198*, 1056−1063.

(12) Johnson, E. S. (2004) Protein modification by SUMO. *Annu. Rev. Biochem. 73*, 355−382.

(13) Kapust, R. B., and Waugh, D. S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci. 8*, 1668−1674.

(14) Kobayashi, H., Yoshida, T., and Inouye, M. (2009) Significant enhanced expression and solubility of human proteins in *Escherichia coli* by fusion with protein S from *Myxococcus xanthus*. *Appl. Environ. Microbiol. 75*, 5356−5362.

(15) LaVallie, E. R., DiBlasio, E. A., Kovacic, S., Grant, K. L., Schendel, P. F., and McCoy, J. M. (1993) A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *Nat. Biotechnol. 11*, 187−193.

(16) Sachdev, D., and Chirgwin, J. M. (2000) Fusions to maltose-binding protein: Control of folding and solubility in protein purification. *Methods Enzymol. 326*, 312−321.

(17) Shen, S. H. (1984) Multiple joined genes prevent product degradation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A. 81*, 4627−4631.

(18) Smith, D. B. (2000) Generating fusions to glutathione S-transferase for protein studies. *Methods Enzymol. 326*, 254−270.

(19) Smith, D. B., and Johnson, K. S. (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene 67*, 31−40.

(20) Sorensen, H. P., Kristensen, J. E., Sperling-Petersen, H. U., and Mortensen, K. K. (2004) Soluble expression of aggregating proteins by covalent coupling to the ribosome. *Biochem. Biophys. Res. Commun. 319*, 715−719.

(21) Vaillancourt, P., Simcox, T. G., and Zheng, C. F. (1997) Recovery of polypeptides cleaved from purified calmodulin-binding peptide fusion proteins. *BioTechniques 22*, 451−453.

(22) Zhan, Y., Song, X., and Zhou, G. W. (2001) Structural analysis of regulatory protein domains using GST-fusion proteins. *Gene 281*, 1−9.

(23) Trabbic-Carlson, K., Meyer, D. E., Liu, L., Piervincenzi, R., Nath, N., LaBean, T., and Chilkoti, A. (2004) Effect of protein fusion on the transition temperature of an environmentally responsive elastin-like polypeptide: A role for surface hydrophobicity? *Protein Eng., Des. Sel. 17*, 57−66.

(24) Trabbic-Carlson, K., Liu, L., Kim, B., and Chilkoti, A. (2004) Expression and purification of recombinant proteins from *Escherichia coli*: Comparison of an elastin-like polypeptide fusion with an oligohistidine fusion. *Protein Sci. 13*, 3274−3284.

(25) Magnan, C. N., Randall, A., and Baldi, P. (2009) SOLpro: Accurate sequence-based prediction of protein solubility. *Bioinformatics 25*, 2200−2207.

(26) Smialowski, P., Martin-Galiano, A. J., Mikolajka, A., Girschick, T., Holak, T. A., and Frishman, D. (2007) Protein solubility: Sequence based prediction and experimental verification. *Bioinformatics 23*, 2536−2542.

(27) Wilkinson, D. L., and Harrison, R. G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Nat. Biotechnol. 9*, 443−448.

(28) Das, P., King, J. A., and Zhou, R. (2011) Aggregation of γ-crystallins associated with human cataracts via domain swapping at the C-terminal β-strands. *Proc. Natl. Acad. Sci. U.S.A. 108*, 10514−10519.

(29) Speed, M. A., Wang, D. I., and King, J. (1996) Specific aggregation of partially folded polypeptide chains: The molecular basis of inclusion body composition. *Nat. Biotechnol. 14*, 1283−1287.

(30) Hoh, J. H. (1998) Functional protein domains from the thermally driven motion of polypeptide chains: A proposal. *Proteins 32*, 223−228.

(31) Mouillon, J. M., Gustafsson, P., and Harryson, P. (2006) Structural investigation of disordered stress proteins. Comparison of full-length dehydrins with isolated peptides of their conserved segments. *Plant Physiol. 141*, 638−650.

(32) Tompa, P., and Kovacs, D. (2010) Intrinsically disordered chaperones in plants and animals. *Biochem. Cell Biol. 88*, 167−174.

(33) Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein. *J. Mol. Graphics Modell. 19*, 26−59.

(34) Vacic, V., Uversky, V. N., Dunker, A. K., and Lonardi, S. (2007) Composition Profiler: A tool for discovery and visualization of amino acid composition differences. *BMC Bioinf. 8*, 211.

(35) Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999) Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inf. Ser. 10*, 30−40.

(36) Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein. *Proteins 42*, 38−48.

(37) Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A. K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins 61* (Suppl. 7), 176−182.

(38) Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinf. 7*, 208.

(39) Xue, B., Dunbrack, R. L., Williams, R. M., Dunker, A. K., and Uversky, V. N. (2010) PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta 1804*, 996−1010.

(40) Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins 41*, 415−427.

(41) Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157*, 105−132.

(42) Kovacs, D., Kalmar, E., Torok, Z., and Tompa, P. (2008) Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. *Plant Physiol. 147*, 381−390.

(43) Narberhaus, F. (2002) α-Crystallin-type heat shock proteins: Socializing minichaperones in the context of a multichaperone network. *Microbiol. Mol. Biol. Rev. 66*, 64−93.

(44) Park, S. M., Jung, H. Y., Kim, T. D., Park, J. H., Yang, C. H., and Kim, J. (2002) Distinct roles of the N-terminal-binding domain and the C-terminal-solubilizing domain of α-synuclein, a molecular chaperone. *J. Biol. Chem. 277*, 28512−28520.

(45) Tompa, P., and Csermely, P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *FASEB J. 18*, 1169−1175.

(46) Weldon, J. E., and Schleif, R. F. (2006) Specific interactions by the N-terminal arm inhibit self-association of the AraC dimerization domain. *Protein Sci. 15*, 2828−2835.

(47) Close, T. J. (1997) Dehydrins: A commonality in the response of plants to dehydration and low temperature. *Physiol. Plant. 100*, 291−296.

(48) Hughes, S., and Graether, S. P. (2011) Cryoprotective mechanism of a small intrinsically disordered dehydrin protein. *Protein Sci. 20*, 42−50.

(49) Nakayama, K., Okawa, K., Kakizaki, T., Honma, T., Itoh, H., and Inaba, T. (2007) *Arabidopsis* Cor15am is a chloroplast stromal protein that has cryoprotective activity and forms oligomers. *Plant Physiol. 144*, 513−523.

(50) Puhakainen, T., Hess, M. W., Makela, P., Svensson, J., Heino, P., and Palva, E. T. (2004) Overexpression of multiple dehydrin genes enhances tolerance to freezing stress in *Arabidopsis*. *Plant Mol. Biol. 54*, 743−753.

(51) Reyes, J. L., Campos, F., Wei, H., Arora, R., Yang, Y., Karlson, D. T., and Covarrubias, A. A. (2008) Functional dissection of hydrophilins during in vitro freeze protection. *Plant, Cell Environ. 31*, 1781−1790.

(52) Rinne, P. L., Kaikuranta, P. L., van der Plas, L. H., and van der Schoot, C. (1999) Dehydrins in cold-acclimated apices of birch (*Betula pubescens* ehrh.): Production, localization and potential role in rescuing enzyme function during dehydration. *Planta 209*, 377−388.

(53) Rorat, T. (2006) Plant dehydrins: Tissue location, structure and function. *Cell. Mol. Biol. Lett. 11*, 536−556.

(54) Wisenieski, M., Webb, R., Balsamo, R., Close, T. J., Yu, X. M., and Griffith, M. (1999) Purification, immunolocalization, cryoprotective, and antifreeze activity of PCA60: A dehydrin from peach (*Prunus persica*). *Physiol. Plant. 105*, 600−608.

(55) Singh, J., Whitwill, S., Lacroix, G., Douglas, J., Dubuc, E., Allard, G., Keller, W., and Schernthaner, J. P. (2009) The use of Group 3 LEA proteins as fusion partners in facilitating recombinant expression of recalcitrant proteins in *E. coli*. *Protein Expression Purif. 67*, 15−22.

(56) Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., and Dunker, A. K. (2007) DisProt: The Database of Disordered Proteins. *Nucleic Acids Res. 35*, D786−D793.

(57) Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., Cortese, M. S., Lawson, J. D., Brown, C. J., Sikes, J. G., Newton, C. D., and Dunker, A. K. (2005) DisProt: A database of protein disorder. *Bioinformatics 21*, 137−140.

(58) Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., and Dunker, A. K. (2007) Intrinsic disorder and functional proteomics. *Biophys. J. 92*, 1439−1456.

(59) Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J., Takayama, S., Brown, C. J., and Dunker, A. K. (2001) The protein non-folding problem: Amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput. 2001,*, 89−100.

(60) Ueda, E. K., Gout, P. W., and Morganti, L. (2003) Current and prospective applications of metal ion-protein binding. *J. Chromatogr., A 988*, 1−23.

(61) Hara, M., Fujinaga, M., and Kuboi, T. (2005) Metal binding by citrus dehydrin with histidine-rich domains. *J. Exp. Bot. 56*, 2695−2703.

(62) Asano, R., Kudo, T., Makabe, K., Tsumoto, K., and Kumagai, I. (2002) Antitumor activity of interleukin-21 prepared by novel refolding procedure from inclusion bodies expressed in *Escherichia coli*. *FEBS Lett. 528*, 70−76.

(63) Kang, W. K., Park, E. K., Lee, H. S., Park, B. Y., Chang, J. Y., Kim, M. Y., Kang, H. A., and Kim, J. Y. (2007) A biologically active angiogenesis inhibitor, human serum albumin-TIMP-2 fusion protein, secreted from *Saccharomyces cerevisiae*. *Protein Expression Purif. 53*, 331−338.

(64) Ouellette, T., Destrau, S., Zhu, J., Roach, J. M., Coffman, J. D., Hecht, T., Lynch, J. E., and Giardina, S. L. (2003) Production and purification of refolded recombinant human IL-7 from inclusion bodies. *Protein Expression Purif. 30*, 156−166.

(65) Campos, F., Zamudio, F., and Covarrubias, A. A. (2006) Two different late embryogenesis abundant proteins from *Arabidopsis thaliana* contain specific domains that inhibit *Escherichia coli* growth. *Biochem. Biophys. Res. Commun. 342*, 406−413.

(66) Jones, L. S., Yazzie, B., and Middaugh, C. R. (2004) Polyanions and the proteome. *Mol. Cell. Proteomics 3*, 746−769.

(67) Volkin, D. B., Tsai, P. K., Dabora, J. M., Gress, J. O., Burke, C. J., Linhardt, R. J., and Middaugh, C. R. (1993) Physical stabilization of acidic fibroblast growth factor by polyanions. *Arch. Biochem. Biophys. 300*, 30−41.

(68) Miller, R. T., Douthart, R. J., and Dunker, A. K. (1993) Learning an objective alphabet of amino acid conformations in protein. *Tech. Protein Chem. 4*, 541−548.

(69) Cao, P., Mei, J. J., Diao, Z. Y., and Zhang, S. (2005) Expression, refolding, and characterization of human soluble BAFF synthesized in *Escherichia coli*. *Protein Expression Purif. 41*, 199−206.

(70) Jin, H. J., Dunn, M. A., Borthakur, D., and Kim, Y. S. (2004) Refolding and purification of unprocessed porcine myostatin expressed in *Escherichia coli*. *Protein Expression Purif. 35*, 1−10.

(71) Zegzouti, H., Li, W., Lorenz, T. C., Xie, M., Payne, C. T., Smith, K., Glenny, S., Payne, G. S., and Christensen, S. K. (2006) Structural and functional insights into the regulation of *Arabidopsis* AGC VIIIa kinases. *J. Biol. Chem. 281*, 35520−35530.

(72) Yang, F., Moss, L. G., and Phillips, G. N., Jr. (1996) The molecular structure of green fluorescent protein. *Nat. Biotechnol. 14*, 1246−1251.

(73) Ormo, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y., and Remington, S. J. (1996) Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science 273*, 1392−1395.

(74) Stepanenko, O. V., Verkhusha, V. V., Kazakov, V. I., Shavlovsky, M. M., Kuznetsova, I. M., Uversky, V. N., and Turoverov, K. K. (2004) Comparative studies on the structure and stability of fluorescent proteins EGFP, zFP506, mRFP1, "dimer2", and DsRed1. *Biochemistry 43*, 14913−14923.

(75) Fukuda, H., Arai, M., and Kuwajima, K. (2000) Folding of green fluorescent protein and the cycle3 mutant. *Biochemistry 39*, 12025−12032.

(76) Reid, B. G., and Flynn, G. C. (1997) Chromophore formation in green fluorescent protein. *Biochemistry 36*, 6786−6791.

(77) Habig, W. H., Pabst, M. J., and Jakoby, W. B. (1974) Glutathione S-transferases. The first enzymatic step in mercapturic acid formation. *J. Biol. Chem. 249*, 7130−7139.

(78) Mannervic, B., and Danielson, U. H. (1988) Glutathione S-transferases: Structure and catalytic activity. *Crit. Rev. Biochem. 23*, 283−337.

(79) Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A. M., Arrowsmith, C. H., Montelione, G. T., and Gerstein, M. (2001) SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res. 29*, 2884−2898.

(80) Goh, C. S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G. T., Zhao, H., and Gerstein, M. (2004) Mining the structural genomics pipeline: Identification of protein

properties that affect high-throughput experimental analysis. *J. Mol. Biol. 336*, 115−130.

(81) Idicula-Thomas, S., and Balaji, P. V. (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci. 14*, 582−592.

(82) 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., and Luo, M. (2004) High-throughput expression of *C. elegans* proteins. *Genome Res. 14*, 2102−2110.

(83) Sharp, K. A., Nicholls, A., Friedman, R., and Honig, B. (1991) Extracting hydrophobic free energies from experimental data: Relationship to protein folding and theoretical models. *Biochemistry 30*, 9686−9697.

(84) Kochendoerfer, G. (2003) Chemical and biological properties of polymer-modified proteins. *Expert Opin. Biol. Ther. 3*, 1253−1261.

(85) Bokor, M., Csizmok, V., Kovacs, D., Banki, P., Friedrich, P., Tompa, P., and Tompa, K. (2005) NMR relaxation studies on the hydrate layer of intrinsically unstructured proteins. *Biophys. J. 88*, 2030−2037.

(86) Tompa, P., Banki, P., Bokor, M., Kamasa, P., Kovacs, D., Lasanda, G., and Tompa, K. (2006) Protein-water and protein-buffer interactions in the aqueous solution of an intrinsically unstructured plant dehydrin: NMR intensity and DSC aspects. *Biophys. J. 91*, 2243−2249.

(87) Soulages, J. L., Kim, K., Arrese, E. L., Walters, C., and Cushman, J. C. (2003) Conformation of a group 2 late embryogenesis abundant protein from soybean. Evidence of poly(L-proline)-type II structure. *Plant Physiol. 131*, 963−975.

(88) Naper, D. H. (1983) Stabilization by attached polymer: Steric stabilization. In *Polymeric stabilization of colloidal dispersions* (Napper, D. H., Ed.) pp 18−30, Academic Press, London.

(89) Milner, S. T. (1991) Polymer Brushes. *Science 251*, 905−914.

(90) Bach, H., Mazor, Y., Shaky, S., Shoham-Lev, A., Berdichevsky, Y., Gutnick, D. L., and Benhar, I. (2001) *Escherichia coli* maltose-binding protein as a molecular chaperone for recombinant intracellular cytoplasmic single-chain antibodies. *J. Mol. Biol. 312*, 79−93.

(91) Receveur-Brechot, V., Bourhis, J. M., Uversky, V. N., Canard, B., and Longhi, S. (2006) Assessing protein disorder and induced folding. *Proteins 62*, 24−45.

(92) Richarme, G., and Caldas, T. D. (1997) Chaperone properties of the bacterial periplasmic substrate-binding proteins. *J. Biol. Chem. 272*, 15607−15612.

(93) Douette, P., Navet, R., Gerkens, P., Galleni, M., Levy, D., and Sluse, F. E. (2005) *Escherichia coli* fusion carrier proteins act as solubilizing agents for recombinant uncoupling protein 1 through interactions with GroEL. *Biochem. Biophys. Res. Commun. 333*, 686−693.

(94) Alsheikh, M. K., Heyen, B. J., and Randall, S. K. (2003) Ion binding properties of the dehydrin ERD14 are dependent upon phosphorylation. *J. Biol. Chem. 278*, 40882−40889.

(95) Chakrabortee, S., Boschetti, C., Walton, L. J., Sarkar, S., Rubinsztein, D. C., and Tunnacliffe, A. (2007) Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function. *Proc. Natl. Acad. Sci. U.S.A. 104*, 18073−18078.

(96) Hara, M., Terashima, S., Fukaya, T., and Kuboi, T. (2003) Enhancement of cold tolerance and inhibition of lipid peroxidation by citrus dehydrin in transgenic tobacco. *Planta 217*, 290−298.

(97) Zhang, Y. B., Howitt, J., McCorkle, S., Lawrence, P., Springer, K., and Freimuth, P. (2004) Protein aggregation during overexpression limited by peptide extensions with large net negative charge. *Protein Expression Purif. 36*, 207−216.

(98) Cohn, E. J., Gross, J., and Johnson, O. C. (1919) The Isoelectric Points of the Proteins in Certain Vegetable Juices. *J. Gen. Physiol. 2*, 145−160.

(99) Loeb, J. (1918) Amphoteric Colloids: II. Volumetric Analysis of Ion-Protein Compounds; the Significance of the Isoelectric Point for the Purification of Amphoteric Colloids. *J. Gen. Physiol. 1*, 237−254.

(100) Shih, Y. C., Prausnitz, J. M., and Blanch, H. W. (1992) Some characteristics of protein precipitation by salts. *Biotechnol. Bioeng. 40*, 1155−1164.

(101) Tompa, P., Fuxreiter, M., Oldfield, C. J., Simon, I., Dunker, A. K., and Uversky, V. N. (2009) Close encounters of the third kind: Disordered domains and the interactions of proteins. *BioEssays 31*, 328−335.

(102) Firestine, S. M., Salinas, F., Nixon, A. E., Baker, S. J., and Benkovic, S. J. (2000) Using an AraC-based three-hybrid system to detect biocatalysts in vivo. *Nat. Biotechnol. 18*, 544−547.

(103) Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., Wilson, I. A., and Baker, D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science 332*, 816−821.

(104) Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S., and Baker, D. (2011) De novo enzyme design using Rosetta3. *PLoS One 6*, e19230.