Protein Engineering, Design & Selection, 2016, vol. 29 no. 6, pp. 231–243
doi: 10.1093/protein/gzw010
Advance Access Publication Date: 9 May 2016
Original Article

OXFORD

Original Article

Modulating non-native aggregation and electrostatic protein–protein interactions with computationally designed single-point mutations

C.J. O'Brien[1], M.A. Blanco[1], J.A. Costanzo[2], M. Enterline[1], E.J. Fernandez[2], A.S. Robinson[1,3,*], and C.J. Roberts[1,*]

[1]Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, DE 19716, USA, [2]Department of Chemical Engineering, University of Virginia, Charlottesville, VA 22903, USA, and [3]Department of Chemical and Biomolecular Engineering, Tulane University, New Orleans, LA 70118, USA

*To whom correspondence should be addressed. E-mail: asr@tulane.edu (A.S.R.); cjr@udel.edu (C.J.R.)
Edited by David Thirumalai

Received 16 September 2015; Revised 29 February 2016; Accepted 28 March 2016

Abstract

Non-native protein aggregation is a ubiquitous challenge in the production, storage and administration of protein-based biotherapeutics. This study focuses on altering electrostatic protein–protein interactions as a strategy to modulate aggregation propensity in terms of temperature-dependent aggregation rates, using single-charge variants of human γ-D crystallin. Molecular models were combined to predict amino acid substitutions that would modulate protein–protein interactions with minimal effects on conformational stability. Experimental protein–protein interactions were quantified by the Kirkwood–Buff integrals (G_{22}) from laser scattering, and G_{22} showed semi-quantitative agreement with model predictions. Experimental initial-rates for aggregation showed that increased (decreased) repulsive interactions led to significantly increased (decreased) aggregation resistance, even based solely on single-point mutations. However, in the case of a particular amino acid (E17), the aggregation mechanism was altered by substitution with R or K, and this greatly mitigated improvements in aggregation resistance. The results illustrate that predictions based on native protein–protein interactions can provide a useful design target for engineering aggregation resistance; however, this approach needs to be balanced with consideration of how mutations can impact aggregation mechanisms.

Key words: computational design, protein aggregation, protein engineering, protein–protein interactions

Introduction

Non-native protein aggregation refers to the net irreversible formation of higher molecular weight species from (partially) unfolded or misfolded protein monomers (Roberts, 2007, 2014; Weiss *et al.*, 2009). Aggregation is a long-standing problem in the field of biological therapeutics due to its prevalence in many aspects of biomolecule production, as well as immunogenicity concerns during patient administration (Wang, 2005; Jiskoot *et al.*, 2010; Wang *et al.*, 2010). Aggregation reduces native protein expression yields, leads to increased loss during purification of proteins, as well as decreases the shelf life of the final product (Weiss *et al.*, 2009; Brummitt *et al.*, 2011c).

Non-native aggregation—hereafter denoted simply as aggregation—is driven by the same molecular-scale forces that drive protein folding. These forces include steric interactions, van der Waals attractions between backbone and side chain atoms, hydrogen bonding, hydrophobic interactions and electrostatic repulsions and attractions; these forces are also combined with losses or gains of chain entropy (Chiti *et al.*, 1999, 2003; Dinner *et al.*, 2000; Kuhlman and Baker, 2000; Dobson, 2001, 2004; DuBay *et al.*, 2004; Roberts, 2014). Therefore, different

contributions must be balanced throughout production and storage to maintain folded-state stability, as well as disfavor aggregation (Wang, 2005; Wang *et al.*, 2010; Roberts, 2014). Interactions can be modulated by changing the solution conditions, pressure or temperature; or by altering the amino acid sequence of the protein (Wang, 2005; Weiss *et al.*, 2009; Wang *et al.*, 2010; van Beers and Bardor, 2012; Roberts, 2014). These interactions may also influence details of the aggregation mechanism(s) for a particular protein in a given solution condition, as well as the properties of the resulting aggregate species (Li *et al.*, 2009, 2010; Sahin *et al.*, 2010, 2012; Brummitt *et al.*, 2011b; Kim *et al.*, 2013).

A number of related aggregation pathways have been identified and characterized for various model proteins (Krebs *et al.*, 2000; Andrews and Roberts, 2007; Weiss *et al.*, 2007; Li *et al.*, 2009, 2010; Sahin *et al.*, 2010, 2012; Brummitt *et al.*, 2011a, b; Kim *et al.*, 2013). Common features of non-native aggregation include the following. Folded or partially unfolded monomers must come together to form reversible oligomers in solution, through native or non-native contacts, that are typically small in size. These oligomers then can undergo structural changes that stabilize them, resulting net irreversible aggregates that might be capable of growth (Roberts, 2007, 2014; Weiss *et al.*, 2009). The smallest net-irreversible aggregate species are denoted as the 'nuclei', particularly when subsequent aggregate growth is much faster than formation of new nuclei. Aggregate growth post-nucleation may involve addition of monomers and/or coalescence of existing aggregates (Weiss *et al.*, 2009; Brummitt *et al.*, 2011b; Kim *et al.*, 2013). Aggregates may also precipitate out of solution if they reach a solubility limit (Kroetsch *et al.*, 2012; Sahin *et al.*, 2012; Kim *et al.*, 2013).

Mechanistic strategies for engineering aggregation resistance are often targeted toward: (i) stabilization of the monomeric, native folded structure to make partial or larger-scale unfolding less favorable, and thereby reducing the concentration of aggregation-prone monomers (Famm *et al.*, 2008; Miller *et al.*, 2009, 2010; Sahin *et al.*, 2011; Costanzo *et al.*, 2014); (ii) identification and removal of aggregation-prone stretches of amino acid sequence that are predicted to be involved in strong contacts that stabilize aggregates (e.g. via β-sheet structures) (Chiti *et al.*, 2002a, b, 2003; de Groot *et al.*, 2006; Sahin *et al.*, 2011; Costanzo *et al.*, 2014); (iii) increasing repulsive protein–protein interactions so as to disfavor the initial stages of protein self-association (Perchiacca and Tessier, 2012; Perchiacca *et al.*, 2012). Experimental methods utilized to employ these approaches may include directed evolution and high throughput screening of libraries of variants to find those with the most desired properties (Proba *et al.*, 1998; Boder *et al.*, 2000; Wörn and Plückthun, 2001; Famm *et al.*, 2008; Miller *et al.*, 2009, 2010); knowledge-based approaches that use information from structures and sequences of similar proteins to imbue favorable traits (Benhar and Pastan, 1995; Ewert *et al.*, 2003a, b; Wang *et al.*, 2009; McConnell *et al.*, 2013); and applying physics- and chemistry-based approaches that rely on models of the underlying molecular interactions (Kuhlman and Baker, 2000; Guerois *et al.*, 2002; Chiti *et al.*, 2003; DuBay *et al.*, 2004; Fernandez-Escamilla *et al.*, 2004; Jordan *et al.*, 2009; Michaelson *et al.*, 2009; Sahin *et al.*, 2011).

Modulating electrostatic interactions in protein monomers has been shown to impart significant changes to the stability and aggregation propensity of proteins. 'Supercharging' a protein's surface by adding many similarly charged residues has been shown to potentially yield highly stable protein variants (Lawrence *et al.*, 2007), increase solubility (Simeonov *et al.*, 2011) and imbue significant aggregation resistance into proteins (Miklos *et al.*, 2012). However, creation of large patches of similar charges can also destabilize the fold or negatively impact function (Miklos *et al.*, 2012). Addition of charged amino acids flanking hydrophobic, aggregation-prone regions in antibody binding loops has been shown to impart increased resistance to aggregation, although selection of mutations requires care to avoid negatively influencing folding or function (Perchiacca *et al.*, 2012). Modulating dipoles and patches of similarly charged amino acids on the surface a protein has also been used to increase aggregation resistance through electrostatic repulsions (Brunsteiner *et al.*, 2013; Roberts and Blanco, 2014). A potential disadvantage of many of these approaches is that they are either specific to the protein class under consideration, and/or they require numerous substitutions in the amino acid sequence of the protein. The latter is more likely to negatively affect conformational stability, biological activity and immunogenicity of the folded protein (Miklos *et al.*, 2012; Raghunathan *et al.*, 2013).

The present work focuses on the strategy of altering electrostatic protein–protein interactions with minimal (i.e. single-point) mutations and minimal alterations to structural stability. Specifically, molecular simulations were used to predict the effects of altering natively charged residues on the surface of a protein one at a time, allowing a change in charge of plus or minus 1 or 2. Molecular simulations yielded predicted values of the second osmotic virial coefficient, B_{22}, which could also be measured directly to assess the accuracy of the predictions. Each possible variant was tested computationally to determine which substitutions were predicted to have the greatest effect on protein–protein interactions and average protein–protein center-to-center distance. RosettaDesign was used to screen promising candidates to avoid those predicted to result in significant changes in conformational stability (i.e. unfolding free energy). Variants that were predicted to provide changes in aggregation propensity with minimal changes in unfolding free energy were identified, and a subset was expressed, purified and experimentally characterized to test the predictions.

Human γ-D crystallin (γDc) is a 173 amino acid, ~20 kDa eye lens protein. It has been extensively characterized in the past both as a disease-related protein (Pande *et al.*, 2000, 2001; Basak *et al.*, 2003; Kosinski-Collins and King, 2003; Kosinski-Collins *et al.*, 2004; Flaugh *et al.*, 2005a, b, 2006; Mills *et al.*, 2007; Papanikolopoulou *et al.*, 2008; Moreau and King, 2009, 2012; Acosta-Sampson and King, 2010) and as a model protein for investigating the role of structural changes and 'hot spot' sequence modifications in mitigating aggregation (Sahin *et al.*, 2011; Costanzo *et al.*, 2014). In acidic solutions, it is much less stable than at neutral pH, and it aggregates on convenient experimental time scales without requiring one to add denaturants or heat to temperatures at or above the boiling point of water (Pande *et al.*, 2000; Flaugh *et al.*, 2006; Sahin *et al.*, 2011; Costanzo *et al.*, 2014). The present work extends those earlier studies and focuses on mildly acidic conditions (pH 5.5) with and without added NaCl. These are representative of the pH and salt conditions for many formulated therapeutic protein products (Wang, 2005), and also allow γDc aggregation rates to be quantified without adding denaturants. Protein variants were compared in terms of experimental B_{22} values and the Kirkwood–Buff integrals (G_{22}), unfolding free energy via chemical denaturation, and temperature-dependent aggregation rates.

Materials and methods

Calculated B_{22} values via Mayer sampling

Molecular simulations using the Mayer sampling algorithm (Singh and Kofke, 2004) were employed to predict which single-charge amino acid substitutions would have the greatest effect on protein–protein interactions, as quantified by B_{22}. The protein was treated at

a coarse-grained level of one bead per amino acid (1bAA), as this was shown previously to provide sufficient accuracy for the purposes of predicting B_{22} for different proteins as a function of pH and solution ionic strength (Blanco *et al*., 2013; Grünberger *et al*., 2013). Complete details of the force field and the Mayer sampling method are provided elsewhere (Singh and Kofke, 2004; Blanco *et al*., 2013; Grünberger *et al*., 2013).

Briefly, the PDB structure of γDc [PDB 1HK0 (Basak *et al*., 2003)] was used to create a 1bAA structural model of a protein monomer (Blanco *et al*., 2013). Each amino acid was specified by the center of its α carbon, bead diameter, charge and relative hydrophobicity. The values of the model parameters for each type of amino acid are tabulated elsewhere (Blanco *et al*., 2013; Grünberger *et al*., 2013). The only adjustable parameters in the model are: (i) the maximum value of the short-range attractive energy between a pair of amino acids (ε_{hp}); (ii) a parameter that scales the magnitude of charge–charge interactions within the implicit solvent approximation (ε_{cc}); and a scaling factor (α) that accounts for the fact that charge screening is not always captured quantitatively by the Debye–Huckel approximation. These three parameters were originally set by multivariate regression against experimental B_{22} values versus salt concentration for wild-type (WT) γDc (Blanco *et al*., 2013). The parameter values used for the present work differ from those described previously to account for the fact that the molecule in the present work contains a 6-His tag that is charged at pH 5.5. The new parameters are $\varepsilon_{hp} = 0.275\ k_BT$ and $\alpha = 2.617\ M^{-1/2}\ nm^{-1}$ with ε_{cc} chosen to maintain the same ratio $\varepsilon_{hp}/\varepsilon_{cc} = 3$ as in previous work (Blanco *et al*., 2013). For reference, the original values of ε_{hp} and α were $\varepsilon_{hp} = 0.375\ k_BT$ and $\alpha = 3.5\ M^{-1/2}\ nm^{-1}$.

Mayer sampling simulations were used to search the available configuration space for two monomers that were allowed to interact with each other over $\sim 10^6$–10^{10} configurations. At each configuration, the interaction energies between all constituent amino acids were summed, and compared with a reference state having a known value of B_{22}. The algorithm biases the choice of configurations to those that contribute most to the value of B_{22}, to allow for more accurate and faster convergence of calculated B_{22} values (Singh and Kofke, 2004). Uncertainties in calculated B_{22} values were estimated from a minimum of three independent simulations.

To evaluate the effects of single amino acid substitutions, each surface residue in the native sequence of γDc with a nominally titratable side chain (Asp, Glu, His, Lys and Arg) was considered as a candidate for mutation. Individually, each of these amino acids was changed to a residue of similar size, but with either zero or opposite charge. Overall, this resulted in 100 different point mutations that were evaluated for their predicted changes of B_{22}. In addition, molecular dynamics simulations were performed to evaluate the average intermolecular distance between two monomers for each candidate mutation (Blanco, 2013). Comparing the predicted B_{22} values and average intermolecular distance of the variants with those for WT γDc allowed selection of the ones that were predicted to make the greatest impact on protein–protein interactions; both more repulsive or more attractive net interactions. However, the 1bAA simulations do not allow the protein to change conformation, and therefore, additional calculations were required to assure that a given mutation would not negatively impact the conformational stability of γDc.

RosettaDesign to assess changes in conformational stability

The fixed backbone design protocol in RosettaDesign 3.0 was used to numerically estimate changes in the conformational stability or unfolding free energy ($\Delta\Delta G_{un}$) for each of the variants of γDc that were predicted from the B_{22} simulations to significantly alter electrostatic repulsions. RosettaDesign is a module of the molecular modeling program Rosetta (Kuhlman and Baker, 2000; Liu and Kuhlman, 2006; Kaufmann *et al*., 2010), that predicts changes in the conformational stability of the folded state of a protein, starting from an atomic structure such as an experimental pdb structure (Liu and Kuhlman, 2006). Briefly, the algorithm allows the side chains to move within rotamer libraries for each specific residue type (Dunbrack and Cohen, 1997; Kuhlman and Baker, 2000; Dantas *et al*., 2003), so as to find configurations with lower energy values. The energy for a given configuration was evaluated using the Rosetta scoring function that includes terms for attractive and repulsive Lennard-Jones potentials, the Lazaridis–Karplus implicit solvation model (Lazaridis and Karplus, 1999), orientation-dependent hydrogen bonding potentials (Kortemme *et al*., 2003), disulfide bonding contributions, torsion potentials from backbone and side chain realignment (Dunbrack and Cohen, 1997), and a reference term for each of the 20 amino acids based on the statistical probability of its presence in a protein sequence (Hu *et al*., 2008).

RosettaDesign searches for the lowest energy score using Monte Carlo optimization with simulated annealing (Liu and Kuhlman, 2006), based on the input sequence and design criteria. In this study, candidate single-point variants that were predicted to have the greatest effect on protein–protein interactions and intermolecular center-to-center distance, as described above, were input into RosettaDesign and the algorithm was permitted to search for a lowest energy configuration. The Rosetta energy scores of candidate variants were compared with that of the WT protein to find amino acid substitutions that minimized the predicted change to conformational stability. Of the variants that were tested in RosettaDesign, ~10% were predicted to have minimal changes to conformational stability compared the WT protein. A subset of these was selected for synthesis and experimental characterization.

Molecular cloning and site-directed mutagenesis

The gene for recombinant human γDc (WT) was obtained from the laboratory of Jonathan King in the form of a pQE-1 vector containing the gene sequence obtained from cDNA with an N-terminal 6-histidine tag fusion. Single-point mutations were introduced into the gene using the QuickChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA, USA) and transformed into XL-10 Gold *Escherichia coli* provided with the kit. Primers used for the mutations are listed in Supplementary Table SI. Cultures of *E.coli* were grown on Luria Berani (LB) agar plates supplemented with 100 μg/ml carbenicillin (Research Products International, Mount Prospect, IL, USA) or in 5 ml LB media supplemented with 50 μg/ml carbenicillin in culture tubes. Plasmids were harvested from *E.coli* using the Wizard Plus SV Miniprep kit (Promega, Madison, WI, USA). WT and mutant gene sequences were verified by gene sequencing.

WT and mutant genes were subcloned into a pET-15b vector, which contains a T7 promoter. The genes were amplified from the pQE-1 vectors using KOD Hot Start Master Mix (EMD Millipore, Billerica, MA, USA) and primers that added an NdeI restriction site to the 5′ end of the gene and BamHI site to the 3′ end. Amplified genes and empty pET15b vector were digested using NdeI and BamHI (New England Biolabs, Ipswich, MA, USA) and complete digestion was verified using gel electrophoresis. The digested WT and mutant genes were ligated into the digested vectors using T4 DNA Ligase (New England Biolabs) and transformed into either DH5α or

NEB5α *E.coli*. Proper integration was verified by digestion with BamHI or NdeI and/or gene sequencing.

Protein expression and purification

pET-15b-γDc WT and pET15b-mutant plasmids were transformed into BL21(DE3) *E.coli* for overexpression. Single colonies were selected from LB agar plates containing 100 μg/ml carbenicillin. Selected colonies were inoculated in 30 ml super broth (SB) media containing 50 μg/ml carbenicillin in 125 ml culture flasks and grown overnight at 37°C, shaking at 250 rpm. Primary cultures were diluted to an OD_{600} of 0.05 into two 650 ml secondary cultures of SB media containing 50 μg/ml carbenicillin in 2800 ml shake flasks. Secondary cultures were allowed to grow at 37°C, 250 rpm until on OD_{600} of 0.8–1.0 was reached. Isopropyl β-D-1-thiogalactopyranoside (Research Products International) was then added to each culture to a concentration of 1 mM to induce protein expression and each culture was grown for an additional 20–24 h at 30°C, 250 rpm.

Cultures were harvested by centrifugation at 5000*g* for 30 min and supernatant was discarded. *Escherichia coli* cell pellets were resuspended in 100 ml lysis buffer (50 mM sodium phosphate, 300 mM sodium chloride, 10 mM imidazole adjusted to pH 8.0) and subjected to cell lysis via at least five passes through a Microfluidics microfluidizer at minimum 8000 psi. The soluble fractions were obtained by centrifuging at 20 000*g* for 45 min and decanting from the insoluble debris. Soluble fractions were incubated on 10 ml Nickel-Nitrilotriacetic acid (Ni-NTA) agarose resin (Qiagen, Germantown, MD, USA) for at least 3 h rotating end-over-end. Ni-NTA resin was subsequently batch washed using 50 mM sodium phosphate, 300 mM sodium chloride, 20 mM imidazole, pH 8.0 buffer five times and transferred to an Econopac gravity flow column (Bio-Rad, Hercules, CA, USA). WT γDc and variants were eluted with a buffer containing 50 mM sodium phosphate, 300 mM sodium chloride and 250 mM imidazole, at pH 8.0. The protein-rich second column volume was immediately dialyzed for at least 12 h against 100 mM sodium phosphate, pH 7.4 to remove imidazole.

Prior to characterization, WT γDc and variants were further purified using Superdex 75 size-exclusion chromatography resin (GE Healthcare Life Sciences, Pittsburgh, PA, USA) packed into a Vantage L column (EMD Millipore) on an Akta Purifier FPLC (GE Healthcare Life Sciences). The mobile phase was 10 mM sodium phosphate adjusted to pH 7.5. For variants that added positive charge to γDc, it was necessary to increase the pH of the mobile phase to 8.0 due to interactions between the protein and the chromatography resin. Fractions containing γDc monomer were pooled and concentrated using a 3 kDa MWCO Amicon filter (EMD Millipore), followed by dialysis to 5 mM acetate pH 5.5 buffer with or without 350 mM sodium chloride. Protein concentration was measured by absorbance at 280 nm using an extinction coefficient of 41 040 M^{-1} (Sahin *et al.*, 2011). All experiments were performed within 3 days of preparation to avoid soluble dimer formation that occurs slowly at 4°C (Pande *et al.*, 2000; Sahin *et al.*, 2011; Costanzo *et al.*, 2014) (see also the Results and Discussion sections).

Differential scanning calorimetry

γDc WT and variants were each diluted to 0.5 mg/ml in 5 mM sodium acetate buffer at pH 5.5 and analyzed in a Microcal VP Differential Scanning Calorimeter (VP-DSC). At least five buffer–buffer scans were performed at a scan rate of 60°C/h prior to addition of the protein sample to establish a baseline over the temperature window of the experiment (20–90°C). Sample scans were performed immediately after buffer–buffer scans to maintain thermal history. Only a single scan was performed for each sample due to the observation of exotherms in each sample after the temperature ramp, which is indicative of irreversible precipitation at the highest temperatures when using this experimental configuration (Roberts *et al.*, 2013). Differential power values were converted to heat capacity using standard methods (Privalov, 1979; Andrews and Roberts, 2007).

Chemical unfolding

Chemical unfolding experiments were performed using increasing concentrations of guanidinium hydrochloride (GdnHCl, Fisher Scientific, Pittsburgh, PA, USA) as a denaturant. A 1.0 mg/ml stock solution of γDc in buffer was diluted 10-fold into varying concentrations of GdnHCl using appropriate combinations of a stock of concentrated GdnHCl solution and buffer. All dilutions were performed by mass, accounting for the known dependence of solution density on GdnHCl concentration (Kawahara and Tanford, 1966). Samples were allowed to incubate for 24 h to reach equilibrium and the intrinsic tryptophan fluorescence was measured using a Jobin Yvon FluoroMax-3 Spectrometer (Horiba, Edison, NJ, USA). An excitation wavelength of 295 nm was used to selectively excite tryptophan residues and obtain emission spectra over the range from 310 to 420 nm. Buffer-only samples at all GdnHCl concentrations were also measured for baseline subtraction.

A characteristic native state peak at 320 nm was observed along with the anomalous quenching pattern characteristic of γDc, in that emission intensity increases, rather than decreases, upon unfolding (Kosinski-Collins and King, 2003; Kosinski-Collins *et al.*, 2004). Because of this, two different methods were used to estimate the fraction of unfolded protein with increasing denaturant concentration, based on either spectral center-of-mass (Andrews and Roberts, 2007; Li *et al.*, 2010; Brummitt *et al.*, 2012) or ratios of peak intensities (Kosinski-Collins and King, 2003; Kosinski-Collins *et al.*, 2004; Flaugh *et al.*, 2005a, b, 2006; Mills *et al.*, 2007; Sahin *et al.*, 2011; Costanzo *et al.*, 2014). Standard two-state (Pace *et al.*, 1989) and three-state (Clark *et al.*, 1993) models for unfolding were tested in all cases.

Spectral center of mass [COM, Equation (1)] was used as one means to quantify the unfolding of the protein with increasing denaturant concentration.

$$\mathrm{COM} = \frac{\sum_\lambda (1/\lambda) I_\lambda}{\sum_\lambda I_\lambda} \quad (1)$$

where λ represents the emission wavelength of a given spectra (in nm) and I_λ the intensity of emitted light at wavelength λ (in counts per second). Alternatively, the ratio of fluorescence emission at 360 nm to that at 320 nm was also used to quantify the fraction of unfolded protein with increasing denaturant concentrations as described previously (Flaugh *et al.*, 2005a, b, 2006; Sahin *et al.*, 2011; Costanzo *et al.*, 2014). All fitting was performed using the nlinfit function of MATLAB (Mathworks, Natick, MA, USA).

Static light scattering

Prior to static light scattering measurements, protein and buffer stocks were filtered an additional time through a sterile 0.22 μm syringe filter (Celltreat, Shirley, MA, USA) to ensure minimal carryover of dust, and stock protein concentrations were verified with UV absorbance after filtration. γDc WT and variants were diluted with buffer to provide protein concentrations ranging from 0.5 to 10 mg/ml in 5 mM sodium acetate pH 5.5, with and without 350 mM sodium chloride. Samples

were then analyzed using a Multiple Angle Laser Light Scattering (MALS) detector (Wyatt Technology, Santa Barbara, CA, USA) outfitted with a microcuvette batch sample holder. Samples were allowed to equilibrate at 25°C and the intensity of light scattered was collected for at least 5 min to ensure sufficient data for high-quality statistics. All data were inspected and any erroneous signals due to scattering from dust were manually eliminated from the averaging processes. Scattering intensity values were converted to values of the excess Raleigh ratio divided by the optical constant (R/K) using instrument calibration procedures described elsewhere (Barnett *et al.*, 2015). The value of dn/dc of 0.187 cm^3/g for γDc was measured previously (Blanco *et al.*, 2013). Equation (2) was regressed against the R/K data as a function of protein concentration (c_2) (Blanco *et al.*, 2011).

$$\frac{R_{90}^{ex}}{K} = M_{2,app}c_2 + M_2 G_{22}(c_2)^2 \quad (2)$$

where M_2 and $M_{2,app}$ (both with units of Da) denote the molecular weight and apparent molecular weight of the protein, respectively. G_{22} denotes the protein–protein Kirkwood–Buff integral and is related to the second osmotic virial coefficient, B_{22}. G_{22} is the more appropriate quantity when there are strong attractions or repulsions, and therefore holds for a wider range of conditions than does B_{22} (Blanco *et al.*, 2011) (see also the Results and Discussion sections). M_2 was determined from the amino acid sequence for each γDc variant, while $M_{2,app}$ and G_{22} are regressed from the data using non-linear least squares fitting in MATLAB. For ideal solutions, $M_{2,app}$ equals M_2, but the two quantities can differ if there are strong protein–solvent and protein–cosolute interactions (Blanco *et al.*, 2011).

Aggregation rates from temperature scanning monomer loss

The temperature scanning monomer loss (TSML) approach is described in detail elsewhere (Brummitt *et al.*, 2011a; Roberts *et al.*, 2013). Briefly, γDc WT and each variant were diluted separately to 0.5 mg/ml in buffer (5 mM sodium acetate, pH 5.5) with or without 350 mM NaCl present. In silica cuvettes, 2.5 ml samples were placed and heated at a constant 60°C/h temperature ramping rate, from an initial temperature of 20°C to a maximum temperature arbitrarily chosen to assure measurable amounts of monomer loss were observed for a given variant. Aliquots were taken at predetermined temperature points during the temperature ramp and immediately quenched on ice to stop further aggregation. Samples were centrifuged at ~9000g for 10 min to remove any insoluble aggregates from solution. The supernatant was analyzed with the Size Exclusion Chromatography method described below.

Monomer fraction values (m) as a function of temperature (T) were fit to a simple asymmetric sigmoidal model that was shown previously to robustly interpolate m as a function of temperature from TSML experiments (Brummitt *et al.*, 2011c; Roberts *et al.*, 2013).

$$m(T) = \exp(-\exp(b(T - T_{ref}))) \quad (3)$$

In Equation (3), b and T_{ref} are fitted parameters. Fitted $m(T)$ curves were used to interpolate the value of m at a desired temperature. This allows one to estimate k_{obs}, the observed or net rate of monomer loss at a given temperature based on Equation (4), which is derived in the Appendix, and holds for conditions where m is close to 1.

$$k_{obs} = k_0 e^{-E_{a,eff}/R(T^{-1} - T_{ref}^{-1})} \quad (4)$$

In Equation (4), $E_{a,eff} = bRT_{ref}^2$ and denotes the apparent activation energy for aggregation (see also, Appendix), and k_0 denotes the value of k_{obs} when the temperature equals T_{ref}. Temperature-dependent aggregation rates based on the TSML experiments were determined for each variant and for the WT protein.

Size exclusion chromatography

Incubated samples and standards were transferred to deactivated borosilicate glass HPLC vials (Waters, Milford, MA, USA). Injections of 0.1 ml were separated with a Tosoh TSK2000xl column (Tosoh Bioscience, King of Prussia, PA, USA) attached to an Agilent 1100 high-performance liquid chromatography system (Agilent Technologies). The system included an in-line variable wavelength detector (Agilent Technologies) set to 280 nm. Monomer fraction was determined by comparing the total area under the monomer peak with that for unheated external standard samples that were identical to the unheated stock solution for a given variant or WT.

Results

Prediction and selection of variants for experimental characterization

Molecular simulations were performed as described in the Materials and methods section to calculate B_{22}, which is reported here in a dimensionless form as B_{22}^* [$B_{22}^* = B_{22}/B_{22}^{HS}$ where B_{22}^{HS} = 80.69 L/mol, or ~0.004 l/g, and represents the 'hard sphere' or purely steric value of B_{22} for γDc that can be calculated from a simple MC simulation for any protein with a known crystal structure (Blanco *et al.*, 2013; Grünberger *et al.*, 2013)]. Charge-variants predicted to exhibit the most significant changes in B_{22}^* were hypothesized to have the greatest effects on modulating aggregation rates and were selected for further screening. Supplementary Table SII contains the charge alterations selected for calculations to assess possible changes in conformational stability. RosettaDesign was used to further narrow the pool of candidate variants by determining residue substitutions with minimal change to conformational stability. Supplementary Table SIII shows the residues predicted by RosettaDesign to satisfy these secondary design criteria. Of these variants, a subset was arbitrarily chosen for experimental characterization that encompasses variants that were predicted to either improve aggregation resistance or reduce it. Opposite-charge variants were predicted to have the most measureable effect on protein–protein interactions. Therefore, variants were chosen primarily from that subset. One variant in which a charged amino acid was substituted for a neutral residue was also chosen for comparison.

Table I lists the first round of variants that were selected for experimental characterization. Two variants, D149K and E17R, were predicted to increase the repulsive protein–protein interactions and therefore were expected to increase resistance to aggregation. The other three variants, R141N, R141E and H22D, were predicted to

Table I. Calculated B_{22}^* values and experimentally determined T_m values from DSC for WT γDc and variants

Variant	Δ Charge	Predicted B_{22}^*	T_m (°C)
WT	–	20.0 ± 1.0	83.3
D149K	+2	24.6 ± 0.6	84.0
E17R	+2	24.5 ± 1.4	80.1
R141N	−1	17.7 ± 0.9	82.0
R141E	−2	15.7 ± 0.8	80.5
H22D	−2	16.1 ± 1.0	78.0

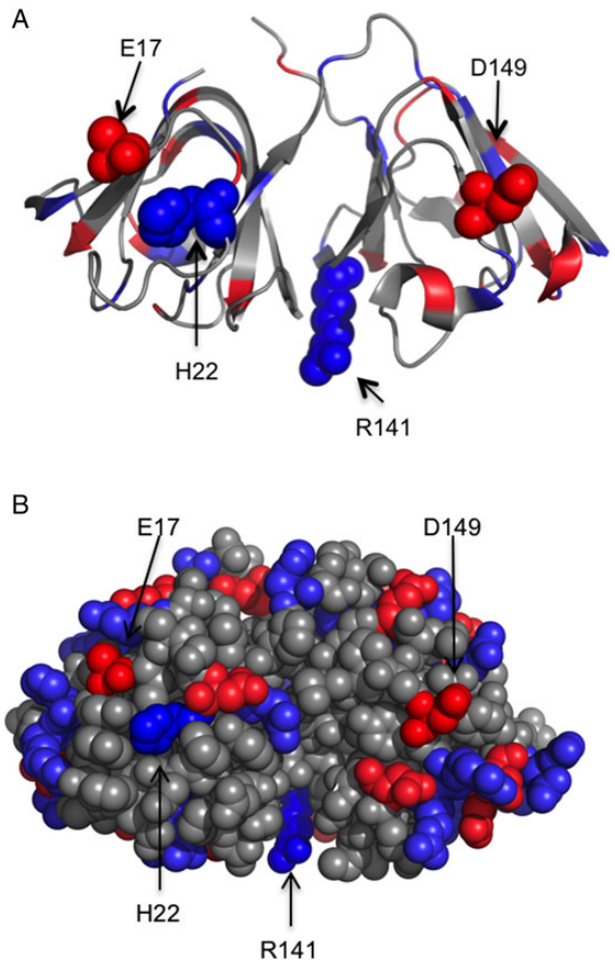

Fig. 1 Residues selected for substitution are all solvent-exposed surface residues, shown both on the protein backbone (**A**) and a space fill model of the protein's surface (**B**). The locations of residues selected for single-charge point mutations on the crystal structure of human γDc (PDB: 1HK0) are indicated with arrows.

have the opposite effect, decreasing repulsive protein–protein interactions and aggregation resistance (i.e. increased aggregation rates). R141N was predicted to have a smaller effect on protein–protein interactions, with only a change in charge of −1. Figure 1 shows the locations of the residues selected for mutation, as well as the distribution of the other charged residues over the surface. Inspection of the crystal structure shows that the residues selected are highly surface exposed and therefore are anticipated to have minimal effect on folding stability, as predicted by RosettaDesign.

Determination of thermal stability using differential scanning calorimetry

The relative thermal stability of each γDc variant was compared with that of the WT protein using differential scanning calorimetry (DSC). The results, shown in Fig. 2, show that D149K exhibited a higher midpoint transition temperature (T_m) than WT, while E17R, R141N, R141E and H22D all had lower T_m values. The T_m values are summarized in Table I. If only unfolding of the protein was occurring upon heating of the sample, these results would suggest that these amino acid substitutions destabilized the folded state of the protein.

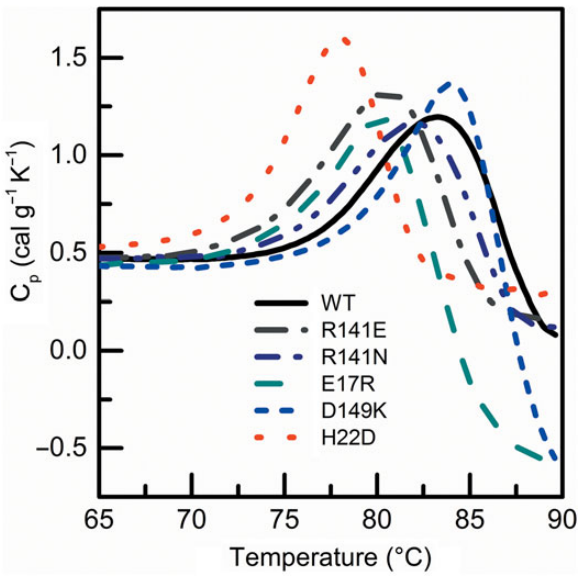

Fig. 2 DSC thermograms for each variant (see legend), compared with the WT (solid curve).

However, precipitation was observed in each sample upon removal from the instrument after completion of the scan, and each thermogram showed a large exotherm above T_m, indicating that irreversible protein aggregation occurred during the temperature scan. This convolutes the results and makes thermodynamic analysis or interpretation of the DSC data unreliable. Theoretical work has shown that irreversible aggregation can cause large shifts of T_m values to lower temperatures, even when changes in conformational stability of the folded state are minimal (Sanchez-Ruiz, 1992). Therefore, the conformational stability of the protein was assessed by chemical denaturation.

Conformational stability from isothermal chemically induced unfolding

Chemical unfolding was performed using guanidinium hydrochloride as a denaturant because the unfolding free energy for γDc is too large for urea to be effective unless one works at much lower pH values (Kosinski-Collins and King, 2003; Kosinski-Collins *et al.*, 2004; Costanzo *et al.*, 2014). Fluorescence spectra showed the characteristic peak of folded γDc at 320 nm, and spectra in the absence of denaturant were indistinguishable for each variant when compared with WT, similar to what was observed in prior work (Pande *et al.*, 2000; Kosinski-Collins and King, 2003; Kosinski-Collins *et al.*, 2004; Sahin *et al.*, 2011; Costanzo *et al.*, 2014). Spectra obtained with increasing concentrations of denaturant showed a red shift along with a Trp quenching pattern seen previously for γDc, specifically the loss of the peak at 320 nm along with an increase in intensity at higher wavelengths ~360 nm (Kosinski-Collins and King, 2003; Kosinski-Collins *et al.*, 2004; Sahin *et al.*, 2011; Costanzo *et al.*, 2014). Illustrative spectra for WT γDc as a function of denaturant concentrations are shown in Fig. 3A. Spectra for all variants are provided in Supplementary Fig. S1 and show that similar patterns were observed in all cases.

Figure 3A shows that there is no isobestic point for the spectra as a function of denaturant concentration, and the changes of the spectra vs. GdnHCl are not consistent with a two-state unfolding transition (Pace *et al.*, 1989). This suggests a folding intermediate may be

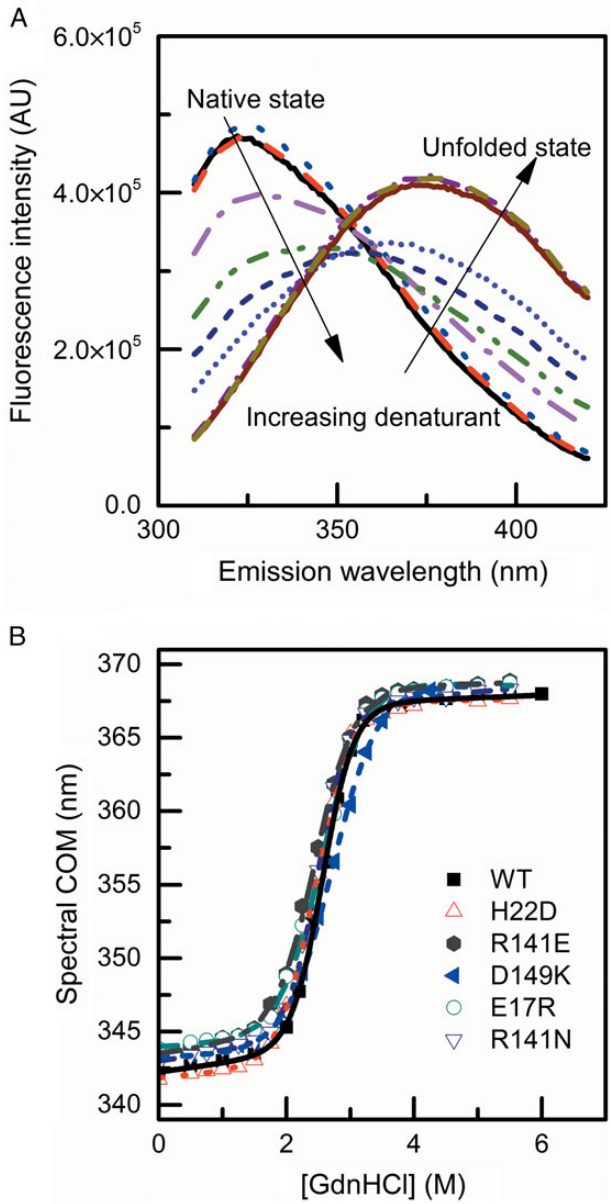

Fig. 3 (**A**) Illustrative intrinsic fluorescence emission spectra of WT γDc as a function of concentration of denaturant. The 0 M gdnHCl spectra represent the native state of γDc, while the 5.5 M spectra and intermediate concentration spectra represent fully unfolded and partially unfolded states, respectively. Arrows indicate increasing denaturant concentration. (**B**) Spectral COM as a function of denaturant concentration for WT and each variant (see legend). Symbols are experimental data and curves are fits to a two-state model (see also, main text).

present. The spectral center of mass (Andrews and Roberts, 2007; Li *et al*., 2010; Brummitt *et al*., 2012) and the ratio of emission intensity at 360–320 nm were each used to quantify the extent of unfolding, as done previously for this protein (Flaugh *et al*., 2005a, b, 2006; Sahin *et al*., 2011; Costanzo *et al*., 2014). Both methods yielded apparent two-state unfolding curves, as evidenced by a single inflection point. Figure 3B illustrates this with the COM results as a function of [GdnHCl]; equivalent results were obtained using intensity ratios (not shown). In all cases, the profiles showed statistically negligible shifts in the midpoint gdnHCl concentration (C_{mid}) for the single discernable transition. The only notable differences were small shifts in the pre-transition and post-transition baselines.

The data were fit to both two-state and three-state models. As anticipated by inspection of the shape of the profiles in Fig. 3B, and consistent with prior work (Kosinski-Collins and King, 2003; Kosinski-Collins *et al*., 2004; Sahin *et al*., 2011; Costanzo *et al*., 2014), the three-state model fits did not provide statistically meaningful improvements over the two-state model fits. Given the observations above regarding (i) the clearly non-two-state unfolding pattern for the full FL spectra (Fig. 3A and Supplementary Fig. S1), (ii) limitations of model-discrimination between two- and three-state models and (iii) the overlap of the profiles in Fig. 3B, it was concluded that the intrinsic fluorescence measurements were not able to distinguish differences in conformational stability (unfolding free energy) between the variants and WT protein. While it might be expected that hydrophilic, single amino acid substitutions on the protein surface would not result in significant changes in stability, the present results cannot rule the possibility of small unfolding free energy differences that were beyond the resolution of the present methods and results. It was noted in previous work (Costanzo *et al*., 2014) that RosettaDesign was most effective in predicting changes in experimental unfolding free energy values if the values are relatively large, and so even the predicted changes that differ slightly for the variants tested here (see Supplementary data) are likely within the statistical uncertainty of the computational methods.

Unfortunately, the need for guanidinium as the denaturant (rather than urea) made it unrealistic to assess the unfolding free energy at low ionic strength, where the surface charge-swap variants might be expected to have the greatest effects on unfolding free energies. As noted earlier, DSC results were convoluted with aggregation and unfolding, and could not provide a robust alternative. As a result, it was not possible to overcome those limitations in the present work. Additional difficulties with the use of DSC to predict effects on aggregation rates at lower temperatures are noted below (see the Discussion section).

Protein–protein interactions from static light scattering

Protein–protein interactions were measured for WT γDc and each variant using static light scattering. Excess Rayleigh ratio values, as a function of protein concentration (see the Materials and methods section), are plotted in Fig. 4A. Strong downward curvature, characteristic of repulsive protein–protein interactions, is seen even at relatively low concentrations in all cases. Fits to Equation (2) allow the interactions to be quantified in terms of G_{22}. The more traditional quantity to report is B_{22}, but recent work has shown that doing so is quantitatively accurate only when the dimensionless product of B_{22} (units of volume/mole or volume/mass) and protein concentration (c_2) is small (absolute value less than ~0.05) (Ben-Naim, 1992; Blanco *et al*., 2011). When the absolute value of c_2B_{22} is large, G_{22} is the more appropriate measure of net protein–protein interactions, and Equation (2) should be used instead of traditional expressions that assume weak interactions and/or extremely low c_2 (Blanco *et al*., 2011).

All of the data for the variants at low ionic strength conditions (Fig. 4A) were consistent with large repulsions. Figure 4A illustrates that even a single amino acid substitution makes a significant difference in the R/K profiles as a function of c_2. Variations that altered the charge by a magnitude of 2 had the greatest effect. It can also be seen that the lowest concentration data points overlap. This is expected at the lowest c_2 values, as it indicates negligible differences in molecular weight, and therefore no convolution of the results due to protein degradation (e.g. fragmentation or aggregation).

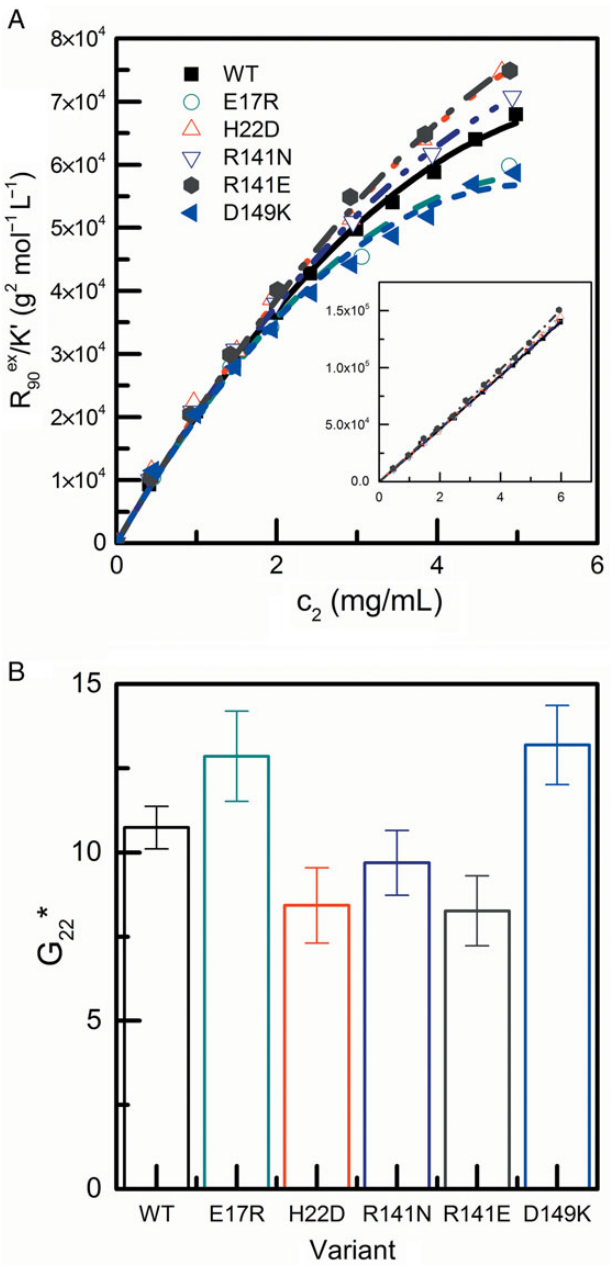

Fig. 4 Protein–protein interactions determined for WT and protein variants via static light scattering. (**A**) Excess Rayleigh ratio as a function of concentration for the WT protein and each variant in 5 mM sodium acetate, pH 5.5 and 5 mM sodium acetate, 350 mM NaCl, pH 5.5 (inset); symbols are experimental values at curves are fits; (**B**) fitted values of G_{22}^* obtained by fitting the data in (A).

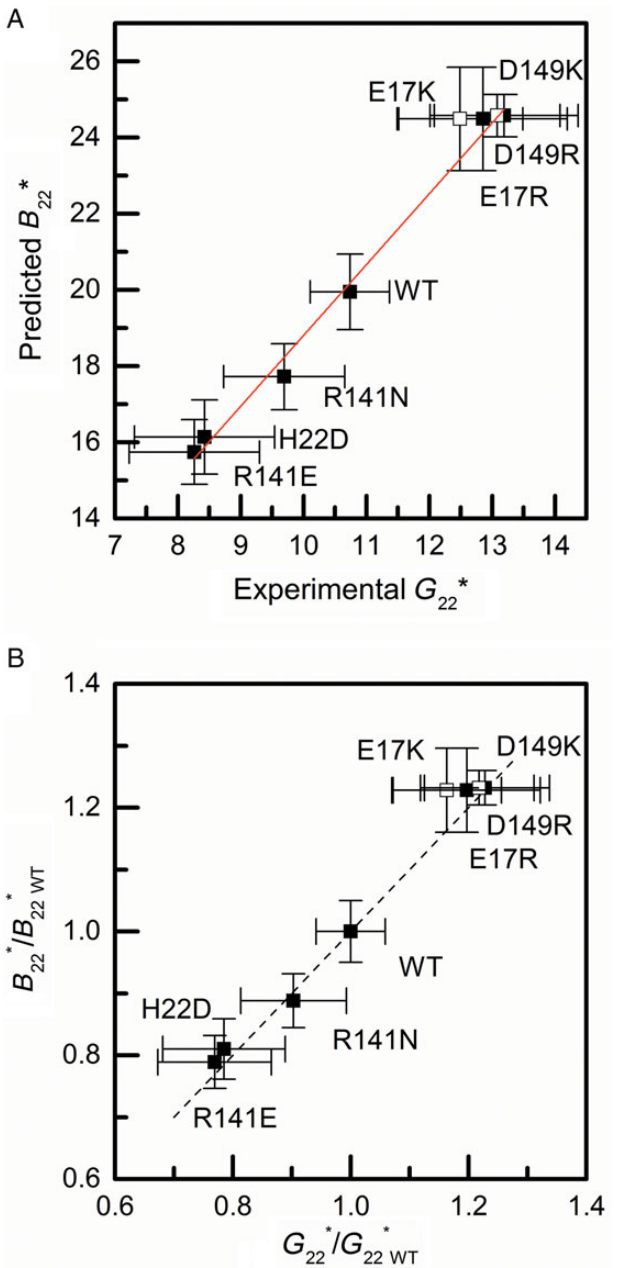

Fig. 5 (**A**) Comparison of predicted B_{22}^* values and the experimental G_{22}^* values for WT and variants at low ionic strength. The line is a linear fit of data (R^2-value of 0.991). At high ionic strength, there is no measured or predicted difference in B_{22}^* or G_{22}^* across WT and these variants (see also, main text). Closed symbols are from the first round of variants, while open symbols are for the second round. (**B**) Comparison of the data from (A) when the B_{22} and G_{22} values are scaled by the respective values for WT. The line corresponds to a perfect one-to-one correlation and is a guide to the eye.

The values of G_{22}^* from fitting the data in Fig. 4A are given in Fig. 4B, where $G_{22}^* = G_{22}/G_{22,\mathrm{ref}}$ was used to compare protein–protein interactions of the variants. $G_{22,\mathrm{ref}}$ is defined as $-2B_{22}^{\mathrm{HS}}$, based on the definition of G_{22} in the limit of low protein concentration (Ben-Naim, 1992). Using this definition, B_{22}^* and G_{22}^* are quantitatively equivalent in the limit of sufficiently low protein concentration or weak protein–protein interactions. Values of $M_{2,\mathrm{app}}$ (not shown) were similar for WT and all variants and were within 5% of the true molecular weight once one accounts for the 6-His tag. From Fig. 4B, it is apparent that the D149K and E17R variants increase repulsive protein–protein interactions, while the R141N, R141E and H22D variants decrease repulsive interactions, relative to WT. The results are consistent with the predicted trends in B_{22} for each amino acid substitution.

Figure 5A shows that there is a linear correlation between the predicted values of B_{22}^* with the experimentally determined values of G_{22}^*. However, the magnitude of the G_{22}^* values was consistently lower than the predicted B_{22}^* values. Note that the predictions are necessarily B_{22} values rather than G_{22} values, because the simulations only deal with

the dilute limit where a given protein only 'feels' interactions with one neighboring protein at a time. If the interactions are sufficiently strong/long-ranged, then proteins will interact significantly with more than one neighboring protein even at seeming dilute conditions. G_{22} captures that difference because it was obtained at finite c_2, while B_{22} is necessarily for $c_2 \to 0$ (Blanco *et al.*, 2011). Based on solution theory, one expects G^*_{22} to be lower in magnitude than B^*_{22} under conditions of strong repulsions if one cannot reach a sufficiently low c_2 for the two quantities to become equal (Ben-Naim, 1992; Blanco *et al.*, 2011). As such, the results in Fig. 5A are taken to represent a reasonable success for the Mayer-sampling CG model predictions. Empirically, it may be possible to phenomenologically correct for this issue by scaling each predicted B_{22} value by the value for WT protein, and similarly for the experimental G_{22} values. Figure 5B shows that doing so provides almost quantitative agreement between the predicted and experiment results (dashed line in the figure shows the result for a perfect one-to-one correlation, as a guide to the eye).

If all of the differences in measured G^*_{22} values are due to electrostatic interactions, then the light scattering profiles should be indistinguishable at high salt concentrations, as a result of charge screening effects (Velev *et al.*, 1998; Blanco *et al.*, 2013). The inset to Fig. 4A shows R/K vs. c_2 for WT, R141N, R141E and H22D with 350 mM NaCl included in solution. As expected, the curves almost completely overlap and show weakly attractive interactions (i.e. minor upward curvature). Fitting the data to Equation (2) provided values of G^*_{22} that were not statistically different from zero in each case (see Supplementary Fig. S2). Consistent with the loss of repulsive interactions due to charge screening, SLS measurements for D149K and E17R at high ionic strength were not tenable due to the formation of a significant (>10%) amount of soluble dimer (measured via SEC) on the time scale of sample preparation. This effect has been observed previously with this protein in various conditions (Pande *et al.*, 2000; Sahin *et al.*, 2011; Mohr *et al.*, 2013; Costanzo *et al.*, 2014).

Temperature-dependent aggregation rates from TSML

Samples quenched at different temperatures during temperature ramping (60°C/h) were analyzed using size exclusion chromatography (see the Materials and methods section). With the exception of the E17R variant (see below), aggregates were insoluble and chromatograms contained a single peak representing the remaining protein monomer. For unheated samples, there were small amounts (<1%) of soluble dimer present. Representative chromatograms are provided in Supplementary Fig. S3B. Figure 6A and B displays the monomer concentration divided by its initial concentration (i.e. fraction monomer remaining, m) as a function of temperature for the low and high ionic strength conditions, respectively. Curves are fits to Equation (3) for the purposes of interpolation (see the Materials and methods section). Values of b and T_{ref} from Equation (3) are tabulated in Supplementary Tables SIV and SV along with the values of the effective activation energy ($E_{a,eff}$).

Previous work showed that the TMSL method illustrated in Fig. 6A provides quantitative or semi-quantitative estimates of aggregation rate vs. temperature if one focuses on the initial-rate regime, so that one does not need to know or assume a detailed rate law or reaction order (Brummitt *et al.*, 2011c; Roberts *et al.*, 2013). Values for k_{obs} from TSML essentially provide the rate of monomer loss on a fractional basis (units of 1/time), and provide temperature-dependent rates with much less material consumption compared with conventional isothermal aggregation kinetic methods (Brummitt *et al.*, 2011c). They should be confined to temperatures near or below (within ~10°C) of the initial downturn in the profiles of m vs. T (see Fig. 6A) (Brummitt *et al.*, 2011c).

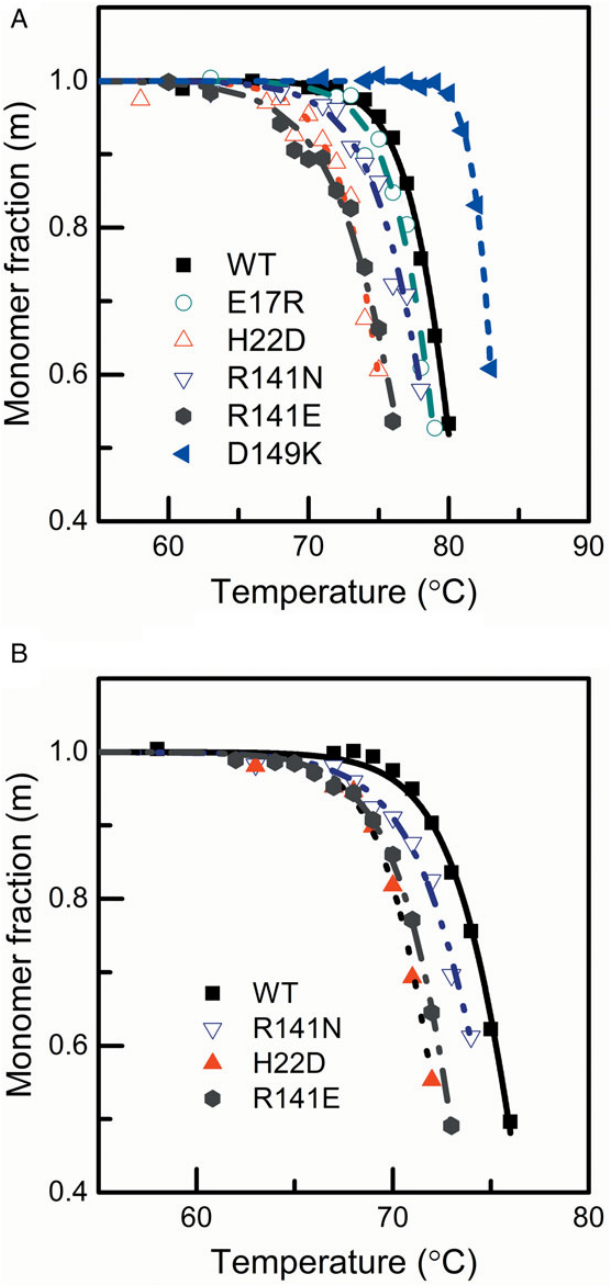

Fig. 6 TSML profiles for quantifying temperature-dependent aggregation rates for WT and each variant. Monomer fraction vs. temperature for the WT protein and each variant at: (**A**) 5 mM sodium acetate, pH 5.5; (**B**) 5 mM sodium acetate, 350 mM NaCl, pH 5.5.

Inspection of Fig. 6A shows that E17R, R141N, R141E and H22D all aggregate faster than WT at a given temperature, indicating reduced aggregation resistance, and vice versa for D149K. Figure 6B is analogous to Fig. 6A, except that all solutions included 350 mM NaCl; data are not included for E17R and D149K for the reasons discussed earlier. Qualitatively, it can be seen that the differences between the TSML profiles for the variants and WT are less pronounced at this higher ionic strength. Equation (4) was regressed for each profile in Fig. 6A and B. The resulting parameters and confidence intervals are reported

in Supplementary Tables SIV and SV. Using Equation (4) and the fitted parameters, one can interpolate or extrapolate values for k_{obs} for a given T and choice of variant or WT, for the solution conditions in Fig. 6A or Fig. 6B.

Discussion

The results in Fig. 5 indicate that the molecular simulations were able to semi-quantitatively or quantitatively predict the experimental behavior of G_{22} for a given choice of charge variant, depending on whether one scales the results by those for the WT protein. Independent of how well the experimental G_{22} values are predicted from theory, a main hypothesis of the present work was that designing variants to alter G_{22} would prove effective in mitigating aggregation rates. To assess how well such a strategy might work, Fig. 7 (filled symbols) shows a comparison of experimental k_{obs} values for $T = 60$°C, with experimental G_{22}^* values. The latter were necessarily determined at room temperature, as aggregation would be too fast to allow G_{22}^* values to be determined at the same temperatures as the aggregation rates.

Figure 7A and B is for low- and high-ionic strength conditions, respectively. Dotted/dashed lines are guides to the eye for where WT values fall in Fig. 7A. Consistent with the above hypothesis, there is a semi-quantitative correlation between G_{22}^* and k_{obs} at low ionic strength, where charge–charge repulsions are anticipated to slow aggregation. At high ionic strength, although a qualitative trend is also visible, it is not discernable once one accounts for the statistical uncertainties (cf., overlap of error bars in Fig. 7B). The fact that there are some differences in aggregation rates for the different variants at high ionic strength indicates that electrostatic repulsions are not the only aspect of the aggregation mechanism that was altered by these point mutations. Substituting a single amino acid necessarily also involves altering short-range interactions as well as intramolecular forces associated with the residue. It is possible that there are small differences in conformational stability that cause those differences in aggregation rate, but as noted in the Results section, the conformational stability differences may be too small to be discernable directly with the methods here. The necessity of using GdnHCl to unfold γDc in the conditions of interest of the present study also presents limitations such as the inability to characterize the effects of ionic strength on conformational stability. A more detailed study of the conformational stability differences with higher resolution techniques that can detect small amounts of partially unfolded proteins is the focus of future work.

DSC measurements were also performed to compare the conformational stability of the variants to WT γDc. However, as noted in the Results section, significant aggregation occurs during the temperature scan that convolutes the measurement of conformational stability. Inspection of the thermograms in Fig. 2 as well as the T_m values in Table I compared with the TSML curves in Fig. 6A and aggregation rate values in Fig. 7A shows that T_m values correlate qualitatively with aggregation rates. There are, however, outliers such as noted below, which make it difficult to conclude with certainty that these apparent T_m values are even qualitatively predictive of aggregation rates (relative to WT) at lower temperatures.

The closed symbols in Fig. 7A are those from the first round of variants. The results show that only one of the two variants that were predicted to slow aggregation proved to be successful. This raised the question of whether the E17R or the D149K variant was the outlier. That is, it was unclear if the position of the substitution was the cause of the discrepancy of E17R from the expected behavior, or if it was a limitation of the charge-swap strategy to improve aggregation resistance. As a result, two additional variants at the E17 and D149 positions were produced and they were characterized at the low ionic strength conditions where the charge-swap strategy was most effective for the first round of variants. Both E17K and D149R fulfilled the design criteria used for the initial set of variants, and they also provided a means to evaluate if the position of the substitution or the choice of arginine vs. lysine altered the observed behavior. Additional details regarding creation of the new variants (e.g. primer sequences) are provided in Supplementary data.

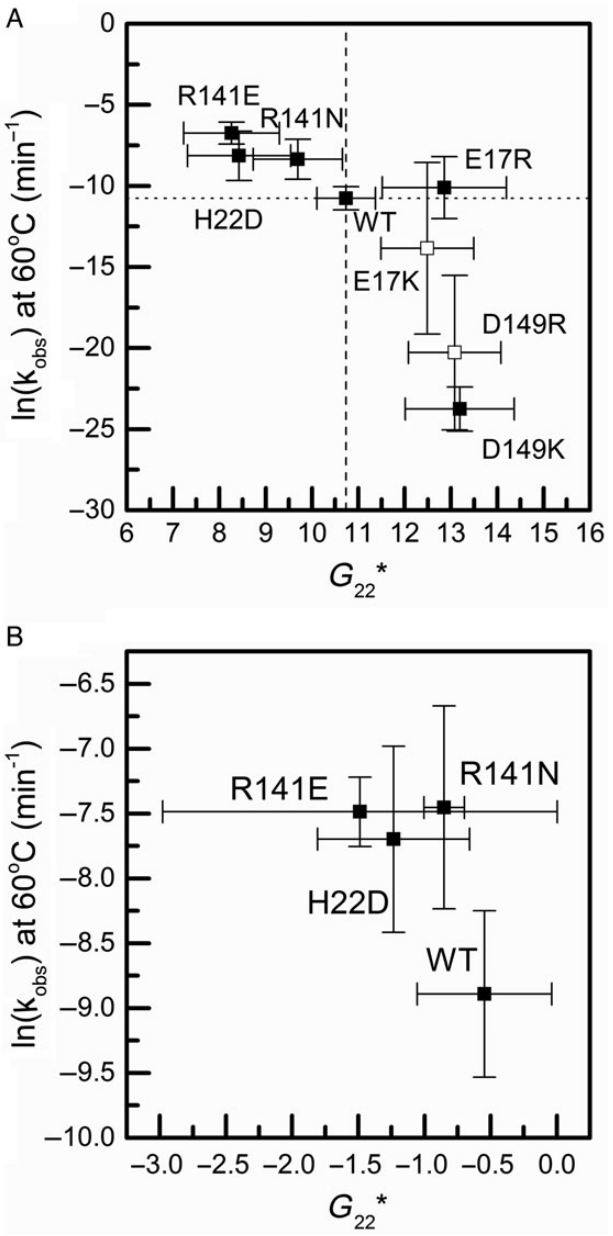

Fig. 7 Comparison of experimental aggregation rates and G_{22}^* values. (**A** and **B**) are for the same solution conditions as in Fig. 6A and B, respectively. k_{obs} values are for $T = 60$°C, while G_{22}^* is for room temperature due to practical constraints (see also, main text). Dotted and dashed lines representing WT values for ln(k_{obs}) at 60°C and G_{22}^*, respectively, are provided as guides for the eye in (A). Open symbols in (A) represent the second round of variants at E17 and D149.

as Fig. 7A and compared with the WT protein. DSC data in Supplementary Fig. S4 show that the T_m for E17K is much lower than that of WT and the T_m value for D149R is slightly lower than WT (see also Supplementary Table SVII). Static light scattering showed increased repulsive protein–protein interactions (Supplementary Fig. S5A) and greater G_{22}^* values (Supplementary Fig. S5B) for both the E17K and D149R variants compared with WT. TSML results (Supplementary Fig. S7) were used to compare aggregation resistance of the new variants with the WT protein and the other variants. Fit parameters from regressing Equation (3) to the data in Supplementary Fig. S7 are tabulated in Supplementary Table SVIII. The results show that the D149R variant showed greatly increased aggregation resistance compared with the WT protein, while E17K provided only marginal improvements. The open symbols in Fig. 7A show the results for D149R and E17K overlaid with those for the earlier variants.

Taken together, the results in Fig. 7A show that, with the exception of E17R and E17K, the variants that were predicted to decrease (increase) aggregation rates correlate semi-quantitatively with increased (decreased) repulsive electrostatic interactions under low-ionic strength conditions. Supplementary Figure S7 shows that the same qualitative conclusion holds for aggregation rates at other temperatures. E17R and E17K were outliers in terms of aggregation rates as a function of electrostatic repulsions, as captured by G_{22}^*. The experimental values of G_{22}^* for both E17R and E17K are more repulsive than that for WT; therefore, this could not simply be a result of less electrostatic repulsion than was expected or predicted for E17R and E17K, relative to WT.

It is hypothesized that E17R and E17K were outliers in Fig. 7A because of something specific to the aggregation process. Previous work with other proteins has shown that changes in net charge, due to changes in pH, can alter the dominant pathway(s) for aggregate growth. Most relevant to the present observations are cases where changes in protein charge state cause aggregates to change from being soluble to insoluble, or vice versa (Kroetsch *et al.*, 2012; Sahin *et al.*, 2012; Kim *et al.*, 2013; Barnett *et al.*, 2015). Theoretical arguments show that monomer loss rates are expected to be higher when aggregates remain soluble, if other key factors remain the same (e.g. temperature, ionic strength, unfolding free energy). This holds because when aggregates are soluble, monomers can be consumed both by the process of nucleating new aggregates and by addition to the soluble aggregates (Weiss *et al.*, 2009; Kim *et al.*, 2013). When aggregates are insoluble, monomer loss occurs only by creation of new aggregates that quickly fall out of solution (Weiss *et al.*, 2009; Kim *et al.*, 2013).

Supplementary Figure S3A shows that a significant soluble aggregate peak in SEC forms over the course of the temperature ramp for E17R. Similar behavior was seen for E17K. This was not present for WT or any of the other variants (see Supplementary Fig. S3B). Additional isothermal kinetic measurements for monomer loss and buildup of soluble aggregates were qualitatively consistent with this conclusion (data not shown); and there was also evidence of the soluble aggregates growing so large that they were eventually filtered irreversibly in SEC. Since both E17 mutations resulted in a significant change in aggregate solubility at pH 5.5, it is possible that the position of the mutation may play a role in the change in aggregate solubility. E17 is in the N-terminal domain of gDc, while D149 is found in the C-terminal domain. Figure 1B shows the distribution charges on the surface of gDc. It is possible that substituting a negatively charged amino acid (E17) with a positively charged, one could potentially lead to patches of positive charge on the surface of one domain. It is possible that this then affects the solubility of aggregates if the N-terminal domain 'projects' out from the aggregates, since the C-terminal domain was previously shown to contain a non-native aggregation 'hot spot' that promotes inter-protein β-sheet formation (Sahin *et al.*, 2011). However, this is only speculation at this stage, and requires additional work to conclusively elucidate such mechanistic details.

Therefore, the examples of E17R and E17K highlight a potential pitfall when attempting to design for aggregation resistance by adding surface charges. That is, one must assure that aggregation mechanisms do not change as a result of the mutations imparted to an aggregation-prone protein. Otherwise, one risks causing unwanted effects in terms of both the aggregation rates and the physical form of aggregates.

It is important to reiterate that only single-amino acid substitutions at residue positions already occupied by residues containing titratable side chains were evaluated in this study. This approach was taken both to narrow the design space and ensure that only already solvent-exposed surface residues were changed. Switching a residue of vastly different properties (hydrophobicity, size) may have had a significantly greater effect on conformational stability. It has been shown previously that introducing multiple like-charged amino acids on a protein's surface ('supercharging') has the potential to impart significant aggregation resistance (Lawrence *et al.*, 2007; Simeonov *et al.*, 2011; Miklos *et al.*, 2012). However, loss of conformational stability and protein activity are also more likely with increased number of amino acid substitutions (Miklos *et al.*, 2012).

Summary and conclusions

A rational and predictive design approach based on electrostatic repulsions between protein monomers was tested here for the case of γDc, with the aid of coarse-grained molecular models. In order to minimize changes in conformational stability and folded-state structures, only single amino-acid substitutions were considered. This would presumably also be more amenable for use in combination with other design strategies that focus more heavily on protein function and activity, as it requires minimal changes to surface residues, and therefore may be more easily tolerated from the perspective of immune responses in patients (Baker *et al.*, 2010; van Beers and Bardor, 2012; Roberts, 2014).

The results show that altering a single charge on the surface of a protein can have a large effect on aggregation rates/aggregation resistance. Single-charge variants showed changes in aggregation rates of one or more orders of magnitude at low ionic strength. As expected, the results are much less pronounced at high ionic strength due to charge screening effects. As such, this strategy is most appropriate for use in improving aggregation resistance from the perspective of *in vitro* stability, manufacturing and produce shelf life. A potential pitfall in this design approach is highlighted via an outlier where the aggregation mechanism changed qualitatively (soluble vs. insoluble aggregates) due to switching from negative to positive charge at position E17. Independent of the relationship between aggregation rates and protein–protein interactions, a coarse-grained molecular modeling approach was also shown to provide semi-quantitative or quantitative predictions of how the net protein–protein interactions (i.e. osmotic second virial coefficients) change for even single-point mutations.

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgements

The authors thank Jonathan King (MIT) for the human γ-D crystallin gene.

Funding

This work was supported by the National Science Foundation (CBET-1264329), the National Institute of Standards and Technology (NIST-70NANB12H239) and the National Institutes of Health (R01-EB006006).

References

Acosta-Sampson,L. and King,J. (2010) *J. Mol. Biol.*, **401**, 134–152.
Andrews,J.M. and Roberts,C.J. (2007) *Biochemistry*, **46**, 7558–7571.
Baker,M., Reynolds,H.M., Lumicisi,B. and Bryson,C.J. (2010) *Self Nonself*, **1**, 314–322.
Barnett,G.V., Razinkov,V.I., Kerwin,B.A., Laue,T.M., Woodka,A.H., Butler, P.D., Perevozchikova,T. and Roberts,C.J. (2015) *J. Phys. Chem. B*, **119**, 5793–5804.
Basak,A., Bateman,O., Slingsby,C., Pande,A., Asherie,N., Ogun,O., Benedek, G.B. and Pande,J. (2003) *J. Mol. Biol.*, **328**, 1137–1147.
Benhar,I. and Pastan,I. (1995) *J. Biol. Chem.*, **270**, 23373–23380.
Ben-Naim,A. (1992) *Statistical Thermodynamics for Chemists and Biochemists*. Plenum Press, New York, NY.
Blanco,M. (2013) Protein–protein interactions and protein cluster formation from scattering experiments and coarse-grained molecular models. Univeristy of Delaware.
Blanco,M.A., Sahin,E., Li,Y. and Roberts,C.J. (2011) *J. Chem. Phys.*, **134**, 225103.
Blanco,M.A., Sahin,E., Robinson,A.S. and Roberts,C.J. (2013) *J. Phys. Chem. B*, **117**, 16013–16028.
Boder,E.T., Midelfort,K.S. and Wittrup,K.D. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 10701–10705.
Brummitt,R.K., Nesta,D.P., Chang,L., Chase,S.F., Laue,T.M. and Roberts,C.J. (2011a) *J. Pharm. Sci.*, **100**, 2087–2103.
Brummitt,R.K., Nesta,D.P., Chang,L., Kroetsch,A.M. and Roberts,C.J. (2011b) *J. Pharm. Sci.*, **100**, 2104–2119.
Brummitt,R.K., Nesta,D.P. and Roberts,C.J. (2011c) *J. Pharm. Sci.*, **100**, 4234–4243.
Brummitt,R.K., Andrews,J.M., Jordan,J.L., Fernandez,E.J. and Roberts,C.J. (2012) *Biophys. Chem.*, **168–169**, 10–18.
Brunsteiner,M., Flock,M. and Nidetzky,B. (2013) *PLoS One*, **8**, e59797.
Chiti,F., Webster,P., Taddei,N., Clark,A., Stefani,M., Ramponi,G. and Dobson, C.M. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 3590–3594.
Chiti,F., Calamai,M., Taddei,N., Stefani,M., Ramponi,G. and Dobson,C.M. (2002a) *Proc. Natl Acad. Sci. USA*, **99**, 16419–16426.
Chiti,F., Taddei,N., Baroni,F., Capanni,C., Stefani,M., Ramponi,G. and Dobson,C.M. (2002b) *Nat. Struct. Biol.*, **9**, 137–143.
Chiti,F., Stefani,M., Taddei,N., Ramponi,G. and Dobson,C.M. (2003) *Nature*, **424**, 805–808.
Clark,A.C., Sinclair,J.F. and Baldwin,T.O. (1993) *J. Biol. Chem.*, **268**, 10773–10779.
Costanzo,J.A., O'Brien,C.J., Tiller,K., Tamargo,E., Robinson,A.S., Roberts,C. J. and Fernandez,E.J. (2014) *Protein Eng. Des. Sel.*, **27**, 157–167.
Dantas,G., Kuhlman,B., Callender,D., Wong,M. and Baker,D. (2003) *J. Mol. Biol.*, **332**, 449–460.
de Groot,N.S., Aviles,F.X., Vendrell,J. and Ventura,S. (2006) *FEBS J.*, **273**, 658–668.
Dinner,A.R., Šali,A., Smith,L.J., Dobson,C.M. and Karplus,M. (2000) *Trends Biochem. Sci.*, **25**, 331–339.
Dobson,C.M. (2001) *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **356**, 133–145.
Dobson,C.M. (2004) *Semin. Cell Dev. Biol.*, **15**, 3–16.
DuBay,K.F., Pawar,A.P., Chiti,F., Zurdo,J., Dobson,C.M. and Vendruscolo,M. (2004) *J. Mol. Biol.*, **341**, 1317–1326.
Dunbrack,R.L. and Cohen,F.E. (1997) *Protein Sci.*, **6**, 1661–1681.
Ewert,S., Honegger,A. and Plückthun,A. (2003a) *Biochemistry*, **42**, 1517–1528.
Ewert,S., Huber,T., Honegger,A. and Plückthun,A. (2003b) *J. Mol. Biol.*, **325**, 531–553.
Famm,K., Hansen,L., Christ,D. and Winter,G. (2008) *J. Mol. Biol.*, **376**, 926–931.
Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) *Nat. Biotechnol.*, **22**, 1302–1306.
Flaugh,S.L., Kosinski-Collins,M.S. and King,J. (2005a) *Protein Sci.*, **14**, 565–581.
Flaugh,S.L., Kosinski-Collins,M.S. and King,J. (2005b) *Protein Sci.*, **14**, 2030–2043.
Flaugh,S.L., Mills,I.A. and King,J. (2006) *J. Biol. Chem.*, **281**, 30782–30793.
Grünberger,A., Lai,P.K., Blanco,M.A. and Roberts,C.J. (2013) *J. Phys. Chem. B*, **117**, 763–770.
Guerois,R., Nielsen,J.E. and Serrano,L. (2002) *J. Mol. Biol.*, **320**, 369–387.
Hu,X., Wang,H., Ke,H. and Kuhlman,B. (2008) *Structure*, **16**, 1799–1805.
Jiskoot,W., Randolph,T.W., Volkin,D.B., *et al.* (2010) *J. Pharm. Sci.*, **101**, 946–954.
Jordan,J.L., Arndt,J.W., Hanf,K., *et al.* (2009) *Proteins*, **77**, 832–841.
Kaufmann,K.W., Lemmon,G.H., DeLuca,S.L., Sheehan,J.H. and Meiler,J. (2010) *Biochemistry*, **49**, 2987–2998.
Kawahara,K. and Tanford,C. (1966) *J. Biol. Chem.*, **241**, 3228–3232.
Kim,N., Remmele,R.L., Liu,D., Razinkov,V.I., Fernandez,E.J. and Roberts,C.J. (2013) *Biophys. Chem.*, **172**, 26–36.
Kortemme,T., Morozov,A.V. and Baker,D. (2003) *J. Mol. Biol.*, **326**, 1239–1259.
Kosinski-Collins,M.S. and King,J. (2003) *Protein Sci.*, **12**, 480–490.
Kosinski-Collins,M.S., Flaugh,S.L. and King,J. (2004) *Protein Sci.*, **13**, 2223–2235.
Krebs,M.R., Wilkins,D.K., Chung,E.W., Pitkeathly,M.C., Chamberlain,A.K., Zurdo,J., Robinson,C.V. and Dobson,C.M. (2000) *J. Mol. Biol.*, **300**, 541–549.
Kroetsch,A.M., Sahin,E., Wang,H.Y., Krizman,S. and Roberts,C.J. (2012) *J. Pharm. Sci.*, **101**, 3651–3660.
Kuhlman,B. and Baker,D. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
Lawrence,M.S., Phillips,K.J. and Liu,D.R. (2007) *J. Am. Chem. Soc.*, **129**, 10110–10112.
Lazaridis,T. and Karplus,M. (1999) *Proteins*, **35**, 133–152.
Li,Y., Weiss,W.F. and Roberts,C.J. (2009) *J. Pharm. Sci.*, **98**, 3997–4016.
Li,Y., Ogunnaike,B.A. and Roberts,C.I. (2010) *J. Pharm. Sci.*, **99**, 645–662.
Liu,Y. and Kuhlman,B. (2006) *Nucleic Acids Res.*, **34**, W235–W238.
McConnell,A.D., Spasojevich,V., Macomber,J.L., *et al.* (2013) *Protein Eng. Des. Sel.*, **26**, 151–163.
Michaelson,J.S., Demarest,S.J., Miller,B.R., *et al.* (2009) *MAbs*, **1**, 128–141.
Miklos,A.E., Kluwe,C., Der,B.S., *et al.* (2012) *Chem. Biol.*, **19**, 449–455.
Miller,B.R., Glaser,S.M. and Demarest,S.J. (2009) *Methods Mol. Biol.*, **525**, 279–289, xiv.
Miller,B.R., Demarest,S.J., Lugovskoy,A., *et al.* (2010) *Protein Eng. Des. Sel.*, **23**, 549–557.
Mills,I.A., Flaugh,S.L., Kosinski-Collins,M.S. and King,J.A. (2007) *Protein Sci.*, **16**, 2427–2444.
Mohr,B.G., Dobson,C.M., Garman,S.C. and Muthukumar,M. (2013) *J. Chem. Phys.*, **139**, 121914.
Moreau,K.L. and King,J. (2009) *J. Biol. Chem.*, **284**, 33285–33295.
Moreau,K.L. and King,J.A. (2012) *PLoS One*, **7**, e37256.
Pace,C., Shirley,B. and Thomson,J. (1989) In Creightoned,T. (ed.), *Protein Structure: A Practical Approach*. IRL Press at Oxford University Press, New York, NY, pp. 311–330.
Pande,A., Pande,J., Asherie,N., Lomakin,A., Ogun,O., King,J.A., Lubsen,N.H., Walton,D. and Benedek,G.B. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 1993–1998.
Pande,A., Pande,J., Asherie,N., Lomakin,A., Ogun,O., King,J. and Benedek,G. B. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 6116–6120.
Papanikolopoulou,K., Mills-Henry,I., Tho,S.L., Wang,Y., Gross,A.A.R., Kirschner,D.A., Decatur,S.M. and King,J. (2008) *Mol. Vis.*, **14**, 81–89.
Perchiacca,J.M. and Tessier,P.M. (2012) *Annu. Rev. Chem. Biomol. Eng.*, **3**, 263–286.
Perchiacca,J.M., Ladiwala,A.R.A., Bhattacharya,M. and Tessier,P.M. (2012) *Protein Eng. Des. Sel.*, **25**, 591–601.
Privalov,P.L. (1979) *Adv. Protein Chem.*, **33**, 167–241.
Proba,K., Wörn,A., Honegger,A. and Plückthun,A. (1998) *J. Mol. Biol.*, **275**, 245–253.
Raghunathan,G., Sokalingam,S., Soundrarajan,N., Madan,B., Munussami,G. and Lee,S.G. (2013) *Mol. Biosyst.*, **9**, 2379–2389.

Roberts,C.J. (2007) *Biotechnol. Bioeng.*, **98**, 927–938.
Roberts,C.J. (2014) *Trends Biotechnol.*, **32**, 372–380.
Roberts,C.J. and Blanco,M.A. (2014) *J. Phys. Chem. B*, **118**, 12599–12611.
Roberts,C.J., Nesta,D.P. and Kim,N. (2013) *J. Pharm. Sci.*, **102**, 3556–3566.
Sahin,E., Grillo,A.O., Perkins,M.D. and Roberts,C.J. (2010) *J. Pharm. Sci.*, **99**, 4830–4848.
Sahin,E., Jordan,J.L., Spatara,M.L., Naranjo,A., Costanzo,J.A., Weiss,W.F., Robinson,A.S., Fernandez,E.J. and Roberts,C.J. (2011) *Biochemistry*, **50**, 628–639.
Sahin,E., Weiss,W.F., Kroetsch,A.M., King,K.R., Kessler,R.K., Das,T.K. and Roberts,C.J. (2012) *J. Pharm. Sci.*, **101**, 1678–1687.
Sanchez-Ruiz,J.M. (1992) *Biophys. J.*, **61**, 921–935.
Simeonov,P., Berger-Hoffmann,R., Hoffmann,R., Strater,N. and Zuchner,T. (2011) *Protein Eng. Des. Sel.*, **24**, 261–268.
Singh,J. and Kofke,D. (2004) *Phys. Rev. Lett.*, **92**, 220601.
van Beers,M.M.C. and Bardor,M. (2012) *Biotechnol. J.*, **7**, 1473–1484.
Velev,O.D., Kaler,E.W. and Lenhoff,A.M. (1998) *Biophys. J.*, **75**, 2682–2697.
Wang,W. (2005) *Int. J. Pharm.*, **289**, 1–30.
Wang,N., Smith,W.F., Miller,B.R., Aivazian,D., Lugovskoy,A.A., Reff,M.E., Glaser,S.M., Croner,L.J. and Demarest,S.J. (2009) *Proteins*, **76**, 99–114.
Wang,W., Nema,S. and Teagarden,D. (2010) *Int. J. Pharm.*, **390**, 89–99.
Weiss,W.F., Hodgdon,T.K., Kaler,E.W., Lenhoff,A.M. and Roberts,C.J. (2007) *Biophys. J.*, **93**, 4392–4403.
Weiss,W.F., Young,T.M. and Roberts,C.J. (2009) *J. Pharm. Sci.*, **98**, 1246–1277.
Wörn,A. and Plückthun,A. (2001) *J. Mol. Biol.*, **305**, 989–1010.

Appendix: Derivation of $E_{a,\mathrm{eff}}$ for Equation (3)

Starting with Equation (3), which is reproduced here as Equation (A1),

$$m(T) = \exp(-\exp(b(T - T_{\mathrm{ref}}))) \tag{A1}$$

the derivative with respect to T gives

$$\frac{\mathrm{d}m}{\mathrm{d}T} = -b\mathrm{e}^{b(T-T_{\mathrm{ref}})_m} \tag{A2}$$

TSML experiments best elucidate initial rate behavior, or where $m > 0.8$, where zeroth-order kinetics can be assumed and the effect of m on $\mathrm{d}m/\mathrm{d}T$ in Equation (A2) is minimal (Brummitt *et al.*, 2011c; Roberts *et al.*, 2013). Accounting for the (constant) scan rate, $\nu = \mathrm{d}T/\mathrm{d}t$, and taking the limit as $T \ll T_{\mathrm{ref}}$ or m approaches 1 gives

$$\lim_{m\to 1}\frac{\mathrm{d}m}{\mathrm{d}t} = \lim_{m\to 1}\frac{\mathrm{d}m}{\mathrm{d}T}\nu = -\nu b e^{b(T-T_{\mathrm{ref}})} \tag{A3}$$

where $\mathrm{d}m/\mathrm{d}t$ can be equated with the observed monomer loss, $-k_{\mathrm{obs}}$, within the limit of initial rate kinetics. Equation (A3) can be transformed to an Arrhenius-like form by multiplying and dividing the exponent by $T_{\mathrm{ref}}T$, giving

$$k_{\mathrm{obs}} = \nu b\mathrm{e}^{-bT_{\mathrm{ref}}T(T^{-1}-T_{\mathrm{ref}}^{-1})} \tag{A4}$$

Comparing Equation (A4) with the standard Arrhenius form

$$k_{\mathrm{obs}} = k_0 e - \frac{E_{a,\mathrm{eff}}}{R}(T^{-1} - T_{\mathrm{ref}}^{-1}) \tag{A5}$$

and simplifying $T_{\mathrm{ref}}T = T_{\mathrm{ref}}^2$ for the small range of temperatures of practical interests for protein solutions gives

$$E_{a,\mathrm{eff}} \approx bRT_{\mathrm{ref}}^2 \tag{A6}$$

and $k_0 = \nu b$. In Equation (A6), R is the gas constant and $E_{a,\mathrm{eff}}$ denotes the effective activation energy for k_{obs} for temperatures close to T_{ref}.