Published in final edited form as:

Biochemistry. 2012 September 18; 51(37): 7250–7262. doi:10.1021/bi300653m.

Sweeping away protein aggregation with entropic bristles: Intrinsically disordered protein fusions enhance soluble expression

Aaron A. Santner^{1,†,‡}, Carrie H. Croy^{1,‡}, Farha H. Vasanwala¹, Vladimir N. Uversky^{1,2,3,††},
Ya-Yue J. Van¹, and A. Keith Dunker^{1,2,*}

¹Molecular Kinetics Inc., Indianapolis, IN 46268, USA

²Department of Biochemistry and Molecular Biology, Indiana University, Indianapolis, IN 46202, USA

³Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

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 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 enable 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 low sequence complexity and 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 physio-chemical properties they did not perturb the structure, conformational stability nor 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.

*Corresponding Author, A.K. Dunker, Molecular Kinetics, 6201 La Pas Trail, Indianapolis, Indiana 46268, akdunker@molecularkinetics.com. Telephone numbers: Indiana University: 317 278 9220; Molecular Kinetics 317 280 8737.

‡These authors contributed equally

†Present Addresses

Cook Medical, Bloomington, IN 47402.

††Department of Molecular Medicine, University of South Florida, Tampa, FL 33612.

SUPPORTING INFORMATION

The supporting information provides computational estimates of the order / disorder statuses for the various solubility-enhancing fusion proteins and entropic bristles used in this study. In addition, the supporting information discusses composition versus sequence as an indicator of structure or disorder, and provides a list of the amino acid sequences of the EBs used herein and also provides the nucleotide sequence of the cloning vector developed for this study. All supporting materials for this publication may be accessed free of charge online at <http://pubs.acs.org>.

Keywords

intrinsic disorder; protein; solubility; aggregation; translational fusion

INTRODUCTION

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 non-membrane 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 more than 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/proteins produced in *Escherichia 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 *Escherichia 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 due to 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 reduced if the fusion tag stimulates chaperone recruitment or if the fusion protein itself acts 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 due to their shifts in the overall amino acid composition towards

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), or “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 strong resemblance to local sequences in Hsp90 (31).

Here we test our hypothesis that IDPs can lead to solubility-enhancement when 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 out-perform several structured solubility-enhancer sequences. Finally, we demonstrate the maintenance of biological function and stability for a few of the EB-target hybrids. Based on 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 were carried out 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 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 datasets containing both structured and disordered proteins, while PONDR® VSL2 gave accuracies of 74% and 78% on the same two datasets (39). Additionally, charge-hydrophobicity distributions (CH-plots) were also analyzed for these proteins using methods as described in Uversky et al. (40). The CH-plot in Figure 1C is a 2D graph plotting the Kyte-Doolittle hydrophobicity 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 fraction. 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 PROSO and SolPro were also run on the various fusion sequences. (25, 26). PROSO (PROtein SOLubility predictor) is a machine-learning approach trained on a 14,200 protein dataset and was originally reported with a prediction accuracy of 72% (26). The SOLpro predictor used a two-tiered SVM strategy trained on 17,408 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 were of 54%, 59% and 74%, respectively (25).

Design of 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). This vector utilizes an N-terminal hexa-histidine 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 sequence sequences provided in Figure S2). 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 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 an LB medium supplemented with 100µg/mL ampicillin at 37°C. The next morning a 150µL aliquot of culture was spun down and re-suspended in LB media containing 0.5M sorbitol and 1mM betaine for the purpose of inducing expression of endogenous *E. coli* chaperone proteins. The fresh cultures were incubated at 37°C until cultures reached an OD600 of 0.4, and then expression was induced by adding 0.2mM IPTG. Induction of protein expression was carried out for six hours at the reduced temperature of 25°C. After this induction period, cells were pelleted by centrifugation, and frozen at -20°C until expression analysis. For soluble protein expression analysis cell pellets

were permeabilized, following the manufacturer's suggested conditions, under isotonic conditions using a solution containing both 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, 5min) capable of pelleting large cellular debris and sub-cellular 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 SDS-PAGE using the NuPAGE Bis-Tris gradient gel system (Invitrogen). The proteins were transferred to PVDF membranes (Invitrogen) and probed with anti-His antibody (Santa Cruz Biotechnologies, 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 the ImageJ software (NIH).

GST Purification and Activity Assay

Following the cell growth and lysis procedure 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, 1mL of soluble lysate containing the GST-fusion protein was incubated with 0.25mL of glutathione resin for 1h at 4°C with mixing. Resin was washed with six volumes of Dulbecco's phosphate buffer (D-PBS), and then eluted with three volumes of D-PBS containing 50mM glutathione. The eluate concentrations were determined by Bradford assay (Coomassie protein assay kit, Thermo Scientific).

GST transferase activity was determined by measuring the coupling of reduced glutathione to a 1-chloro-2, 4-dinitrobenzene (CDNB, Sigma) substrate by observing increasing absorbance at 340nm. Specifically, the reaction was initiated when ~20pmoles of the enriched GST-fusion protein was added to a 1mL quartz cuvette containing 2mM Glutathione, 1mM CDNB in Dulbecco's PBS. Using a kinetics program the change in the 340nm reading was measured every 30s for 5mins on a Varian Cary Eclipse Fluorescence spectrophotometer. The GST specific activity was then calculated as the μ moles of CDNB converted per minute per pmol 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 above the various GFP fusions were partially purified using TALON® Superflow™ Metal Affinity Resin (Clontech). Specifically, 1mL of soluble lysate containing the GFP-fusion protein was incubated with 0.1mL of TALON® resin for 20min at 25°C with rotation. The resin was recovered in a column and washed with ten volumes of wash buffer (50mM Tris pH8.0, 150mM NaCl, 5mM Imidazole), and then

eluted with three volumes of buffer containing 150mM Imidazole. The eluate concentrations were determined by 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 to 72hrs. GFP unfolding was monitored using fluorescence spectrophotometry (Cary Eclipse, Varian), the excitation wavelength was 395nm and emission was detected at 510nm.

Enterokinase Cleavage

To demonstrate that enterokinase can cleave the DDDDKS consensus sequence located between the EB and target sequences, 5µg of purified EB-GFP fusion protein was incubated with 1.5U of recombinant enterokinase protease (Novagen) over 24hours. The efficiency of the digestion was monitored at the discrete time points of 0, 2, 4, 8, and 24hours 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 aim we first analyzed the primary sequences 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 anti-aggregant 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 *A. thaliana* dehydrin proteins. Focusing on the disorder predictors values presented in Table 1 first, PONDR® predictors VLXT and VSL2 verify high “percent disorder” values for *A. thaliana* 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 ninety residue stretch in the central part of the protein (residues 90–132 and 138–179). Further examination of dehydrins by measuring the mean hydropathy indicate that all 6 dehydrins lie on the disordered side of the boundary with 5/6 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 in bulky hydrophobic and aromatic residues which 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 have both a deficiency in several order-promoting residues (W, C, and F), and enrichment of 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 of functional importance as members of the family have been shown to bind several divalent metals using a conserved His containing sequence, e.g. the HKGEHHS GDHH core sequence for Cu²⁺ binding by citrus dehydrin CuCOR15 (53, 60, 61).

Figure 1A also demonstrates that dehydrins have favorable soluble expression characteristics by multiple predictor algorithms. The early Wilkinson and Harrison solubility model found only the COR47, ERD10 and ERD14 dehydrin sequences favor 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) mean net negative. Though 5/6 dehydrins have favorable enrichments in “turn-forming” residues for (Gly for LTI30, Rab18, and Xero1; and Pro for ERD10, 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, PROSO and SOLpro (25, 26), both of which use a machine learning approach, broadly show more favorable solubility scores for the dehydrin proteins than the W-H 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 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 yield higher levels of soluble protein than the disordered LTI30 fusion. After minimal success with the neutral Rab18 and positive LTI30 dehydrins, we proceeded with evaluating the ability of the negatively charged dehydrins to aid soluble protein expression (data not shown). Overall ERD10 and ERD14 showed a 74% and 54% rate of success 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 ten proteins where multiple experimental sets were collected to enable statistical evaluations. In six of the ten cases, ERD10 significantly (p-value <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 wide-ranging design of the potentially solubilizing sequences.

Design of artificial EB fusion polypeptides

Designing artificial EB sequences which 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 attempted in recombinant expression systems (65). All polypeptides were designed to be 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 verifies their localization within the trapezium of extended IDP proteins. Finally, pilot expression studies verified that our artificial EBs displayed no obvious cellular toxicity.

To explain our design rationale, we will briefly describe how the *de novo* EB templates shown in Table 2 were created. The residues at the far-right side of Figure 1A represent the most disorder-promoting residues (E, P, Q, and S) and were chosen as constituents of EB60A, in the proportion 2E:2P:1Q:1S. The rationale for the proportions was the following: 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 because it was not only hydrophilic, but it exhibits one of the largest conformational variabilities of the twenty amino acids (68). Based on such considerations, 360-nucleotide long sequences were randomly generated with codon optimization for expression in *E. coli*. This synthetic gene encodes the 120-residue-long polypeptide which serves as the basis for our *de novo* EB sequence fusion. Since serines are common sites for posttranslational modification, we also designed an EB60B series of *de novo* EBs, in which serines were replaced by the disorder-neutral residue G (EB60B 2E:2P:1Q:1G). Additionally, as we generated EBs with higher negative charge densities and increased lengths, in order to avoid potential issues with expression problems associated with high sequence redundancy, we added a larger subset of disordered-residues, for example, EB144 uses 1D:2E:2P:1Q:2S:1G. Finally, EB250 was designed with the 1D:2E:2P:1Q:2S:1G 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 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 EBs fusions used to present the soluble expression efficiency and activity studies below (Figure S2 gives the exact polypeptide sequence for each EB).

The same analyses carried out on the six dehydrins (Table 1, above) 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 fusion to enhance soluble protein expression

The extent of soluble expression enhancement provided by the four compositionally unique EB fusion sequences (Figure S2) was determined using the widely used *Escherichia coli* recombinant system. The 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 aggregate 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 3 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 or EB144 and EB250, in 75%, 95% and 100% (p-value = 0.05) of the test candidates, respectively. Interestingly, we found that length was a more important determinant than composition. Specifically, the two EBs of sixty amino acids in length showed similar performance, but were less successful than the longer 144 and 250 amino acid fusion. Follow-up studies in which EBs of identical composition but varying sequence order and length will be carried out to further compare the affects of composition, primary sequence, and length on solubility. Although Table 4 supports that variation in soluble expression levels exist for a given target, the wide-spread success of EB144 and EB250 support the notion that a universal IDP-based solubilizer could be a reasonable goal. However, in consideration of 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 physio-chemical 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 in comparison to the MBP-hybrid. Next, we selected four of our translatable gene targets (2141, CAT 9, FLCN, 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 other fusion systems.

The EB-fusions have very distinct physio-chemical properties compared to the commonly used solubility-enhancing structured protein. Thus, the question arises 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. 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 fluorescent protein with similar spectral properties as the original non-tagged 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 the 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 the increasing GndHCl concentrations for periods of time up to several days at room temperature. Consistent with previous findings, 6xHis-GFP fluorescence was found to increase slightly when in the presence of low GndHCl concentrations (<2M) 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 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.

Transferase activity of recombinant Glutathione-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 reported were normalized to the GST-active standard purchased from Biovision (Cat #1243-1), and reported on a per mole basis due to the size variation among different EB fusions. Lysates purified from different clones and grown on different days showed similar results 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 the 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. Since 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 observed 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, or EBs, to aggregation-prone protein leads to almost universal improvement of protein solubility, with longer and more negatively charged EBs showing slight enhancements compared to shorter and less charged EBs. More work is needed to understand secondary affects 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 suggest that, compared to water-soluble, structured proteins, EBs have particular properties that significantly improve the solubility of the fused construct.

Computer algorithms identify not only 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 residues 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 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 given 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 to a first approximation 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; and 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 for their expected relative contribution to each of these three factors.

Compared to a structured protein of 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 due to the presence of polyproline II-type helices, which have higher solvent-accessible surface areas compared to other types of secondary structural elements (31, 87).

Compared to a structured protein of the same number of amino acids, an IDP would provide a much larger excluded volume. In fact, an unstructured polymer that enhances solubility has been given the special name of entropic bristle (88) or EB. Indeed, EBs based on a variety of polymers have been used to reduce aggregation of particles such as latex particles in paints and to stabilize a wide variety of other colloidal products (30). 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 growing evidence that the disordered regions within chaperones are responsible for their support of protein folding (reviewed in (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 the structured protein fusion tags.

The work presented here revealed that dehydrin protein family members, ERD10 and ERD14, are effective fusion partners for improving the solubility of aggregation-prone proteins expressed in *E. coli* (Fig 2). Since 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 assessing these novel polypeptides, it became clear that the EB polypeptides with net positive charges were not only ineffectual as solubility-enhancers but were in some cases detrimental to the overall solubility of the fusion proteins (data not shown). Perhaps their association with negatively phospholipid head groups 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 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 confirmation 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 free from interference. This mechanism of improving solubility is distinct from 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 (Fig. 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 (Fig. 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, and so on. 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 affects 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 carried out the first *de novo* design of disordered domains that carry out a pre-specified, biologically useful function, namely solubility enhancement. Much more sophisticated feats of protein engineering than described herein have enabled scientists to manipulate pre-existing sequences for the purposes altering both structure and function of known proteins (examples, (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 pre-identified biological function.

Overall, the EB technology described here is a promising new tool that can help to overcome problems associated with protein over-expression 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Dr. Stephen Randall of the Department of Biology at IUPUI (Indianapolis, IN) for providing the *A. thaliana* dehydrin clones.

Funding Sources

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

REFERENCES

1. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA,

- Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH. Structural proteomics of an archaeon. *Nat Struct Biol.* 2000; 7:903–909. [PubMed: 11017201]
2. Christendat D, Yee A, Dharamsi A, Kluger Y, Gerstein M, Arrowsmith CH, Edwards AM. Structural proteomics: prospects for high throughput sample preparation. *Prog Biophys Mol Biol.* 2000; 73:339–345. [PubMed: 11063779]
3. Braun P, Hu Y, Shen B, Halleck A, Koundinya M, Harlow E, LaBaer J. Proteome-scale purification of human proteins from bacteria. *Proc Natl Acad Sci U S A.* 2002; 99:2654–2659. [PubMed: 11880620]
4. Abrahmsen L, Moks T, Nilsson B, Uhlen M. Secretion of heterologous gene products to the culture medium of *Escherichia coli*. *Nucleic Acids Res.* 1986; 14:7487–7500. [PubMed: 3020516]
5. Gottesman S, Zipser D. Deg phenotype of *Escherichia coli* lon mutants. *J Bacteriol.* 1978; 133:844–851. [PubMed: 146704]
6. Nilsson B, Abrahmsen L, Uhlen M. Immobilization and purification of enzymes with staphylococcal protein A gene fusion vectors. *EMBO J.* 1985; 4:1075–1080. [PubMed: 2990908]
7. Chatterjee DK, Esposito D. Enhanced soluble protein expression using two new fusion tags. *Protein Expr Purif.* 2006; 46:122–129. [PubMed: 16146696]
8. Davis GD, Elisee C, Newham DM, Harrison RG. New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol Bioeng.* 1999; 65:382–388. [PubMed: 10506413]
9. di Guan C, Li P, Riggs PD, Inouye H. Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene.* 1988; 67:21–30. [PubMed: 2843437]
10. Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J. Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. *BMC Biotechnol.* 2004; 4:32. [PubMed: 15598350]
11. Itakura K, Hirose T, Crea R, Riggs AD, Heyneker HL, Bolivar F, Boyer HW. Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin. *Science.* 1977; 198:1056–1063. [PubMed: 412251]
12. Johnson ES. Protein modification by SUMO. *Annu Rev Biochem.* 2004; 73:355–382. [PubMed: 15189146]
13. Kapust RB, Waugh DS. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 1999; 8:1668–1674. [PubMed: 10452611]
14. Kobayashi H, Yoshida T, Inouye M. Significant enhanced expression and solubility of human proteins in *Escherichia coli* by fusion with protein S from *Myxococcus xanthus*. *Appl Environ Microbiol.* 2009; 75:5356–5362. [PubMed: 19542330]
15. LaVallie ER, DiBlasio EA, Kovacic S, Grant KL, Schendel PF, McCoy JM. A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *Biotechnology (N Y).* 1993; 11:187–193. [PubMed: 7763371]
16. Sachdev D, Chirgwin JM. Fusions to maltose-binding protein: control of folding and solubility in protein purification. *Methods Enzymol.* 2000; 326:312–321. [PubMed: 11036650]
17. Shen SH. Multiple joined genes prevent product degradation in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 1984; 81:4627–4631. [PubMed: 6379648]
18. Smith DB. Generating fusions to glutathione S-transferase for protein studies. *Methods Enzymol.* 2000; 326:254–270. [PubMed: 11036647]
19. Smith DB, Johnson KS. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene.* 1988; 67:31–40. [PubMed: 3047011]
20. Sorensen HP, Kristensen JE, Sperling-Petersen HU, Mortensen KK. Soluble expression of aggregating proteins by covalent coupling to the ribosome. *Biochem Biophys Res Commun.* 2004; 319:715–719. [PubMed: 15184041]
21. Vaillancourt P, Simcox TG, Zheng CF. Recovery of polypeptides cleaved from purified calmodulin-binding peptide fusion proteins. *Biotechniques.* 1997; 22:451–453. [PubMed: 9067020]
22. Zhan Y, Song X, Zhou GW. Structural analysis of regulatory protein domains using GST-fusion proteins. *Gene.* 2001; 281:1–9. [PubMed: 11750122]

23. Trabbic-Carlson K, Meyer DE, Liu L, Piervincenzi R, Nath N, LaBean T, Chilkoti A. Effect of protein fusion on the transition temperature of an environmentally responsive elastin-like polypeptide: a role for surface hydrophobicity? *Protein Eng Des Sel*. 2004; 17:57–66. [PubMed: 14985538]
24. Trabbic-Carlson K, Liu L, Kim B, Chilkoti A. Expression and purification of recombinant proteins from *Escherichia coli*: Comparison of an elastin-like polypeptide fusion with an oligohistidine fusion. *Protein Sci*. 2004; 13:3274–3284. [PubMed: 15557268]
25. Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*. 2009; 25:2200–2207. [PubMed: 19549632]
26. Smialowski P, Martin-Galiano AJ, Mikolajka A, Girschick T, Holak TA, Frishman D. Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*. 2007; 23:2536–2542. [PubMed: 17150993]
27. Wilkinson DL, Harrison RG. Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology (N Y)*. 1991; 9:443–448. [PubMed: 1367308]
28. Das P, King JA, Zhou R. Aggregation of gamma-crystallins associated with human cataracts via domain swapping at the C-terminal beta-strands. *Proc Natl Acad Sci U S A*. 2011; 108:10514–10519. [PubMed: 21670251]
29. Speed MA, Wang DI, King J. Specific aggregation of partially folded polypeptide chains: the molecular basis of inclusion body composition. *Nat Biotechnol*. 1996; 14:1283–1287. [PubMed: 9631094]
30. Hoh JH. Functional protein domains from the thermally driven motion of polypeptide chains: a proposal. *Proteins*. 1998; 32:223–228. [PubMed: 9714161]
31. Mouillon JM, Gustafsson P, Harryson P. Structural investigation of disordered stress proteins. Comparison of full-length dehydrins with isolated peptides of their conserved segments. *Plant Physiol*. 2006; 141:638–650. [PubMed: 16565295]
32. Tompa P, Kovacs D. Intrinsically disordered chaperones in plants and animals. *Biochem Cell Biol*. 2010; 88:167–174. [PubMed: 20453919]
33. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. *J Mol Graph Model*. 2001; 19:26–59. [PubMed: 11381529]
34. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*. 2007; 8:211. [PubMed: 17578581]
35. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform*. 1999; 10:30–40.
36. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins*. 2001; 42:38–48. [PubMed: 11093259]
37. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*. 2005; 7(61 Suppl):176–182. [PubMed: 16187360]
38. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 2006; 7:208. [PubMed: 16618368]
39. Xue B, Dunbrack RL, Williams RM, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta*. 2010; 1804:996–1010. [PubMed: 20100603]
40. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*. 2000; 41:415–427. [PubMed: 11025552]
41. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982; 157:105–132. [PubMed: 7108955]
42. Kovacs D, Kalmar E, Torok Z, Tompa P. Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. *Plant Physiol*. 2008; 147:381–390. [PubMed: 18359842]

43. Narberhaus F. Alpha-crystallin-type heat shock proteins: socializing minichaperones in the context of a multichaperone network. *Microbiol Mol Biol Rev.* 2002; 66:64–93. table of contents. [PubMed: 11875128]
44. Park SM, Jung HY, Kim TD, Park JH, Yang CH, Kim J. Distinct roles of the N-terminal-binding domain and the C-terminal-solubilizing domain of alpha-synuclein, a molecular chaperone. *J Biol Chem.* 2002; 277:28512–28520. [PubMed: 12032141]
45. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. *Faseb J.* 2004; 18:1169–1175. [PubMed: 15284216]
46. Weldon JE, Schleif RF. Specific interactions by the N-terminal arm inhibit self-association of the AraC dimerization domain. *Protein Sci.* 2006; 15:2828–2835. [PubMed: 17132863]
47. Close TJ. Dehydrins: a commonality in the response of plants to dehydration and low temperature. *Physiologia Plantarum.* 1997; 100:291–296.
48. Hughes S, Graether SP. Cryoprotective mechanism of a small intrinsically disordered dehydrin protein. *Protein Sci.* 2011; 20:42–50. [PubMed: 21031484]
49. Nakayama K, Okawa K, Kakizaki T, Honma T, Itoh H, Inaba T. Arabidopsis Cor15am is a chloroplast stromal protein that has cryoprotective activity and forms oligomers. *Plant Physiol.* 2007; 144:513–523. [PubMed: 17384167]
50. Puhakainen T, Hess MW, Makela P, Svensson J, Heino P, Palva ET. Overexpression of multiple dehydrin genes enhances tolerance to freezing stress in Arabidopsis. *Plant Mol Biol.* 2004; 54:743–753. [PubMed: 15356392]
51. Reyes JL, Campos F, Wei H, Arora R, Yang Y, Karlson DT, Covarrubias AA. Functional dissection of hydrophilins during in vitro freeze protection. *Plant Cell Environ.* 2008; 31:1781–1790. [PubMed: 18761701]
52. Rinne PL, Kaikuranta PL, van der Plas LH, van der Schoot C. Dehydrins in cold-acclimated apices of birch (*Betula pubescens* ehrh.): production, localization and potential role in rescuing enzyme function during dehydration. *Planta.* 1999; 209:377–388. [PubMed: 10550618]
53. Rorat T. Plant dehydrins--tissue location, structure and function. *Cell Mol Biol Lett.* 2006; 11:536–556. [PubMed: 16983453]
54. Wisniewski M, Webb R, Balsamo R, Close TJ, Yu XM, Griffith M. Purification, immunolocalization, cryoprotective, and antifreeze activity of PCA60: A dehydrin from peach (*Prunus persica*). *Pyhsiologia Plantarum.* 1999; 105:600–608.
55. Singh J, Whitwill S, Lacroix G, Douglas J, Dubuc E, Allard G, Keller W, Schernthaner JP. The use of Group 3 LEA proteins as fusion partners in facilitating recombinant expression of recalcitrant proteins in *E. coli*. *Protein Expr Purif.* 2009; 67:15–22. [PubMed: 19364534]
56. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* 2007; 35:D786–D793. [PubMed: 17145717]
57. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK. DisProt: a database of protein disorder. *Bioinformatics.* 2005; 21:137–140. [PubMed: 15310560]
58. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. *Biophys J.* 2007; 92:1439–1456. [PubMed: 17158572]
59. Williams RM, Obradovic Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput.* 2001:89–100. [PubMed: 11262981]
60. Ueda EK, Gout PW, Morganti L. Current and prospective applications of metal ion-protein binding. *J Chromatogr A.* 2003; 988:1–23. [PubMed: 12647817]
61. Hara M, Fujinaga M, Kuboi T. Metal binding by citrus dehydrin with histidine-rich domains. *J Exp Bot.* 2005; 56:2695–2703. [PubMed: 16131509]
62. Asano R, Kudo T, Makabe K, Tsumoto K, Kumagai I. Antitumor activity of interleukin-21 prepared by novel refolding procedure from inclusion bodies expressed in *Escherichia coli*. *FEBS Lett.* 2002; 528:70–76. [PubMed: 12297282]

63. Kang WK, Park EK, Lee HS, Park BY, Chang JY, Kim MY, Kang HA, Kim JY. A biologically active angiogenesis inhibitor, human serum albumin-TIMP-2 fusion protein, secreted from *Saccharomyces cerevisiae*. *Protein Expr Purif*. 2007; 53:331–338. [PubMed: 17368046]
64. Ouellette T, Destrau S, Zhu J, Roach JM, Coffman JD, Hecht T, Lynch JE, Giardina SL. Production and purification of refolded recombinant human IL-7 from inclusion bodies. *Protein Expr Purif*. 2003; 30:156–166. [PubMed: 12880763]
65. Campos F, Zamudio F, Covarrubias AA. Two different late embryogenesis abundant proteins from *Arabidopsis thaliana* contain specific domains that inhibit *Escherichia coli* growth. *Biochem Biophys Res Commun*. 2006; 342:406–413. [PubMed: 16487487]
66. Jones LS, Yazzie B, Middaugh CR. Polyanions and the proteome. *Mol Cell Proteomics*. 2004; 3:746–769. [PubMed: 15143156]
67. Volkin DB, Tsai PK, Dabora JM, Gress JO, Burke CJ, Linhardt RJ, Middaugh CR. Physical stabilization of acidic fibroblast growth factor by polyanions. *Arch Biochem Biophys*. 1993; 300:30–41. [PubMed: 7678726]
68. Miller RT, Douthart RJ, Dunker AK. Learning an objective alphabet of amino acid conformations in protein. *Curr. Tech. Prot. Chem*. 1993; 4:541–548.
69. Cao P, Mei JJ, Diao ZY, Zhang S. Expression, refolding, and characterization of human soluble BAFF synthesized in *Escherichia coli*. *Protein Expr Purif*. 2005; 41:199–206. [PubMed: 15802238]
70. Jin HJ, Dunn MA, Borthakur D, Kim YS. Refolding and purification of unprocessed porcine myostatin expressed in *Escherichia coli*. *Protein Expr Purif*. 2004; 35:1–10. [PubMed: 15039059]
71. Zegzouti H, Li W, Lorenz TC, Xie M, Payne CT, Smith K, Glenny S, Payne GS, Christensen SK. Structural and functional insights into the regulation of *Arabidopsis* AGC VIIIa kinases. *J Biol Chem*. 2006; 281:35520–35530. [PubMed: 16973627]
72. Yang F, Moss LG, Phillips GN Jr. The molecular structure of green fluorescent protein. *Nat Biotechnol*. 1996; 14:1246–1251. [PubMed: 9631087]
73. Ormo M, Cubitt AB, Kallio K, Gross LA, Tsien RY, Remington SJ. Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science*. 1996; 273:1392–1395. [PubMed: 8703075]
74. Stepanenko OV, Verkhusha VV, Kazakov VI, Shavlovsky MM, Kuznetsova IM, Uversky VN, Turoverov KK. Comparative studies on the structure and stability of fluorescent proteins EGFP, zFP506, mRFPI1, "dimer2", and DsRed1. *Biochemistry*. 2004; 43:14913–14923. [PubMed: 15554698]
75. Fukuda H, Arai M, Kuwajima K. Folding of green fluorescent protein and the cycle3 mutant. *Biochemistry*. 2000; 39:12025–12032. [PubMed: 11009617]
76. Reid BG, Flynn GC. Chromophore formation in green fluorescent protein. *Biochemistry*. 1997; 36:6786–6791. [PubMed: 9184161]
77. Habig WH, Pabst MJ, Jakoby WB. Glutathione S-transferases. The first enzymatic step in mercapturic acid formation. *J Biol Chem*. 1974; 249:7130–7139. [PubMed: 4436300]
78. Mannervic B, Danielson UH. Glutathione S-transferases - structure and catalytic activity. *Crit Rev Biochem*. 1988; 23:283–337. [PubMed: 3069329]
79. Bertone P, Kluger Y, Lan N, Zheng D, Christendat D, Yee A, Edwards AM, Arrowsmith CH, Montelione GT, Gerstein M. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res*. 2001; 29:2884–2898. [PubMed: 11433035]
80. Goh CS, Lan N, Douglas SM, Wu B, Echols N, Smith A, Milburn D, Montelione GT, Zhao H, Gerstein M. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J Mol Biol*. 2004; 336:115–130. [PubMed: 14741208]
81. Iidicula-Thomas S, Balaji PV. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci*. 2005; 14:582–592. [PubMed: 15689506]
82. Luan CH, Qiu S, Finley JB, Carson M, Gray RJ, Huang W, Johnson D, Tsao J, Reboul J, Vaglio P, Hill DE, Vidal M, Delucas LJ, Luo M. High-throughput expression of *C. elegans* proteins. *Genome Res*. 2004; 14:2102–2110. [PubMed: 15489332]

83. Sharp KA, Nicholls A, Friedman R, Honig B. Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry*. 1991; 30:9686–9697. [PubMed: 1911756]
84. Kochendoerfer G. Chemical and biological properties of polymer-modified proteins. *Expert Opin Biol Ther*. 2003; 3:1253–1261. [PubMed: 14640951]
85. Bokor M, Csizmok V, Kovacs D, Banki P, Friedrich P, Tompa P, Tompa K. NMR relaxation studies on the hydrate layer of intrinsically unstructured proteins. *Biophys J*. 2005; 88:2030–2037. [PubMed: 15613629]
86. Tompa P, Banki P, Bokor M, Kamasa P, Kovacs D, Lasanda G, Tompa K. Protein-water and protein-buffer interactions in the aqueous solution of an intrinsically unstructured plant dehydrin: NMR intensity and DSC aspects. *Biophys J*. 2006; 91:2243–2249. [PubMed: 16798808]
87. Soulages JL, Kim K, Arrese EL, Walters C, Cushman JC. Conformation of a group 2 late embryogenesis abundant protein from soybean. Evidence of poly (L-proline)-type II structure. *Plant Physiol*. 2003; 131:963–975. [PubMed: 12644649]
88. Naper, DH. Stabilization by attached polymer: steric stabilization, In: Napper, DH., editor. *Polymeric stabilization of colloidal dispersions*. London: Academic Press; 1983. p. 18–30.
89. Milner ST. Polymer Brushes. *Science*. 1991; 251:905–914. [PubMed: 17847384]
90. Bach H, Mazor Y, Shaky S, Shoham-Lev A, Berdichevsky Y, Gutnick DL, Benhar I. Escherichia coli maltose-binding protein as a molecular chaperone for recombinant intracellular cytoplasmic single-chain antibodies. *J Mol Biol*. 2001; 312:79–93. [PubMed: 11545587]
91. Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B, Longhi S. Assessing protein disorder and induced folding. *Proteins*. 2006; 62:24–45. [PubMed: 16287116]
92. Richarme G, Caldas TD. Chaperone properties of the bacterial periplasmic substrate-binding proteins. *J Biol Chem*. 1997; 272:15607–15612. [PubMed: 9188448]
93. Douette P, Navet R, Gerkens P, Galleni M, Levy D, Sluse FE. Escherichia coli fusion carrier proteins act as solubilizing agents for recombinant uncoupling protein 1 through interactions with GroEL. *Biochem Biophys Res Commun*. 2005; 333:686–693. [PubMed: 15961060]
94. Alsheikh MK, Heyen BJ, Randall SK. Ion binding properties of the dehydrin ERD14 are dependent upon phosphorylation. *J Biol Chem*. 2003; 278:40882–40889. [PubMed: 12917402]
95. Chakrabortee S, Boschetti C, Walton LJ, Sarkar S, Rubinsztein DC, Tunnacliffe A. Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function. *Proc Natl Acad Sci U S A*. 2007; 104:18073–18078. [PubMed: 17984052]
96. Hara M, Terashima S, Fukaya T, Kuboi T. Enhancement of cold tolerance and inhibition of lipid peroxidation by citrus dehydrin in transgenic tobacco. *Planta*. 2003; 217:290–298. [PubMed: 12783337]
97. Zhang YB, Howitt J, McCorkle S, Lawrence P, Springer K, Freimuth P. Protein aggregation during overexpression limited by peptide extensions with large net negative charge. *Protein Expr Purif*. 2004; 36:207–216. [PubMed: 15249042]
98. Cohn EJ, Gross J, Johnson OC. The Isoelectric Points of the Proteins in Certain Vegetable Juices. *J Gen Physiol*. 1919; 2:145–160. [PubMed: 19871798]
99. Loeb J. Amphoterics Colloids : Ii. Volumetric Analysis of Ion-Protein Compounds; the Significance of the Isoelectric Point for the Purification of Amphoterics Colloids. *J Gen Physiol*. 1918; 1:237–254. [PubMed: 19871741]
100. Shih YC, Prausnitz JM, Blanch HW. Some characteristics of protein precipitation by salts. *Biotechnol Bioeng*. 1992; 40:1155–1164. [PubMed: 18601066]
101. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN. Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*. 2009; 31:328–335. [PubMed: 19260013]
102. Firestine SM, Salinas F, Nixon AE, Baker SJ, Benkovic SJ. Using an AraC-based three-hybrid system to detect biocatalysts in vivo. *Nat Biotechnol*. 2000; 18:544–547. [PubMed: 10802623]
103. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*. 2011; 332:816–821. [PubMed: 21566186]

104. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. De novo enzyme design using Rosetta3. PLoS One. 2011; 6:e19230. [PubMed: 21603656]

Figure 1. Evaluation of intrinsic disorder in the members of the *Arabidopsis thaliana* dehydrin family

A. POND[®] VLXT analysis of ERD10. **B.** CH-plot analysis of: dehydrins (squares) ERD10 (grey), ERD14 (blue), COR47 (yellow), Rab18 (green), Xero1 (cyan), LTI30 (red); ordered proteins from DISPROT (light grey triangles), disordered proteins from DISPROT (light grey diamonds). The black delineates the boundary above which proteins which most extended, disordered proteins locate. **C.** Compositional profiling of dehydrins: Disprot (black), ERD10 (grey), ERD14 (blue), COR47 (yellow), Rab18 (green), Xero1 (cyan), LTI30 (red).

Figure 2. Analyzing the solubilization capabilities of dehydrins

A. SDS-PAGE analysis of CTLA4 solubility alone or fused to ERD10, ERD14, LTI30

dehydrins. **B.** The ability of ERD10 fusion (black bars) to enhance the percentage of soluble protein yield compared to the 6x-His fusion control (grey bars) for 10 insoluble proteins.

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 (dark bars) with Trx, GST, NusA and MBP fusions (light bars) marketed to enhance soluble protein expression. **B.** Western blot analysis of the protein TIMP2 hybridized to the various fusion partners (anti-6x His blot, G-18 from Santa Cruz Biotechnologies).

Figure 4. Effect of fusion of various EBs on conformational stability of GFP and transferase activity of GST

A. Coomassie gel image showing the enrichment of GFP after purification on a NiNTA column. Lanes contain the following samples: 1. Molecular weight standard (Invitrogen); 2. 6xHis-GFP (29.3 kD); 3. 6xHis-MBP-GFP (69.5 kD); 4. 6xHis-EB60A-GFP (36.1 kD); 5. 6xHis-EB60B-GFP (36.0 kD); 6. 6xHis-EB144-GFP (44.4 kD); 7. 6xHis-EB250-GFP (55.5 kD). **B.** The fluorescence unfolding curve of GFP constructs fused to: 6xHis (◆), MBP (■), EB60A (▲), EB60B (▼), EB144 (right-facing triangle), EB250 (left-facing triangle). **C.** A

table reporting the transferase activity of purified GST-fusions toward the synthetic CDNB substrate. Please note the following: (*) the molecular weight is based on the dimer weight of GST; (#) the GST transferase activity assay was measured in units of $\mu\text{mol}_{\text{CDNB}} \bullet \text{min}^{-1} \bullet \text{pmol}_{\text{GST}}^{-1}$ to correct for differences in the molecular weight of the various fusions.

Figure 5. Removal of the EB fusions via proteolytic cleavage using enterokinase (EK)

A. Coomassie gel image visualizing the cleavage of the 6x-MBP fusion from GFP at 0, 2, 4, 8 and 24 hour time points. **B.** Coomassie gel image visualizing the cleavage of the 6x-EB60A fusion from GFP at 0, 2, 4, 8 and 24 hour time points. **C.** Coomassie gel image visualizing the cleavage of the 6x-EB250 fusion from GFP at 0, 2, 4, 8 and 24 hour time points.

Table 1
Compilation of characteristics, disorder and solubility predictors for dehydrin proteins originating from *A. thaliana*.

Protein	Accession Number	MW (kD)	Net Charge	Length (AA)	Percent ^d Disorder		Mean ^b	WH ^c		ProSol ^d		SolPro ^e	
					VLXT	VSL2		Hydropathy	Solubility Predictor	Solubility Predictor	Solubility Predictor	Solubility Predictor	Solubility Predictor
ERD10	NP_564114	29.4	-15	259	64%	100%	0.3529	75% soluble	75% soluble	0.72 soluble	0.81 soluble	0.81 soluble	0.81 soluble
ERD14	NP_177745	20.8	-9	185	64%	100%	0.3595	65% soluble	65% soluble	0.32 insoluble	0.82soluble	0.82soluble	0.82soluble
COR47	NP_195554	18.0	-5	163	60%	100%	0.319	52% insoluble	52% insoluble	0.83 soluble	0.92 soluble	0.92 soluble	0.92 soluble
Rab18	CAA48178	18.5	0	186	80%	100%	0.3686	97% insoluble	97% insoluble	0.55 insoluble	0.93 soluble	0.93 soluble	0.93 soluble
XERO1	NP_190667	13.4	+3	128	60%	100%	0.5939	97% insoluble	97% insoluble	0.86 soluble	0.93 soluble	0.93 soluble	0.93 soluble
LT130	NP_190666	20.9	+6	193	21%	100%	0.3697	94% insoluble	94% insoluble	0.86 soluble	soluble	soluble	soluble

^a. Percentage of amino acids found disordered using the POND[®] 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 values 0–0.6 being insoluble to values 0.6–1.0 being soluble (26).

^e. The probability value for the soluble/insoluble designation range from 0.5–1.0 (25).

Table 2

Artificial entropic bristle (EB) characteristics

EB fusion	A.A. Composition	EB length	MW (kD)	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

Table 3

Compilation of disorder and solubility predictions for various protein fusions

Protein	Percent ^a Disordered		Mean ^b	WH ^c		ProSol ^d	SolPro ^e
	VLXT	VSL2		Solubility Predictor	Solubility Predictor		
EB60A	100%	100%	0.2281	97% soluble	0.88 soluble	0.97 soluble	0.97 soluble
EB60B	100%	100%	0.2133	97% soluble	0.89 soluble	1.00 soluble	1.00 soluble
EB144	100%	100%	0.2640	86% insoluble	0.88 soluble	0.97 soluble	0.97 soluble
EB250	100%	100%	0.3424	92% insoluble	0.87 soluble	0.97 soluble	0.97 soluble
GST	12%	14%	0.4572	58% soluble	0.58 insoluble	0.78 insoluble	0.78 insoluble
MBP	11%	17%	0.4640	52% insoluble	0.77 soluble	0.94 soluble	0.94 soluble
NusA	47%	20%	0.4423	95% soluble	0.66 soluble	0.58 soluble	0.58 soluble
Trx	6.4%	12%	0.028	72.6% soluble	0.36 insoluble	0.89 soluble	0.89 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 values 0–0.6 being insoluble to values 0.6–1.0 being soluble (26).^e. The probability value for the soluble/insoluble designation range from 0.5–1.0 (25).

Table 4

Summary of Solubility Data for twenty test protein candidates

Protein	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 %
CATA9	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 <0.05	na	65%	75%	75%	95%	100%
% p-value vs. MBP <0.05	20%	na	50%	50%	85%	85%