Published in final edited form as:

ACS Chem Biol. 2018 October 19; 13(10): 2855–2861. doi:10.1021/acscchembio.8b00696.

Systematic evaluation of soluble protein expression using a fluorescent unnatural amino acid reveals no reliable predictors of tolerability

Zachary M. Hostetler[†], John J. Ferrie[‡], Marc R. Bornstein[†], Itthipol Sungwienwong[‡], E. James Petersson^{*‡}, and Rahul M. Kohli^{**†}

[†]Department of Medicine, Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States

[‡]Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States

Abstract

Improvements in genetic code expansion have made preparing proteins with diverse functional groups almost routine. Nonetheless, unnatural amino acids (Uaas) pose theoretical burdens on protein solubility, and determinants of position-specific tolerability to Uaas remain underexplored. To broadly examine associations, we systematically assessed the effect of substituting the fluorescent Uaa, acridonylalanine, at more than fifty chemically, evolutionarily, and structurally diverse residues in two bacterial proteins—LexA and RecA. Surprisingly, properties that ostensibly contribute to Uaa tolerability—like conservation, hydrophobicity, or accessibility—demonstrated no consistent correlations with resulting protein solubility. Instead, solubility closely depended on the location of the substitution within the overall tertiary structure, suggesting that intrinsic properties of protein domains, and not individual positions, are stronger determinants of Uaa tolerability. Consequently, those who seek to install Uaas in new target proteins should consider broadening, rather than narrowing, the types of residues screened for Uaa incorporation.

Keywords

Genetic code expansion; nonsense codon suppression; protein solubility; non-canonical amino acids; SOS response

^{*}Corresponding Authors: ejpetersson@sas.upenn.edu. ^{**} rkohli@penmedicine.upenn.edu.

Author Contributions

Z.M.H., J.J.F., M.R.B., I.S., E.J.P., and R.M.K. designed the experiments. Z.M.H performed all the experiments with assistance from M.R.B. J.J.F. performed the Rosetta simulations. I.S. synthesized Acd. Z.M.H. and J.J.F. performed the data analysis with input from all authors. Z.M.H., E.J.P., and R.M.K. wrote the manuscript with input from all authors.

Supporting Information

Supporting Information Available: The following material is available free of charge via the internet. Experimental methods, supplemental figures and tables, and associated references.

Notes

The authors declare no competing financial interest.

Technological advances in genetic code expansion have encouraged the design of proteins with a wide range of reactive residues, post-translational modifications, photocaged groups, or intrinsic fluorophores.^{1–3} Nonsense codon suppression using orthogonal tRNA/aminoacyl-tRNA synthetase pairs enables direct incorporation of chemically diverse unnatural amino acids (Uaas, also known as non-canonical amino acids) into proteins *in vivo*. Many efforts have sought to boost the efficiency of Uaa incorporation, including evolving more efficient aminoacyl-tRNA synthetases and recoding the *E. coli* genome to remove competing translational release factors.^{4,5} Although these developments can improve total yields of modified proteins, factors governing the position-dependent effects of Uaa substitution on protein solubility remain understudied.

Recent reports have demonstrated that the position of a Uaa can affect the level of total protein expressed, both in cell-free and cell-based systems.^{6–10} Investigations of 20 positions in IFN- α and 33 positions in VSV glycoprotein revealed varying total protein yields, from 0 to 95% of wildtype.^{11,12} Despite these observations, explanations for position-dependent differences in total amounts of Uaa-containing proteins have been limited, and no studies have explicitly addressed UAA incorporation versus the resulting protein solubility.

Unnatural amino acid mutagenesis could hypothetically operate under well-accepted principles that govern the effects of natural amino acid mutation. For example, substitution of a nonpolar for a polar residue within the hydrophobic core generally destabilizes proteins, whereas mutations on the solvent-exposed surface less frequently affect solubility.^{13,14} Unsurprisingly, evolutionarily-conserved residues largely disfavor mutation.^{15–17} Substituting bulkier and more chemically-diverse Uaas into a protein can restrict function¹⁸ and therefore could pose similar burdens on folding and solubility. Nevertheless, the applicability of principles of natural amino acid mutagenesis to Uaa mutagenesis remains unknown.

Suggested guidelines or approaches for choosing Uaa-tolerant sites have been proposed. Some groups favor residues with structural similarity to the Uaa.⁹ Others assert that candidate positions should be first assessed for mutational tolerability with natural amino acids¹⁰ or that proteins should be thoroughly screened by random incorporation of Uaas into protein-GFP fusions to reveal positions that label with high efficiency.^{19,20} Nonetheless, the feasibility of using position-specific properties to increase soluble protein expression remains untested.

To address these open questions, we aimed to explore factors that impact Uaa incorporation and soluble protein production. By employing an intrinsically fluorescent Uaa, acridonylalanine (Acđ),^{6,21,22} we directly detect labeled protein in cell lysate samples, overcoming the inability of past studies to measure levels of both total and soluble expressed protein. Our systematic survey of more than fifty sites across two proteins reveals that while incorporation efficiency is relatively similar, protein solubility, and by extension Uaa tolerability, varies widely across different positions. However, most position-specific physicochemical, evolutionary, and structural properties, some of which have been previously suggested to improve yield, were minimally predictive; instead, solubility more strongly associated with the identity of the protein domain. After controlling for this domain

effect, we found that only a few factors, such as a tolerance for aromatic residues, moderately trended with protein solubility. To our knowledge, this work currently represents the most systematic effort evaluating predictive factors for producing soluble Uaa-containing proteins.

RESULTS AND DISCUSSION

The bacterial protein LexA, a multi-domain repressor of the DNA damage response, has characteristics that made it well-suited to this broader survey. Wild-type *E. coli* LexA is well-behaved in overexpression and has previously tolerated selective unnatural amino acid (Uaa) incorporation.²² Additionally, the availability of protein crystal structures and a multiple sequence alignment for LexA enabled retrieval of position-specific properties from databases or servers that require these data as inputs (Table S1). For every position in LexA, we calculated established metrics across different classes of properties: physicochemical, such as hydrophobicity; evolutionary, such as conservation; and structural, such as solvent accessibility (Table 1). Using these metrics, we selected 32 positions spanning both domains of LexA, deliberately avoiding known deleterious mutants as well as the most conserved or hydrophobic positions (Figure 1a, Table S2). Our selected positions sample the remaining metrics well (Figure 1b, Figure S1, and Figure S2), indicating that this series is well-positioned to explore how aromatic, accessible, or poorly-conserved residues might differentially tolerate Uaa incorporation.

Historically, measuring Uaa incorporation efficiencies *in vivo* has overlooked protein solubility issues, while labeling Uaa-containing proteins *in vitro* has suffered from incomplete sample recovery and detection. Crucially, we chose to measure both total and soluble protein levels by using the fluorescent Uaa acridonylalanine (Acd, Figure 1c), which already possesses an optimized tRNA/tRNA synthetase pair for *in vivo* incorporation.^{21,22} This system offers several advantages. First, Acd incorporation occurs during protein overexpression without post-translational labeling. Second, measurements of Acd fluorescence at the expected size on an SDS-PAGE gel are directly proportional to levels of protein with successfully-incorporated Acd. Finally, gel-based detection of Acd demonstrates a broad dynamic range, enabling us to detect quantitative differences in the expression of Acd-containing LexA mutants (Figure S3).

Expression levels for a single protein can range widely due to experimental variability, making quantitative comparison between different proteins difficult. To overcome this challenge, we overexpressed the 32 LexA mutants in the presence of both Acd and the Acd-specific tRNA/tRNA synthetase using autoinduction media for consistency in the timing and duration of protein production. Following overexpression, we measured fluorescence intensity levels of Acd-containing LexA protein in both the whole cell lysate and soluble fraction (Figure 1d). The use of purified Acd-containing LexA as a standard enabled quantitative and reproducible comparisons of protein amounts across independent experiments (Figure S4).

Parallel overexpression of all 32 LexA mutants allowed us to investigate how amounts of total expressed Acd-labeled LexA proteins differed (Table S3). A plot of logarithmically-

transformed total protein amounts shows uniformly high protein expression (mean = 3.1) with minor variability (SD = 0.16) (Figure 2a). While past studies have suggested that the identity of nucleotides surrounding the stop codon can impact nonsense codon suppression efficiencies,^{23–25} we did not observe this relationship (Figure S5). Rather, the small 4.5-fold difference between measurements of the lowest and highest-expressing samples suggests that changing the position of Acd does not substantially alter Acd incorporation rates *in vivo*, and that incorporation is not a major bottleneck with regards to solubility.

Recognizing the consistency in total levels of expressed protein, we next evaluated whether levels of soluble protein differed. A distribution of logarithmically-transformed soluble protein amounts (Figure 2a) reveals more variability (mean = 2.2, SD = 0.86). Measured soluble protein amounts ranged nearly 40-fold from the lowest detectable measurements to the highest, a ten-fold increase over the range of total protein amounts. Because both measurements are paired, we can isolate the position-dependent effect of Acd incorporation on solubility by calculating the soluble fraction of total protein, which should exclude variability due to differences in total protein production. The soluble fractions of Acd-labeled LexA mutants still vary considerably, from 0% to nearly 70% of total protein expressed (Figure 2b, Table S3). This result not only corroborates previous observations of position-dependent effects on total protein expression,^{11,12} but it also establishes the heightened sensitivity of protein solubility to Uaa incorporation.

Observing that the position of Acd can substantially impact protein solubility, we next asked which of the properties that ostensibly affect Uaa tolerability might correlate with solubility. We fitted the soluble fraction as a response variable to each property in individual linear regression models (Table S4 and Table S5). For almost all of the properties we evaluated, the explained variability (adj. R^2) was about 5% or less, indicating that if any property-specific effect exists, it is insubstantial and likely below our ability to detect with a sample size of 32 (Figure S6 and Figure S7). We note that particular properties—such as accessibility, conservation, and hydrophobicity—did not explain any substantial variation in our data, despite past suggestions that choosing accessible, less-conserved, and chemically-similar residues may yield more soluble Uaa-containing protein (Figure 2c).

Conspicuously, several highly-correlated properties each explained around 50% of the variability in our data, including individual residue position (adj. R^2 = 0.53), secondary structure (adj. R^2 = 0.45), and overall protein domain (adj. R^2 = 0.53) (Figure 2d and Figure 2e). Specifically, we obtained more soluble protein when Acd was incorporated within the first 74 residues of LexA, which includes all three of the α -helices that comprise the N-terminal domain. By contrast, Acd incorporation within the β -sheets of the C-terminal domain resulted in much lower proportions of soluble protein. The nearly uniform secondary structure composition of each domain limited our ability to interpret whether Acd tolerability is due to local secondary structure effects or global protein domain stabilities.

Excluding the effect that secondary or tertiary structure has on protein solubility could reveal minor trends obscured in the overall dataset. To address this possibility, in individual linear regression models, we fitted each property along with protein domain as explanatory factors for the soluble fraction (Table S6 and Table S7). By controlling for domain, we could detect

a minor correlation between the soluble fraction and the evolutionary tolerance of any given position to an aromatic residue (Figure 2f). However, remaining factors—including notable ones such as conservation, hydrophobicity, and accessibility—either did not explain any substantial variation in the data or demonstrated inconsistent trends between domains (Figure 2c). Consequently, our extended LexA analysis reaffirmed that the tolerability of a protein domain to Acd—or possibly the tolerability of a secondary structure type—overwhelmingly determines soluble protein expression.

Studying Acd incorporation in a distinct protein scaffold with mixed α/β character could help dissect the similar effects we observed from the highly-correlated domain and secondary structure factors with LexA. Thus, we extended our survey to RecA, a bacterial ATPase that binds LexA to suppress its repressor function.²⁶ We selected positions in *E. coli* RecA that satisfied one or more criteria: high accessibility, low conservation, few inter-residue contacts, or prior functional tolerance to mutation (Figure 3a).²⁷ After expressing these mutants with Acd and measuring protein amounts, we again observed greater variability in logarithmically-transformed soluble protein levels (mean = 3.42, SD = 0.40) compared to total protein levels (mean = 3.72, SD = 0.17) (Figure 3b, Figure 3c, Table S8). Similar to LexA, most properties examined did not explain much variation in the fractions of soluble protein (Figure 3d), with the exception that solubility modestly trended with domain type and tolerance to aromatics (Table S9). However, unlike in LexA, no clear relationship existed between protein solubility and type of secondary structure (Figure 3e), a result consistent with a more limited prior survey of GFP.⁸ This survey in RecA bolsters a model in which the intrinsic Uaa tolerability of a protein domain remains the key obstacle for the production of soluble protein.

Searching for easily-determined properties that correlate with Acd tolerability may have eliminated from consideration more complicated properties with higher predictive ability. Additionally, linear regression modeling may have over-simplified the inter-dependence of certain properties and protein solubility. Previously, Rosetta modeling has predicted the ΔG associated with a particular mutation and identified tolerated mutations within a protein.^{28–30} Speculating that Rosetta modeling could recapitulate our experimental results, we used the Rosetta Modeling Suite to simulate the resulting energy associated with Acd incorporation in LexA or RecA. However, we observed no significant correlations between simulated energies and soluble fractions of LexA or RecA (Figure S8 and Figure S9). Incidentally, we noted that nearly all high-energy positions in LexA experimentally yielded insoluble protein and may therefore have been useful in filtering out those positions; however, we did not observe a similar energy threshold effect for RecA. Accordingly, further refinement towards predicting Uaa incorporation using Rosetta is required in order to recapitulate experimental data and exclude higher-energy and lower-solubility mutants.

CONCLUSION

The expression of soluble protein is a major bottleneck for the study of protein function. Here, we leveraged the fluorescence of Acd to study how protein solubility is impacted by Uaa mutagenesis. In two bacterial proteins, we demonstrated the dramatic impact that Uaa position has on protein solubility. Surprisingly, a number of amino acid properties that

purportedly contribute to Uaa tolerability—including low evolutionary conservation, similar hydrophobic character, or high surface accessibility—were unreliable predictors of protein solubility. Instead, these inconsistent relationships suggest that consideration of specific amino acid features for successful Uaa mutagenesis is less critical than previously thought. Rather, we speculate that the Uaa tolerability of a protein domain may matter more. Our results also emphasize a continued need to explore, through theory and experiment, the steric and chemical burdens different Uaas pose to the expression of soluble protein. In the absence of reliable predictors or refined simulation algorithms for Uaa tolerability, a chemical biologist pursuing Uaa incorporation in a new protein, as of now, should broaden rather than narrow the types of residues screened for Uaa tolerability when possible.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank members of the Kohli and Petersson laboratories for general advice, and we are grateful to E. Schutsky for input in preparing the manuscript. This work was supported by the National Institutes of Health (R01-GM127593 to R.M.K. and E.J.P.) and the National Science Foundation (NSF, CHE-1708759 to E.J.P.). Z.M.H. was supported by the NIH Chemistry Biology Interface Training Program (T32-GM071399). J.J.F. was supported by the NSF Graduate Research Fellowship Program (DGE-1321851). I.S. was supported by the Royal Thai Foundation.

REFERENCES

- (1). Young TS, and Schultz PG (2010) Beyond the Canonical 20 Amino Acids: Expanding the Genetic Lexicon. *J. Biol. Chem* 285, 11039–11044. [PubMed: 20147747]
- (2). Neumann-Staubitz P, and Neumann H (2016) The use of unnatural amino acids to study and engineer protein function. *Curr. Opin. Struct. Biol* 38, 119–128. [PubMed: 27318816]
- (3). Xiao H, and Schultz PG (2016) At the Interface of Chemical and Biological Synthesis: An Expanded Genetic Code. *Cold Spring Harb. Perspect. Biol* 8, a023945. [PubMed: 27413101]
- (4). Chatterjee A, Sun SB, Furman JL, Xiao H, and Schultz PG (2013) A Versatile Platform for Single- and Multiple-Unnatural Amino Acid Mutagenesis in *Escherichia coli*. *Biochemistry* 52, 1828–1837. [PubMed: 23379331]
- (5). Lajoie MJ, Rovner AJ, Goodman DB, Aerni H-R, Haimovich AD, Kuznetsov G, Mercer J. a, Wang HH, Carr P. a, Mosberg J. a, Rohland N, Schultz PG, Jacobson JM, Rinehart J, Church GM, and Isaacs FJ (2013) Genomically recoded organisms expand biological functions. *Science* 342, 357–360. [PubMed: 24136966]
- (6). Hamada H, Kameshima N, Szymanska A, Wegner K, Lankiewicz Ł, Shinohara H, Taki M, and Sisido M (2005) Position-specific incorporation of a highly photodurable and blue-laser excitable fluorescent amino acid into proteins for fluorescence sensing. *Bioorg. Med. Chem* 13, 3379–3384. [PubMed: 15848750]
- (7). Goerke AR, and Swartz JR (2009) High-level cell-free synthesis yields of proteins containing site-specific non-natural amino acids. *Biotechnol. Bioeng* 102, 400–416. [PubMed: 18781689]
- (8). Albayrak C, and Swartz JR (2013) Cell-free co-production of an orthogonal transfer RNA activates efficient site-specific non-natural amino acid incorporation. *Nucleic Acids Res.* 41, 5949–5963. [PubMed: 23589624]
- (9). Hammill JT, Miyake-Stoner S, Hazen JL, Jackson JC, and Mehl RA (2007) Preparation of site-specifically labeled fluorinated proteins for ¹⁹F-NMR structural characterization. *Nat. Protoc* 2, 2601–2607. [PubMed: 17948003]

- (10). Hino N, Hayashi A, Sakamoto K, and Yokoyama S (2006) Site-specific incorporation of non-natural amino acids into proteins in mammalian cells with an expanded genetic code. *Nat. Protoc* 1, 2957–2962. [PubMed: 17406555]
- (11). Zhang B, Xu H, Chen J, Zheng Y, Wu Y, Si L, Wu L, Zhang C, Xia G, Zhang L, and Zhou D (2015) Development of next generation of therapeutic IFN- α 2b via genetic code expansion. *Acta Biomater.* 19, 100–111. [PubMed: 25769229]
- (12). Zheng Y, Yu F, Wu Y, Si L, Xu H, Zhang C, Xia Q, Xiao S, Wang Q, He Q, Chen P, Wang J, Taira K, Zhang L, and Zhou D (2015) Broadening the versatility of lentiviral vectors as a tool in nucleic acid research via genetic code expansion. *Nucleic Acids Res.* 43, e73. [PubMed: 25765642]
- (13). Lau KF, and Dill KA (1990) Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. U. S. A* 87, 638–642. [PubMed: 2300551]
- (14). Markiewicz P, Kleina LG, Cruz C, Ehret S, and Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. *J. Mol. Biol* 240, 421–433. [PubMed: 8046748]
- (15). Lim WA, and Sauer RT (1989) Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* 339, 31–36. [PubMed: 2524006]
- (16). Campbell-Valois F-X, Tarassov K, and Michnick SW (2005) Massive sequence perturbation of a small protein. *Proc. Natl. Acad. Sci. U. S. A* 102, 14988–14993. [PubMed: 16214894]
- (17). Romero PA, Tran TM, and Abate AR (2015) Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Natl. Acad. Sci. U. S. A* 112, 7159–7164. [PubMed: 26040002]
- (18). Luo J, Uprety R, Naro Y, Chou C, Nguyen DP, Chin JW, and Deiters A (2014) Genetically encoded optochemical probes for simultaneous fluorescence reporting and light activation of protein function with two-photon excitation. *J. Am. Chem. Soc* 136, 15551–15558. [PubMed: 25341086]
- (19). Reddington SC, Baldwin AJ, Thompson R, Brancale A, Tippmann EM, and Jones DD (2015) Directed evolution of GFP with non-natural amino acids identifies residues for augmenting and photoswitching fluorescence. *Chem. Sci* 6, 1159–1166. [PubMed: 29560203]
- (20). Arpino JAJ, Baldwin AJ, McGarrity AR, Tippmann EM, and Jones DD (2015) In-frame amber stop codon replacement mutagenesis for the directed evolution of proteins containing non-canonical amino acids: identification of residues open to bio-orthogonal modification. *PLoS One* 10, e0127504. [PubMed: 26011713]
- (21). Speight LC, Muthusamy AK, Goldberg JM, Warner JB, Wissner RF, Willi TS, Woodman BF, Mehl R. a, and Petersson EJ (2013) Efficient synthesis and in vivo incorporation of acridon-2-ylalanine, a fluorescent amino acid for lifetime and Förster resonance energy transfer/ luminescence resonance energy transfer studies. *J. Am. Chem. Soc* 135, 18806–18814. [PubMed: 24303933]
- (22). Sungwienwong I, Hostetler ZM, Blizzard RJ, Porter JJ, Driggers CM, Mbengi LZ, Villegas JA, Speight LC, Saven JG, Perona JJ, Kohli RM, Mehl RA, and Petersson EJ (2017) Improving target amino acid selectivity in a permissive aminoacyl tRNA synthetase through counter-selection. *Org. Biomol. Chem* 15, 3603–3610. [PubMed: 28397914]
- (23). Miller JH, and Albertini AM (1983) Effects of surrounding sequence on the suppression of nonsense codons. *J. Mol. Biol* 164, 59–71. [PubMed: 6188840]
- (24). Pott M, Schmidt MJ, and Summerer D (2014) Evolved sequence contexts for highly efficient amber suppression with noncanonical amino acids. *ACS Chem. Biol* 9, 2815–2822. [PubMed: 25299570]
- (25). Xu H, Wang Y, Lu J, Zhang B, Zhang Z, Si L, Wu L, Yao T, Zhang C, Xiao S, Zhang L, Xia Q, and Zhou D (2016) Re-exploration of the Codon Context Effect on Amber Codon-Guided Incorporation of Noncanonical Amino Acids in *Escherichia coli* by the Blue-White Screening Assay. *Chembiochem* 17, 1250–1256. [PubMed: 27028123]
- (26). Culyba MJ, Mo CY, and Kohli RM (2015) Targets for Combating the Evolution of Acquired Antibiotic Resistance. *Biochemistry* 54, 3573–3582. [PubMed: 26016604]

- (27). McGrew DA, and Knight KL (2003) Molecular design and functional organization of the RecA protein. *Crit. Rev. Biochem. Mol. Biol* 38, 385–432. [PubMed: 14693725]
- (28). Smith CA, and Kortemme T (2011) Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PLoS One* 6, e20451. [PubMed: 21789164]
- (29). Kellogg EH, Leaver-Fay A, and Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* 79, 830–838. [PubMed: 21287615]
- (30). Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL, Das R, Baker D, Kuhlman B, Kortemme T, and Gray JJ (2017) The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput* 13, 3031–3048. [PubMed: 28430426]

Figure 1: Scanning a variety of positions in LexA for Acd tolerability.

(a) Positions chosen for Uaa incorporation in the LexA dimer. Chosen positions are depicted in yellow, α -helices in blue, and β -sheets in green. (b) Principal component analysis (PCA) of LexA positions determined by multiple structural, evolutionary, and physicochemical properties (see methods). All residues in LexA were scored and plotted against the first two principal components, with positions chosen for Uaa incorporation highlighted in yellow. Arrow segments represent a few notable variables among those used in PCA loaded onto the plotted data. (c) Chemical structure of Acd with indicated excitation and emission peaks. (d) Acd-labeled LexA samples visualized in 15% SDS-PAGE gels by Coomassie staining (left) or UV excitation (right). Lanes 1–3 show purified LexA standards. Lanes 4–11 show paired total and soluble fractions from four individual mutants as representative examples.

Figure 2: Features associated with soluble Acd-labeled LexA proteins.

(a) Smoothed density plots of $\log_{10}$ -transformed amounts of total protein or soluble protein. (b) Average $\log_{10}$ -transformed soluble protein amounts overlaid on average $\log_{10}$ -transformed total protein amounts for each mutant. Error bars indicate the standard deviation from three individual replicates each derived from separate clones. (c) Plots of the average fraction of soluble protein as a function of three selected parameters: conservation, hydrophobicity, and accessibility. Other parameters were also examined (Figures S6 and S7). Fits for the entire LexA dataset to individual linear regression models yield best fit lines (solid black) and 95% confidence intervals (shaded gray). Fits of data from each separate LexA domain yield best fit lines for the NTD (dashed green) or CTD (dashed blue). (d) Boxplots comparing the average fraction of soluble protein against either domain or secondary structure, with individual averages overlaid. Differences between groups were evaluated using Tukey's HSD test for multiple pairwise comparisons (** = p-value < 0.01; *** = p-value < 0.001). (e) Plot of the average fraction of soluble protein as a function of position in the LexA sequence, with error bars indicating the standard deviation from three replicates. Above, the secondary and tertiary structure of LexA is indicated; α -helices are depicted as green ovals and β -sheets as blue rectangles. (f) Separate boxplots for each LexA

domain indicating the relationship between average fraction of soluble protein and evolutionary tolerance at each position to tryptophan, as one example of an aromatic residue.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3: Features associated with soluble Acd-labeled RecA proteins.

(a) Positions chosen for Acd incorporation in RecA. Chosen positions are depicted in yellow, α -helices in blue, and β -sheets in green. (b) Smoothed density plots of $\log_{10}$ -transformed amounts of total protein or soluble protein. (c) Average $\log_{10}$ -transformed soluble protein amounts overlaid on average $\log_{10}$ -transformed total protein amounts for each mutant. Error bars indicate the standard deviation from three individual replicates each derived from separate clones. (d) Plots of the average fraction of soluble protein as a function of three selected parameters: conservation, hydrophobicity, and accessibility. Fits to individual linear regression models yield best fit lines (solid black) and 95% confidence intervals (shaded gray). (e) Boxplots comparing the average fraction of soluble protein against domain or secondary structure, with individual averages overlaid.

Table 1:List of properties examined for association with Uaa tolerability^a

Property	Details
<i>Physicochemical</i>	
Hydrophobicity	Discrete number describing experimentally-determined hydrophobic indices (usually kcal/mol)
Similar to Phe, Trp, or Tyr	Discrete number calculated from a substitution matrix similarity score table
Volume	Size of residue ($\text{\AA}^3$)
<i>Evolutionary</i>	
Conservation	Calculated score describing the degree of conservation from a multiple sequence alignment
Tolerance to Phe, Trp, or Tyr	Presence or absence of a particular residue substitution within a multiple sequence alignment
<i>Structural</i>	
Solvent Accessible Area	Surface area of residue exposed to solvent ($\text{\AA}^2$)
Accessibility	Ratio of solvent accessible area relative to the theoretical maximum surface area of a residue
Fractional Loss of Accessible Area	Area lost when a residue is buried upon folding ($\text{\AA}^2$)
Surrounding Hydrophobicity	Numerical sum of local hydrophobic indices assigned to residues within 8 $\text{\AA}$
Average hydrophobic gain/ratio	Total numerical increase or a ratio describing the difference in local surrounding hydrophobicity between unfolded to folded state
Position	Residue number in primary sequence of protein
Secondary/tertiary structure	Categorical assignment to secondary structure type or classification into a protein domain
Nearby contacts	Discrete number of contacts within 8 or 14 $\text{\AA}$, either using C_α or C_β atoms
Noncovalent contacts	Presence or absence of interaction with another residue through a H-bond, cation- π , hydrophobic, or polar contact
Long Range Order	Presence or absence of contacts with residues close in space but far in sequence
Surrounding Residues	Number of residues within 8 $\text{\AA}$ contextualized by sequence position

^aRefer to Table S1 for more details and references to relevant databases