molecular pharmaceutics

Article

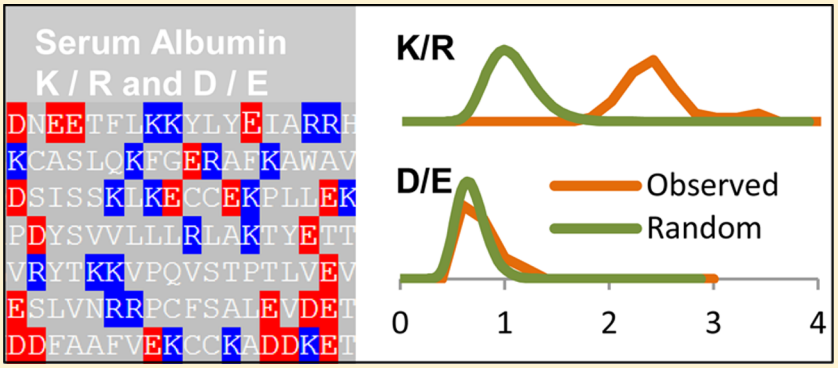
pubs.acs.org/molecularpharmaceutics

Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design

Jim Warwicker,*[,†] Spyros Charonis,[‡] and Robin A. Curtis[‡]

[†]Faculty of Life Sciences, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, U.K.

[‡]School of Chemical Engineering and Analytical Sciences, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, U.K.



ABSTRACT: Prediction and engineering of protein solubility is an important but imprecise area. While some features are routinely used, such as the avoidance of extensive non-polar surface area, scope remains for benchmarking of sequence and structural features with experimental data. We study properties in the context of experimental solubilities, protein gene expression levels, and families of abundant proteins (serum albumin and myoglobin) and their less abundant paralogues. A common feature that emerges for proteins with elevated solubility and at higher expression and abundance levels is an increased ratio of lysine content to arginine content. We suggest that the same properties of arginine that give rise to its recorded propensity for specific interaction surfaces also lead to favorable interactions at nonspecific contacts, and thus lysine is favored for proteins at relatively high concentration. A survey of protein therapeutics shows that a significant subset possesses a relatively low lysine to arginine ratio, and therefore may not be favored for high protein concentration. We conclude that modulation of lysine and arginine content could prove a useful and relatively simple addition to the toolkit available for engineering protein solubility in biotechnological applications.

KEYWORDS: *protein aggregation, bioinformatics, solubility prediction, biologics, amino acid side chain charge*

INTRODUCTION

Protein solubility has been a crucial property leading up to and through the advent of recombinant protein production. Prior to the era of recombinant protein expression, the natural abundance of proteins largely determined which were studied in detail. When it became possible to overexpress proteins, and subsequently with whole genome sequencing, proteins could be chosen on the basis of the underlying science in question, but high level expression of soluble protein is not guaranteed.[1] This is related to the problem of avoiding protein aggregates when developing and formulating proteins in the increasingly important pharmaceutics area of biologics.[2]

Several properties have been used for correlating and predicting protein solubility. A corollary of the oil-drop model for protein folding[3] is that non-polar regions could lead to protein−protein interactions, a common theme throughout the analysis of protein structure and function.[4] One approach is to rank non-polar patches,[5] which are expected to mediate nonspecific interactions and influence colloidal stability. Such patches are a key component of the statistical aggregation propensity (SAP) method that has been used to redesign antibodies for improved stability.[6,7] Whereas non-polar patch analysis must be based upon a 3D structure or model, a popular approach is to look at the propensity of protein sequence stretches to form amyloid.[8] Several online tools are available to predict the amyloid propensity, which is related to the likelihood of β-structure formation, e.g., PASTA,[9] TANGO,[10] and Zyggregator.[11] It is assumed that partial unfolding will expose such sequences to potential protein−protein interactions, mediated by β-strands. This approach is therefore linked to the conformational stability of a folded protein, and irreversible aggregation. Returning to the colloidal stability properties probed by 3D-based patch analysis, parallel

Received: August 10, 2013
Revised: November 14, 2013
Accepted: November 20, 2013
Published: November 20, 2013

to the reduction of non-polar surface area is the introduction of surface charge. Several reports indicate that negative charge is preferred over positive charge for properties related to protein solubility, such as aggregation resistance,[12−16] consistent with recent work suggesting that protein solubility correlates with negative charge.[17] The detail of protein solubility modification by charge is likely to be a case- and condition-dependent combination of factors that include reduction of non-polar regions,[18] overall net charge repulsion,[19,20] and attraction from positive and negative regions in an anisotropic charge distribution.[21,22] It is clear from the engineering of overcharged proteins,[23,24] and from a tradition of handling proteins away from their pI's to prevent aggregation, that net charge can be a major factor in determining solubility. It is believed that at least a part of the effect of supercharging lies in preventing aggregation of partially unfolded states,[24] analogous to the avoidance of β-forming sequence regions.

This report focuses on 3D and sequence-based charge properties and their correlation with available data for protein solubility, and mRNA and protein abundances. Using a data set for *Escherichia coli* protein solubility in a cell-free expression system,[25] the structural feature correlating best with solubility was a lack of large positively charged surface patches, where the difference in positive patch signatures for the separation of soluble and insoluble data sets was similar to that seen for DNA binding versus non-DNA-binding proteins (Chan, Curtis, and Warwicker, in press). As a result, it was suggested that interactions between expressed proteins and nucleic acid (mRNAs, tRNAs) may lead to insolubility, as intermediates in an unknown mechanism. Consistent with this hypothesis, there was no equivalent separation for negatively charged patches. Although the observations may be specific for solubility in expression systems, and not necessarily relevant for concentrated protein solutions low in nucleic acid, the difference between positive and negative charge is intriguing. It leads to the question of whether amino acid side chains normally positively charged at physiological pH (Lys and Arg) contribute equally to the observed correlation with solubility. A limited analysis indicated that they do not, with Lys being favored over Arg for high solubility (Chan, Curtis, and Warwicker, in press). The current study explores this simple observation, which is based on sequence properties. If Lys is favored over Arg for proteins that have evolved to function at higher concentration, it might be expected that protein (or mRNA) levels show a correlation with Lys to Arg ratio. Additionally, extracellular proteins at high abundance would have a high Lys to Arg ratio, and paralogues of high abundance proteins, themselves at lower concentration, a lower Lys to Arg ratio. These questions are addressed, alongside structural features where 3D information is available. Results bear on whether the Lys to Arg ratio could be a useful tool in protein design for solubility, including in biopharmaceutical applications.

METHODS AND DATA SETS

Structural Data Sets. Various sets of protein structures were obtained from the Protein Data Bank (PDB).[26] Complete antibody structures are sparse in the PDB, but many structures of Fab fragments are present. A search in the PDB for structures containing the term Fab was followed by removal of entries with 100% sequence identity (i.e., precise copies). Since the structures of many Fab fragments are obtained in combination with an antigen, it was necessary to remove just the antibody component. Rather than analyze 670 coordinate sets with molecular graphics, only H (heavy) and L (light) chains were selected, followed by inspection for chains of the expected size and a coordinate file header that contains reference to an antibody Fab fragment, resulting in 408 HL chain combinations. While this procedure will have missed some genuine Fab fragments with nonstandard chain nomenclature, it maintains the bulk of the data set, and has the advantage of identifying a biological unit (a single Fab fragment) from what may be a more extensive entry in the PDB (e.g., crystal asymmetric unit). A text search at the PDB for either scFv or single chain antibody, with a filter at 100% sequence identity gives 36 entries. With the small data set (in comparison with Fab's) and the introduction of a linker peptide preventing the general use of the HL chain nomenclature, a graphics screen was employed to ascertain the chains associated with genuine scFv structures. This resulted in a data set of 24 scFv's. Structure-based calculations with both Fab's and scFv's covered the biological units, with combinations of protein chains where appropriate.

In order to compare computed structural features for therapeutic proteins with a background set of human proteins, a search for human proteins in the PDB was made, followed by a filtering with the PISCES[27] tool for crystal structures with sequence identity less than 30%, and length within 40−10000 amino acids. In this case, the data set is too large to reliably screen for biological units, and calculations are made on single protein chains (2073). A data set of structures for therapeutic proteins that have reached the market was prepared from searches in the PDB for entries corresponding to a recent review of biologics.[28] One coordinate entry was used for each biologic, and the analysis was again based on chains (62) rather than biological units, maintaining consistency with the background set of human protein structures. For estimation of the concentration of a marketed (or previously marketed) biologic at the point of delivery, the DailyMed resource (http://dailymed.nlm.nih.gov) was searched for preparation and delivery guidelines for each therapeutic protein. Where possible, the mass of protein delivered in a given volume of fluid was computed, with reference to the wider literature as appropriate to identify a relationship between units of activity and protein mass. The correlation between estimated delivery concentration and ratio of Lys content to Arg content (KR-ratio) was calculated for the resultant 46 protein chain set of biologics, and the 30 protein chains that exclude Fab fragments of therapeutic mAb's.

Sequence Data Sets. With the focus on sequence-based KR-ratio, there is no need to restrict analysis to those proteins that can be annotated with 3D structure, when comparing with solubility data for cell-free expression of *E. coli* proteins.[25] This gives 2931 data points overall from the experimental studies of solubility[25,29] that can be cross-referenced with KR-ratio values for *E. coli* K-12 ORFs obtained from the UCSC genome browser (http://genome.ucsc.edu).[30] Subsets from the experimental data are formed for low solubility (<30%) and high solubility (>70%) proteins, which are then compared. The Pearson correlation coefficient (r), a test for linear correlation between two properties, is used along with p-value calculation for the given number of data points. Where subsets are compared for the likelihood that they could arise from the same underlying distribution by chance, the Mann−Whitney test is used.

Expression data for *E. coli* K-12 were obtained from the relative abundance of mRNAs (log units to base 2) for growth in LB medium.[31] The mRNA abundance data were cross-referenced to KR-ratio, yielding matches for 1561 proteins. Yeast mRNA abundance data were taken from a study of *Saccharomyces cerevisiae* grown in steady state culture,[32] and transformed to log (base 2) units. Cross-referencing to a set of yeast proteins obtained from the *Saccharomyces* Genome Database (http://www.yeastgenome.org)[33] gave 5455 proteins with matched KR-ratio and mRNA abundance. Human mRNA levels were used for a set of housekeeping genes, with expression averaged across tissues,[34] giving mRNA abundance (log base 2) and KR-ratio for 2007 proteins. A single protein abundance set was also included, for a human blastoma cell line,[35] with protein levels cross-referenced to KR-ratio for 749 proteins. For the analysis of expression data, a lower limit of 100 amino acids was placed on proteins to prevent particularly large or small KR-ratio values arising simply from the low numbers of Lys and Arg sites.

Calculations. Structure-based properties, the maximum positive contoured patch size (in grid points) and the maximum ratio of non-polar solvent accessible surface area (SASA) to polar SASA for a patch, were derived for PDB subsets. A finite difference Poisson−Boltzmann method[36] was used to calculate electrostatic potential. Charges are assigned in a protein−solvent−counterion system in a continuum electrostatics framework (the Poisson−Boltzmann equation), that is inscribed on a Cartesian grid and solved (for electrostatic potential) using the numerical method of finite differences. Negatively charged Asp and Glu side chains and C-termini and positively charged Arg and Lys side chains and N-termini were included. Potential was contoured at $+kT/e$ on a shell around the protein (k is the Boltzmann constant, T = 300 K, and e is the electronic charge), with an ionic strength of 0.15 M. The size of the largest contoured positive patch is taken for each protein. The ratio of non-polar SASA to polar SASA is calculated for patches drawn out by 13 Å spheres centered on all non-hydrogen atoms,[5] with the highest such ratio stored for a protein. A 1.4 Å radius solvent probe was used to generate the solvent accessible surfaces. KR-ratio is simply the ratio of the number of Lys residues to the number of Arg residues in a given protein sequence, and could therefore be obtained for all proteins, irrespective of 3D structure availability. By analogy, DE-ratio is calculated from the relative occurrence of Asp and Glu residues. In order to identify orthologues and paralogues in the albumin and myoglobin families, BLAST[37] searches were made, seeded with the human protein and filtered for the subfamily in question (e.g., myoglobin, cytoglobin, and neuroglobin for the myoglobin family). Sequences were then retrieved and analyzed for KR-ratio. Comparison is made between the distribution of KR-ratio values found within a set of homologues and that expected if Lys and Arg were randomly populated for a given number of Lys + Arg sites (n, derived from the human variant), and an overall probability of Lys at each Lys or Arg site (p, equal to 1 − probability of Arg) taken from the occurrence of Lys and Arg in the human proteome. An expected random distribution of KR-ratio is then calculated from the corresponding (p, n) binomial. An equivalent analysis was made for DE-ratio. While the evolution of sequences will not necessarily lead to population around a random distribution, comparison of KR-ratio and DE-ratio may be indicative of selection pressure.

Threshold values for the three calculated properties—maximum positive patch size, maximum ratio of non-polar to polar SASA in a patch, and KR-ratio (in a sequence)—were derived in previous work (Chan, Curtis, and Warwicker, in press) as those values that best separate the most and least soluble subsets for cell-free expression of *E. coli* proteins,[25] when these subsets are plotted as cumulative distributions. For the cell-free expression data, the thresholds obtained in this way are 3000 grid points for maximum positive patch size, 4.5 for the ratio of non-polar to polar SASA, and 1.2 for the ratio of Lys to Arg in a sequence (Chan, Curtis, and Warwicker, in press). The parameters used to construct electrostatic potential grids are fixed, so that 3000 grid points always represent the same surface area, from protein to protein. With a grid step of 0.6 Å, a two-dimensional grid element is 0.36 Å^2 and 3000 grid points relates to roughly 1000 Å^2 surface area.

For all mRNA expression and protein level data sets, subsets were formed for all gene products with KR-ratio < 0.5 and for those with KR-ratio > 2.0. Limiting values of KR-ratio at 0.5 and 2.0 were chosen on the basis of their separation on either side of the threshold KR-ratio of 1.2. In practice, each subset defined in this way contains between 10 and 20% of the total number of proteins in a data set, depending on the data set. Abundances were then plotted as cumulative percentages of proteins, against expression, to examine whether the lower and upper KR-ratio subsets follow the same distributions.

RESULTS

Distribution of Charge and Non-Polar Features for Fab and scFv Structures. Three properties were computed for Fab and scFv sets from the PDB (Figure 1). These properties were chosen because they each show some separation of the most and least soluble subsets for cell-free expression of *E. coli* proteins,[25] with the relevant thresholds (Chan, Curtis, and Warwicker, in press) shown in Figure 1. For the structure-based positive patch size (Figure 1A), neither Fab's nor scFv's show a clear preference to lie on either side of the threshold. Fab's are representative of antibodies circulating in blood at high protein concentrations, along with serum albumin.[38] It is not clear that maximum positive patch size is relevant for solubility of proteins (Fab fragments of antibodies) at high circulatory concentration. Otherwise, evolutionary pressure would have pushed Fab fragments more toward the region lower than the threshold in Figure 1A. This argument will apply less to scFv's, which are excised from Fab's and re-engineered.[39] The maximum ratio of non-polar SASA to polar SASA lies largely under the threshold for both Fab's and scFv's (Figure 1B), consistent with a role for non-polar patches in protein insolubility, whether in expression or at high concentration of secreted protein. Interestingly, the sequence-based KR-ratio is largely above the threshold for Fab's but not for scFv's (Figure 1C). Again, this might indicate an evolutionary pressure on antibody sequences (and Fab fragments) to maintain a relatively high ratio of Lys to Arg content. Subsequent sections study KR-ratio in more detail.

KR-Ratio and Solubility in Cell-Free Expression. For a sequence-based property, it is possible to extend calculations for the cell-free expression data set beyond proteins that can be annotated with 3D structure to all proteins that can be cross-referenced. Solubility and KR-ratio for 2931 *E. coli* proteins correlate with r = 0.22, $p < 1 \times 10^{-8}$. The least and most soluble groups are clearly separated ($p = 1.98 \times 10^{-35}$), where higher KR-ratio associates with higher solubility (Figure 2). Of

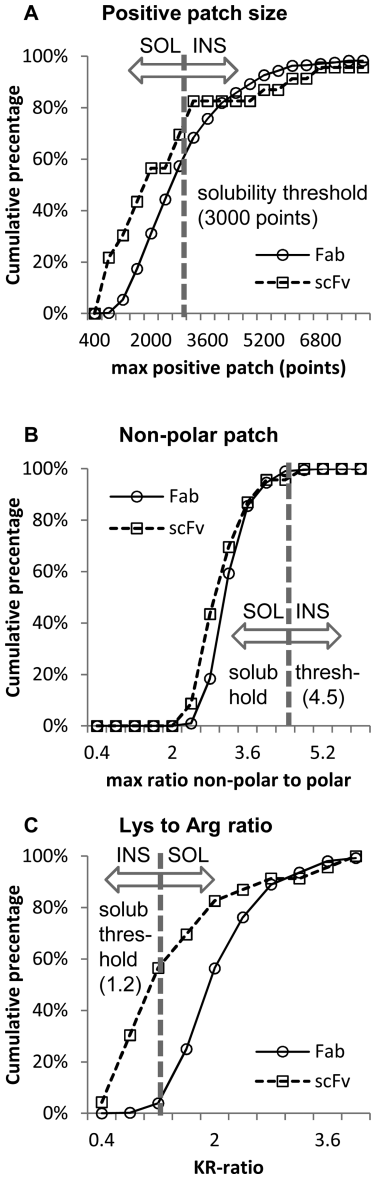


Figure 1. Calculated features plotted as cumulative percentages for Fab and scFv sets, compared with thresholds that best separate insoluble and soluble *E. coli* proteins in cell-free expression.[25] The direction of predicted solubility and insolubility, relative to threshold, is drawn for each panel. (A) Maximum positive potential patch size, measured in the number of patch points (threshold is 3000 points, more soluble below threshold). (B) Maximum ratio of non-polar to polar SASA for a patch (threshold ratio is 4.5, more soluble below threshold). (C) Ratio of Lys to Arg content (KR-ratio) for each protein (threshold is 1.2, more soluble above threshold).

the two systems, cell-free expression and Fab's (representing proteins that circulate at relatively high concentration), some properties may relate to solubility in both. This appears to be the case for KR-ratio and the maximum of the ratio of non-polar to polar surface area, whereas the maximum positive patch appears to be more relevant to cell-free expression. Studying KR-ratio, a sequence-based property, allows proteome-wide coverage of protein levels.

KR-Ratio and mRNA Levels for *E. coli*, Yeast, and Human Proteins. Transcriptome data and mRNA levels have provided the most widespread indication of protein expression because they have been more readily available than the

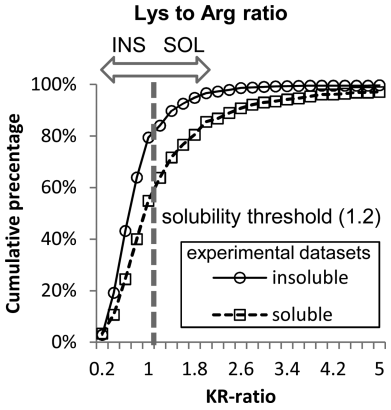


Figure 2. Separation of KR-ratio for soluble and insoluble subsets of proteins, from cell-free expression of *E. coli* proteins.[25] Predicted soluble and insoluble regions are indicated, relative to the threshold value (1.2).

corresponding proteomic data. While this situation is changing,[40] and the correlation between mRNA and protein levels can be poor,[41] bioinformatics studies still tend to use mRNA abundances as a proxy for protein levels. This approach is taken for expression data in *E. coli*, yeast, and human (Figure 3A,B,C), but additionally with proteome data added for human blastoma cells (Figure 3D). In all cases, added complications of cell-specific conditions exist, which are multiplied as species developmental complexity increases. Correlation coefficients for KR-ratio and (log) mRNA abundances are $r = 0.175$, $n = 1561$, and $p < 10^{-8}$ for *E. coli*; $r = 0.081$, $n = 5455$, and $p < 10^{-8}$ for yeast; and $r = 0.145$, $n = 2007$, and $p < 10^{-8}$ for human. A second approach is to form subsets of proteins, for each organism, with KR-ratio < 0.5 and KR-ratio > 2.0, and then compare expression levels of these subsets. Cumulative plots of mRNA levels are distinguishable for *E. coli* ($p = 8.78 \times 10^{-16}$), yeast ($p = 1.36 \times 10^{-5}$), and human ($p = 4.95 \times 10^{-4}$) cases. When human protein levels (from blastoma cells) are used in place of the log (mRNA) values, correlation with KR-ratio gives $r = 0.167$ and $p = 4.33 \times 10^{-6}$ for 749 data pairs. Lower and higher KR-ratio subsets for protein levels are separated ($p = 0.015$). These results are consistent with a selection pressure for lower Arg content relative to Lys in more highly expressed proteins.

KR-Ratio in Abundant Human Proteins and Their Less Abundant Paralogues. Human serum albumin (HSA) and antibodies are abundant circulating proteins.[38] Fab fragments, representing the bulk of structural information available for antibodies, have been analyzed already in this study. HSA has lower abundance paralogue families afamin, vitamin D-binding protein (DBP), and α-fetoprotein. While there are many myoglobin structures available, most of these relate to variants and mutants, and analysis is restricted to the sequence level. Two myoglobin paralogues, cytoglobin and neuroglobin, were found only within the genomic era,[42] since they are present at much lower concentrations than myoglobin. Sequences for HSA, myoglobin, and their paralogues were obtained from searching the human variants against the RefSeq database (www.ncbi.nlm.nih.gov/refseq/) with BLAST,[37] and subsequent inspection to collate only the relevant subfamily. In order to estimate a random expectation for occurrence of both KR-ratio and DE-ratio in serum albumins and myoglobin, a binomial expansion was used in each case. This gives the KR-ratio and DE-ratio distributions expected if the KR and DE sites

dx.doi.org/10.1021/mp4004749 | *Mol. Pharmaceutics* 2014, 11, 294−303

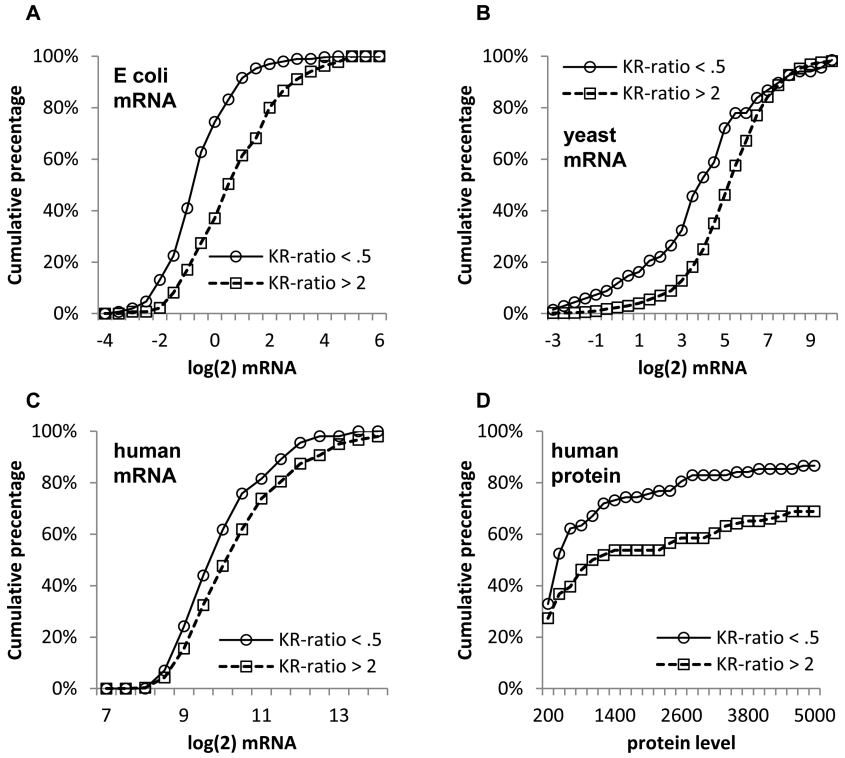


Figure 3. Protein and mRNA levels are more abundant at higher KR-ratio. Subsets of proteins are formed for KR-ratio < 0.5 and KR-ratio > 2.0, and the cumulative percentage of these subsets plotted as the expression level increased. Log (to base 2) of mRNA expression is used. (A) mRNA levels in *E. coli*.[31] (B) Yeast mRNA.[32] (C) mRNA abundance for protein-coding genes that are expressed across a series of tissues.[34] (D) Protein levels in human blastoma cells.[35]

of HSAs and myoglobins are populated at random, given the overall (proteome) KR-ratio and DE-ratio values.

Whereas the observed KR-ratio distribution for serum albumins is above that expected by chance (Figure 4A), the two distributions for DE-ratio lie close to each other (Figure 4A,B). Distributions of KR-ratio for the serum albumin paralogues (Figure 4C) are generally below that for serum albumin itself, consistent with their circulation at lower levels, estimated at 60 μg/mL for afamin,[43] 3 mg/mL for α-fetoprotein,[44] and 0.3 mg/mL for DBP.[45] DBP is interesting, since it is bimodal in KR-ratio, straddling the distribution for serum albumin.

KR-ratio and DE-ratio distributions for myoglobin follow the same general behavior as that for serum albumins (Figure 5A,B) but now with observed KR-ratio far exceeding random and DE-ratio roughly in line with that expected from a random selection of D and E residues. These results are consistent with evolutionary pressure on Lys relative to Arg content for myoglobins but not for Asp relative to Glu. Both cytoglobin and neuroglobin are present at concentrations <0.1 mg/mL,[46,47] and their KR-ratio distributions are far lower than that for myoglobin (Figure 5A,C).

KR-Ratio and Other Properties Compared between Biologics and Background Sets of Human Proteins. We are interested in comparing the properties studied here between human proteins in general and proteins that have been marketed as therapeutics. In order to analyze the structure-based features, as well as sequence-based, a set of 2073 protein chains from the PDB, non-redundant at 30% sequence identity was used. The data set for marketed biologics contained 62 protein chains. There was no significant separation for these two structural sets, in terms of the maximum positive charge patch and maximum ratio of non-polar to polar SASA. Of 62 biologic protein chains, 20 lie above the 3000 points threshold for a positive patch, but as seen earlier, it is not clear that this property has a general correlation with solubility. For a non-polar patch, which may be a more general correlate with solubility, just 8 of the 62 biologic chains lie above the threshold of 4.5.

Figure 6 compares KR-ratio (Figure 6A) and DE-ratio (Figure 6B) for the biologics and human PDB structural sets, in addition to the set of human ORFs. Distributions are similar for DE-ratio and more varied for KR-ratio. Variation in the biologics plot toward higher KR-ratio is largely due to the Fab fragments. There will have been less evolutionary pressure toward higher solubility for biologics that occur naturally at low concentration, unlike antibodies and serum albumins. It is therefore important to identify features that could be relevant for solubility engineering of biologics (and recombinant protein production in general). There are several biologic protein chains that have a KR-ratio of <1 (Figure 6A). The current study indicates that such a low KR-ratio could contribute to reduced solubility. If this were the case, then it might be expected that biologic concentration at the point of delivery (a value that can be estimated from available data) would show some correlation with KR-ratio. Clearly, there will be several factors in the complex considerations of design and delivery schemes for protein therapeutics, but solubility is one such factor.[48,49] For 46 protein chains, including Fab fragments of mAb's, testing the hypothesis that delivery concentration is positively correlated with KR-ratio gives $r = 0.158$ ($p = 0.147$), i.e., insignificant at the 5% level. These values when Fab's are removed, leaving 30 protein chains, are $r = 0.411$ and $p = 0.012$, i.e., significant. The Fab's have relatively high KR-ratios, and

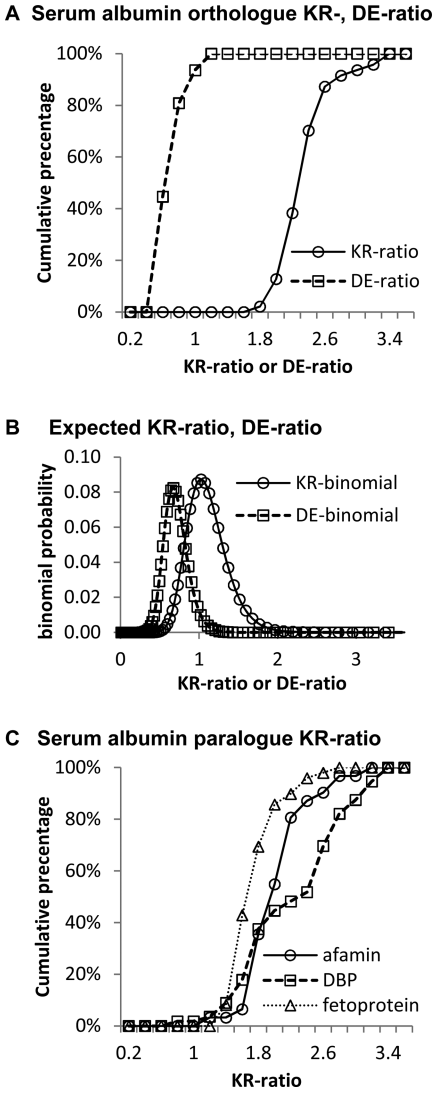


Figure 4. KR-ratio and DE-ratio for serum albumins and paralogues, compared with those expected from random sampling. (A) Cumulative percentage plots of KR-ratio and DE-ratio for 47 serum albumins. (B) Probability distributions plotted against KR-ratio and DE-ratio, for binomial distributions given the number of Lys + Arg (83) and Asp + Glu (97) sites in HSA, and the probability of Lys or Arg calculated from occurrence over the whole human proteome. (C) Serum albumin paralogue KR-ratio distributions for 31 afamins, 56 vitamin D-binding proteins, and 49 α-fetoproteins.

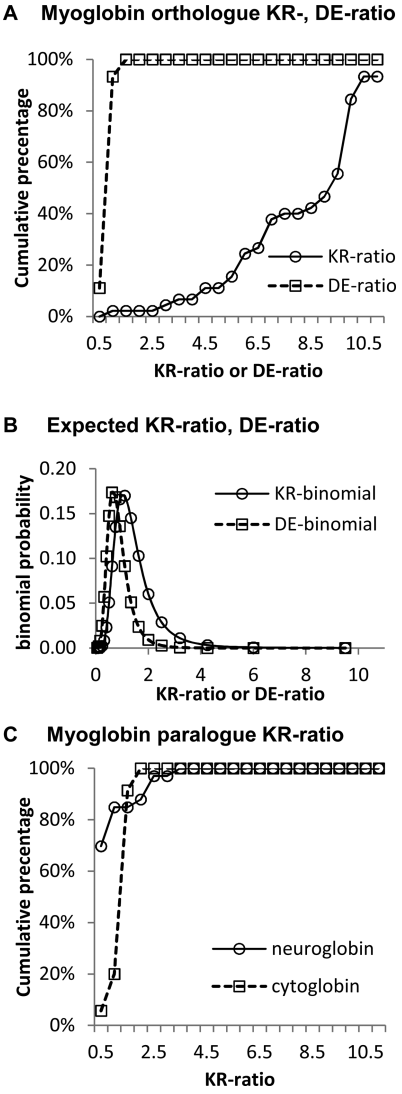


Figure 5. KR-ratio and DE-ratio for myoglobins and paralogues, compared with those expected from random sampling. (A) Cumulative percentage plots of KR-ratio and DE-ratio for 45 myoglobins. (B) Probability distributions plotted against KR-ratio and DE-ratio, for binomial distributions given the number of Lys + Arg (21) and Asp + Glu (21) sites in myoglobin, and the probability of Lys or Arg calculated from occurrence over the whole human proteome. (C) Myoglobin paralogue KR-ratio distributions for 33 neuroglobins and 35 cytoglobins.

their corresponding mAb's tend to be delivered at relatively high concentration. Since there exists a significant correlation for the generally lower delivery concentrations, when mAb's are excluded, there is scope to investigate whether engineering KR-ratio to higher values could contribute to future biologic design.

DISCUSSION

The finding that largest positive patch could be added to non-polar surface in the list of structure-based features that correlate with insolubility (Chan, Curtis, and Warwicker, in press) in a cell-free expression system[25] led to the question of whether these relationships were more ubiquitous. This could be investigated with threshold values obtained from analysis of low and high solubility subsets of the cell-free expression data. With many structural representatives of antibody Fab fragments in the PDB, and since antibodies typically circulate at high concentration, these presented a good case with which to examine protein solubility in a physiological environment, with the application of thresholds. While non-polar surface area does transfer to Fab fragments, the positive charge surface features do not (Figure 1). Of interest in the current work is the emergence of KR-ratio as a possible novel correlate to protein solubility in a physiological environment (Figure 1C). A sequence-based property offers more scope (than a structure-based feature) for investigation against protein expression data. A data set of scFv structures was included alongside Fab's for comparison. Here, the only feature which is generally toward the low solubility side of the relevant threshold is non-polar surface area. This appears to be a general property that should be considered when designing for solubility of natively structured proteins. If KR-ratio also turns out to be ubiquitous

dx.doi.org/10.1021/mp4004749 | *Mol. Pharmaceutics* 2014, 11, 294−303

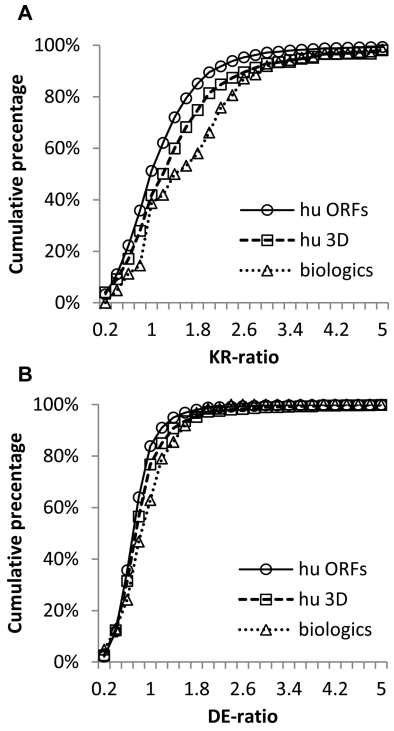


Figure 6. KR-ratio and DE-ratio for biologics compared with human ORFs and a set of human protein structures (3D). (A) Cumulative percentage distributions for KR-ratio. (B) DE-ratio.

for solubility, then several members of the scFv set could be improved in this respect (Figure 1C).

Comparison of KR-ratio against mRNA levels for *E. coli*, yeast, and humans shows a correlation, and separation of lower and upper KR-ratio subsets, in each case (Figure 3). Messenger RNAs for proteins with higher KR-ratio are, on average, present at higher levels, consistent with a selection pressure for greater lysine content over arginine, when expression is higher. Although correlation between mRNA levels and protein copy numbers is imprecise, the relationship with KR-ratio is consistent over varied species. Moreover, analysis of proteins expressed across many tissues in humans exhibits the same sense of correlation between gene expression and KR-ratio (Figure 3D). As more quantitative proteomic data become available, further examination may reveal other sequence (or structure) features that consistently correlate with protein expression level, revealing selection pressures.

KR-ratio has been compared with expected values for random KR assignment given the proteome K, R content, and with equivalent analysis for DE-ratio. In the case of serum albumin, itself a therapeutic protein, and myoglobin, while DE-ratios are in line with the expected distributions across species, KR-ratios are almost uniformly high (Figures 4 and 5). A fruitful area to study is comparison of proteins known to be present at high concentration with paralogues that have lower physiological concentrations. This approach can be conveniently applied to serum albumin and myoglobin, with the lower concentration paralogues generally at lower KR-ratio (very substantially so for cytoglobin and neuroglobin in comparison to myoglobin). The one exception is a bimodal KR-ratio distribution for vitamin D-binding protein. It is not clear what might lie behind this bimodality. The two KR-ratio regions for DBP are largely separated by species phylogeny (not shown). One complication that arises with DBP is the balance between the normal circulating vitamin D transport role and transport of vitamin D from the epidermis[50] in those species for which synthesis in the skin is a major contributor to the vitamin D pool. Possibly the bimodal KR-ratio distribution for DBP reflects this distinction. A recent report shows that the net charge of myoglobin has evolved to higher values in animals with greater diving ability, suggesting that prevention of aggregation at higher myoglobin concentrations underlies this observation.[51] No investigation of Arg, Lys composition was reported. It will be interesting to examine how KR-ratio, which is already very high for extant myoglobins (Figure 5A), varies for these myoglobins, including inferred ancestral sequences.[51]

For the charge-based properties presented in this work, it is important to consider possible confounding factors. The correlation seen between maximal positively charged patch and solubility in cell-free expression (Chan, Curtis, and Warwicker, in press) does not account for how different counterions and their binding to proteins could influence solubility. Calculations of positive patches are made with a simple model for 0.15 M monovalent counterions, and do not incorporate specific ion binding. This remains an area to develop computationally, and could be particularly important for cases such as biologics where formulation and solution conditions are highly variable. For KR-ratio, identification of proteins that are soluble at high concentration but with low KR-ratio will help to refine the model. One example is the subfamily of eye lens proteins, γ-crystallins, with KR-ratio as low as 0.048 for human γD-crystallin. Crystallins maintain solubility at high concentration and without protein turnover, and have been extensively studied.[52] An interesting observation is that, in the human γD-crystallin cystal structure, Arg side chains are involved in extensive charge networks with acidic side chains,[53] presumably reinforcing the folded state stability. It may be that selection for Arg over Lys in γ-crystallins increases intramolecular interactions, and reduces Arg influence on intermolecular interactions through sequestering the side chains into the intramolecular charge networks. Clearly, there is more to understand about the relative contributions of different basic (and acidic) side chains to stability and solubility.

The analysis to this point suggests that, when assessed over large data sets, KR-ratio may be a correlate of physiological protein concentration, that has hitherto gone unnoticed. Differences between Arg and Lys side chain interactions have been reported. Arginine is known to be over-represented in functional protein−protein and protein−nucleic acid interfaces.[4] A preference of Arg over Lys for (cation−pi) interactions with aromatic groups has been observed.[54,55] It has been found that Lys to Arg mutations can increase crystallization propensity, thought to be through differential conformational mobility.[56] Arginine can be an important component of additive solutions that stabilize against protein aggregation.[57] It will be important to establish the molecular basis of interactions and the extent to which the mechanisms by which Arg stabilizes as an additive[58,59] has features in common with Arg in protein−protein interfaces, both specific and nonspecific. An emerging result that requires synthesis into a design framework is the utility of negative charge in improving solubility.[17] Several studies have reported supercharging of proteins (either positively or negatively), with consequent improvement of solubility. It has been reported that prevention of aggregation from partially unfolded states contributes to the effect of supercharging.[24] Our work suggests that, in terms of positive charge, Lys should be more effective than Arg in

dx.doi.org/10.1021/mp4004749 | *Mol. Pharmaceutics* 2014, 11, 294−303

promoting solubility by supercharging. To compare charge roles in solubility, it should be possible to simply swap charges around (negative to positive, Lys to Arg).

A correlation between KR-ratio and solubility will have implications for protein expression in biotechnology, and for modulation of therapeutic protein properties. Non-mAb biologics show a correlation between KR-ratio and estimated concentration at the point of delivery. Monoclonal antibodies are generally soluble at relatively high concentration (estimated at between 2 and 150 mg/mL for delivery of the therapeutic mAb's studied here), and the lack of a correlation for this subset of biologics may indicate that solubility is not a limiting factor. Figure 6 indicates that several (non-mAb) biologics are suboptimal in terms of the correlations observed in the current work between KR-ratio and solubility. Other than a few examples such as serum albumin and mAb's, proteins that have crucial functions for therapeutic use will not necessarily have evolved characteristics suitable for maintaining solubility at high concentrations. There may be cases where solubility has been a limiting factor in biologic development, and where alteration of KR-ratio could contribute to overcoming such problems, particularly if it is incorporated early in the design process. Additionally, there may be scope later in development to counter a low KR-ratio through formulation. If a correlation between KR-ratio and solubility is confirmed experimentally, then a relatively simple design tool could be developed (e.g., a subset of nonessential Arg residues substituted with Lys). KR-ratio is a sequence-based property, without reference to 3D arrangement of amino acids. It may therefore parallel the study of sequence regions with amyloid propensity, and their proposed role in promoting aggregation from partially unfolded states.

■ CONCLUSION

Protein solubility is a key property for biotechnological and biopharmaceutical applications, where a protein may be required to be soluble when removed from its physiological environment. Some of the properties associated with solubility such as the avoidance of non-polar interactions and the importance of charged groups and net charge are well-known. Information is emerging concerning the relative roles of positive and negative charge, and now in the current work the positively charged side chains of Lys and Arg. Analysis of experimental solubility data, of mRNA and protein levels, as well as a study of paralogue families expressed at different physiological concentrations, suggests that higher Lys (and lower Arg) content correlates with protein solubility. Since many proteins of biotechnological and biopharmaceutical interest will not be expressed at high concentration naturally, we conclude that scope exists for simple adjustment of Arg and Lys content to enhance protein production and solubility. An experimental program in this direction will establish whether the correlation is general and, if it is, the underlying molecular details.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +44 (0)161 306 4490. Fax: +44 (0)161 275 5082. E-mail: jim.warwicker@manchester.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.C. is supported by PhD funding from the EPSRC Centre for Doctoral Training in Innovative Manufacturing in Emergent Macromolecular Therapies, centred at UCL. The authors thank Rose Keeling for discussion.

■ REFERENCES

(1) Esposito, D.; Chatterjee, D. K. Enhancement of soluble protein expression through the use of fusion tags. *Curr. Opin. Biotechnol.* **2006**, *17*, 353−358.

(2) den Engelsman, J.; Garidel, P.; Smulders, R.; Koll, H.; Smith, B.; Bassarab, S.; Seidl, A.; Hainzl, O.; Jiskoot, W. Strategies for the assessment of protein aggregates in pharmaceutical biotech product development. *Pharm. Res.* **2011**, *28*, 920−933.

(3) Kauzmann, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **1959**, *14*, 1−63.

(4) Jones, S.; Marin, A.; Thornton, J. M. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **2000**, *13*, 77−82.

(5) Cole, C.; Warwicker, J. Side-chain conformational entropy at protein-protein interfaces. *Protein Sci.* **2002**, *11*, 2860−2870.

(6) Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. L. Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 11937−11942.

(7) Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. L. Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B* **2010**, *114*, 6614−6624.

(8) Chiti, F.; Dobson, C. M. Amyloid formation by globular proteins under native conditions. *Nat. Chem. Biol.* **2009**, *5*, 15−22.

(9) Trovato, A.; Seno, F.; Tosatto, S. C. The PASTA server for protein aggregation prediction. *Protein Eng., Des. Sel.* **2007**, *20*, 521−523.

(10) Linding, R.; Schymkowitz, J.; Rousseau, F.; Diella, F.; Serrano, L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **2004**, *342*, 345−353.

(11) Tartaglia, G. G.; Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **2008**, *37*, 1395−1401.

(12) Arbabi-Ghahroudi, M.; To, R.; Gaudette, N.; Hirama, T.; Ding, W.; MacKenzie, R.; Tanha, J. Aggregation-resistant VHs selected by in vitro evolution tend to have disulfide-bonded loops and acidic isoelectric points. *Protein Eng., Des. Sel.* **2009**, *22*, 59−66.

(13) Dudgeon, K.; Rouet, R.; Kokmeijer, I.; Schofield, P.; Stolp, J.; Langley, D.; Stock, D.; Christ, D. General strategy for the generation of human antibody variable domains with increased aggregation resistance. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 10879−10884.

(14) Kvam, E.; Sierks, M. R.; Shoemaker, C. B.; Messer, A. Physico-chemical determinants of soluble intrabody expression in mammalian cell cytoplasm. *Protein Eng., Des. Sel.* **2010**, *23*, 489−498.

(15) Wayne, N.; Bolon, D. N. Charge-rich regions modulate the anti-aggregation activity of Hsp90. *J. Mol. Biol.* **2010**, *401*, 931−939.

(16) Perchiacca, J. M.; Bhattacharya, M.; Tessier, P. M. Mutational analysis of domain antibodies reveals aggregation hotspots within and near the complementarity determining regions. *Proteins* **2011**, *79*, 2637−2647.

(17) Kramer, R. M.; Shende, V. R.; Motl, N.; Pace, C. N.; Scholtz, J. M. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys. J.* **2012**, *102*, 1907−1915.

(18) Ho, J. G.; Middelberg, A. P. Estimating the potential refolding yield of recombinant proteins expressed as inclusion bodies. *Biotechnol. Bioeng.* **2004**, *87*, 584−592.

(19) Chi, E. Y.; Krishnan, S.; Kendrick, B. S.; Chang, B. S.; Carpenter, J. F.; Randolph, T. W. Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci.* **2003**, *12*, 903−913.

(20) Olsen, S. N.; Andersen, K. B.; Randolph, T. W.; Carpenter, J. F.; Westh, P. Role of electrostatic repulsion on colloidal stability of Bacillus halmapalus alpha-amylase. *Biochim. Biophys. Acta* **2009**, *1794*, 1058−1065.
(21) Yadav, S.; Laue, T. M.; Kalonia, D. S.; Singh, S. N.; Shire, S. J. The influence of charge distribution on self-association and viscosity behavior of monoclonal antibody solutions. *Mol. Pharmaceutics* **2012**, *9*, 791−802.
(22) Saluja, A.; Badkar, A. V.; Zeng, D. L.; Kalonia, D. S. Ultrasonic rheology of a monoclonal antibody (IgG2) solution: implications for physical stability of proteins in high concentration formulations. *J. Pharm. Sci.* **2007**, *96*, 3181−3195.
(23) Lawrence, M. S.; Phillips, K. J.; Liu, D. R. Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* **2007**, *129*, 10110−10112.
(24) Der, B. S.; Kluwe, C.; Miklos, A. E.; Jacak, R.; Lyskov, S.; Gray, J. J.; Georgiou, G.; Ellington, A. D.; Kuhlman, B. Alternative computational protocols for supercharging protein surfaces for reversible unfolding and retention of stability. *PLoS One* **2013**, *8*, e64363.
(25) Niwa, T.; Ying, B. W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; Taguchi, H. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4201−4206.
(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(27) Wang, G.; Dunbrack, R. L., Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **2005**, *33*, W94−98.
(28) Dimitrov, D. S. Therapeutic proteins. *Methods Mol. Biol.* **2012**, *899*, 1−26.
(29) Niwa, T.; Kanamori, T.; Ueda, T.; Taguchi, H. Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 8937−8942.
(30) Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*, 996−1006.
(31) Bernstein, J. A.; Khodursky, A. B.; Lin, P. H.; Lin-Chao, S.; Cohen, S. N. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 9697−9702.
(32) Lee, D.; Smallbone, K.; Dunn, W. B.; Murabito, E.; Winder, C. L.; Kell, D. B.; Mendes, P.; Swainston, N. Improving metabolic flux predictions using absolute gene expression data. *BMC Syst. Biol.* **2012**, *6*, 73.
(33) Cherry, J. M.; Hong, E. L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; Chan, E. T.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S. R.; Fisk, D. G.; Hirschman, J. E.; Hitz, B. C.; Karra, K.; Krieger, C. J.; Miyasato, S. R.; Nash, R. S.; Park, J.; Skrzypek, M. S.; Simison, M.; Weng, S.; Wong, E. D. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **2012**, *40*, D700−705.
(34) Chang, C. W.; Cheng, W. C.; Chen, C. R.; Shu, W. Y.; Tsai, M. L.; Huang, C. L.; Hsu, I. C. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One* **2011**, *6*, e22859.
(35) Vogel, C.; Abreu Rde, S.; Ko, D.; Le, S. Y.; Shapiro, B. A.; Burns, S. C.; Sandhu, D.; Boutz, D. R.; Marcotte, E. M.; Penalva, L. O. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **2010**, *6*, 400.
(36) Warwicker, J. Continuum dielectric modelling of the protein-solvent system, and calculation of the long-range electrostatic field of the enzyme phosphoglycerate mutase. *J. Theor. Biol.* **1986**, *121*, 199−210.
(37) Altschul, S. F.; Koonin, E. V. Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases. *Trends Biochem. Sci.* **1998**, *23*, 444−447.
(38) Andersen, J. T.; Sandlie, I. The versatile MHC class I-related FcRn protects IgG and albumin from degradation: implications for development of new diagnostics and therapeutics. *Drug Metab. Pharmacokinet.* **2009**, *24*, 318−332.
(39) Demarest, S. J.; Glaser, S. M. Antibody therapeutics, antibody engineering, and the merits of protein stability. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 675−687.
(40) Rodriguez-Suarez, E.; Whetton, A. D. The application of quantification techniques in proteomics for biomedical research. *Mass Spectrom. Rev.* **2013**, *32*, 1−26.
(41) Nie, L.; Wu, G.; Culley, D. E.; Scholten, J. C.; Zhang, W. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit. Rev. Biotechnol.* **2007**, *27*, 63−75.
(42) Pesce, A.; Bolognesi, M.; Bocedi, A.; Ascenzi, P.; Dewilde, S.; Moens, L.; Hankeln, T.; Burmester, T. Neuroglobin and cytoglobin. Fresh blood for the vertebrate globin family. *EMBO Rep.* **2002**, *3*, 1146−1151.
(43) Jerkovic, L.; Voegele, A. F.; Chwatal, S.; Kronenberg, F.; Radcliffe, C. M.; Wormald, M. R.; Lobentanz, E. M.; Ezeh, B.; Eller, P.; Dejori, N.; Dieplinger, B.; Lottspeich, F.; Sattler, W.; Uhr, M.; Mechtler, K.; Dwek, R. A.; Rudd, P. M.; Baier, G.; Dieplinger, H. Afamin is a novel human vitamin E-binding glycoprotein characterization and in vitro expression. *J. Proteome Res.* **2005**, *4*, 889−899.
(44) Olsson, M.; Lindahl, G.; Ruoslahti, E. Genetic control of alpha-fetoprotein synthesis in the mouse. *J. Exp. Med.* **1977**, *145*, 819−827.
(45) Lauridsen, A. L.; Vestergaard, P.; Nexo, E. Mean serum concentration of vitamin D-binding protein (Gc globulin) is related to the Gc phenotype in women. *Clin. Chem.* **2001**, *47*, 753−756.
(46) Fago, A.; Hundahl, C.; Malte, H.; Weber, R. E. Functional properties of neuroglobin and cytoglobin. Insights into the ancestral physiological roles of globins. *IUBMB Life* **2004**, *56*, 689−696.
(47) Burmester, T.; Hankeln, T. What is the function of neuroglobin? *J. Exp. Biol.* **2009**, *212*, 1423−1428.
(48) Caravella, J.; Lugovskoy, A. Design of next-generation protein therapeutics. *Curr. Opin. Chem. Biol.* **2010**, *14*, 520−528.
(49) Yamniuk, A. P.; Ditto, N.; Patel, M.; Dai, J.; Sejwal, P.; Stetsko, P.; Doyle, M. L. Application of a kosmotrope-based solubility assay to multiple protein therapeutic classes indicates broad use as a high-throughput screen for protein therapeutic aggregation propensity. *J. Pharm. Sci.* **2013**, *102*, 2424−2439.
(50) Haddad, J. G.; Matsuoka, L. Y.; Hollis, B. W.; Hu, Y. Z.; Wortsman, J. Human plasma transport of vitamin D after its endogenous synthesis. *J. Clin. Invest.* **1993**, *91*, 2552−2555.
(51) Mirceta, S.; Signore, A. V.; Burns, J. M.; Cossins, A. R.; Campbell, K. L.; Berenbrink, M. Evolution of mammalian diving capacity traced by myoglobin net surface charge. *Science* **2013**, *340*, 1234192.
(52) Bloemendal, H.; de Jong, W.; Jaenicke, R.; Lubsen, N. H.; Slingsby, C.; Tardieu, A. Ageing and vision: structure, stability and function of lens crystallins. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 407−485.
(53) Basak, A.; Bateman, O.; Slingsby, C.; Pande, A.; Asherie, N.; Ogun, O.; Benedek, G. B.; Pande, J. High-resolution X-ray crystal structures of human gammaD Crystallin (1.25 A) and the R58H mutant (1.15 A) associated with aculeiform cataract. *J. Mol. Biol.* **2003**, *328*, 1137−1147.
(54) Martis, R. L.; Singh, S. K.; Gromiha, M. M.; Santhosh, C. Role of cation-pi interactions in single chain 'all-alpha' proteins. *J. Theor. Biol.* **2008**, *250*, 655−662.
(55) Shah, D.; Li, J.; Shaikh, A. R.; Rajagopalan, R. Arginine-aromatic interactions and their effects on arginine-induced solubilization of aromatic solutes and suppression of protein aggregation. *Biotechnol. Prog.* **2012**, *28*, 223−231.
(56) Czepas, J.; Devedjiev, Y.; Krowarsch, D.; Derewenda, U.; Otlewski, J.; Derewenda, Z. S. The impact of Lys−>Arg surface mutations on the crystallization of the globular domain of RhoGDI. *Acta Crystallogr., Sect. D* **2004**, *60*, 275−280.
(57) Golovanov, A. P.; Hautbergue, G. M.; Wilson, S. A.; Lian, L. Y. A simple method for improving protein solubility and long-term stability. *J. Am. Chem. Soc.* **2004**, *126*, 8933−8939.

(58) Arakawa, T.; Ejima, D.; Tsumoto, K.; Obeyama, N.; Tanaka, Y.; Kita, Y.; Timasheff, S. N. Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects. *Biophys. Chem.* **2007**, *127*, 1−8.
(59) Shukla, D.; Trout, B. L. Understanding the synergistic effect of arginine and glutamic acid mixtures on protein solubility. *J. Phys. Chem. B* **2011**, *115*, 11831−11839.