Investigating protein aggregation mechanism and evaluating mutational strategies to reduce aggregate formation

---

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Joseph Anthony Costanzo

December

2012

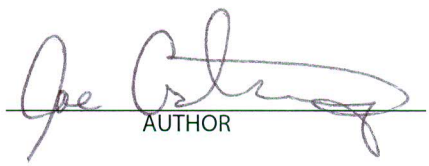
APPROVAL SHEET

The dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy



AUTHOR

The dissertation has been read and approved by the examining committee:

Erik Fernandez, Ph.D.

Advisor

Michael Shirts, Ph.D.

Inchan Kwon, Ph.D.

John O'Connell, Ph.D.

Christopher Roberts, Ph.D.

Christopher Stroupe, Ph.D.

Accepted for the School of Engineering and Applied Science:

Dean, School of Engineering and Applied Science

December

2012

Abstract

Non-native protein aggregation is a critical problem for biopharmaceuticals as it can compromise the biological activity and/or elicit an undesired immune response. Also, non-native aggregation can lead to amyloidosis, commonly associated with several neurodegenerative diseases. Thus, understanding the cause(s) of aggregation and developing tools or strategies to prevent aggregation are critical to improving human health.

A theoretical evaluation of mutational strategies that were intended to reduce protein aggregation by 1) conformationally stabilizing a single domain, 2) a domain interface, or 3) reducing the intrinsic aggregation propensity (IAP) of sub-sequences was conducted using the multi-domain protein, human γD crystallin (γD-crys). The IAP is defined here as the reactivity of sub-sequences to form intermolecular contacts that stabilize an aggregated state. The protein design program, RosettaDesign, and several empirical, sequence-based aggregation predictors were implemented to identify candidate variants, and nine variants were characterized experimentally. Afterwards, the effectiveness of the mutational strategies and computational design algorithms was assessed.

Given that only a small fraction of protein sequences are capable of folding, the observation that 3 of 9 candidate variants proved to be less aggregation-prone than wild type demonstrates promise for this general approach. The results suggested the IAP is another molecular property beyond conformational stability that needs to be considered in such protein design efforts. Further, each mutational strategy showed potential for deterring aggregation, and the computational algorithms demonstrated an *a priori* ability to identify aggregation-resistant variants for experimental evaluation. Improved success rates could make such design tools central to development of new biopharmaceuticals.

This work also utilized experimental and computational approaches to investigate the aggregation mechanism of γD-crys and A4V-human superoxide dismutase-1 (A4V-hSOD1), an hSOD1 variant associated with amyotrophic lateral sclerosis (ALS). For γD-crys, the aggregation of three variants displaying different aggregation behavior were examined relative to wild type using hydrogen-deuterium exchange with mass spectrometry (HX-MS). The more aggregation-resistant variants, H22T and S130P, formed flexible, less-structured aggregates; however, a more aggregation-prone variant, S130T, and wild type formed well-structured, amyloid-like aggregates. A potential aggregation contact within residues N125-L133 was identified both computationally and experimentally for all γD-crys species tested.

RosettaDesign was also utilized to investigate the aggregation mechanism of A4V-hSOD1 by identifying residues that could abolish much of the aggregation induced by the A4V variant. An intra-domain steric clash between residue F20 and V4 was observed in crystal structures of A4V-hSOD1, and was hypothesized to destabilize the protein and thereby instigate aggregation. RosettaDesign and the aggregation predictors identified F20G and F20A as candidate variants that could prevent the clash, and improve the conformational stability and/or the aggregation propensity of A4V-hSOD1. Experimental results showed F20A and F20G variants could indeed restabilize A4V-hSOD1 and reduce its aggregation. This result shows that eliminating the intra-domain steric clash is effective in reducing A4V-hSOD1 aggregation. Further, the correlation between computational design and experimental results demonstrates the potential of using these design algorithms to investigate protein aggregation mechanisms.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1.Background Information and Significance

Non-native protein aggregation

Proper manufacturing and commercialization of protein-based therapeutics involves a variety of challenging issues that must be understood. Specifically, protein aggregation is a common problem observed in nearly every biotechnology production or manufacturing process and often compromises the biological activity of the molecule [Vazquez-Rey 2011, Wang 2005, Mahler 2009, Weiss 2009, Cromwell 2006, Chi 2003]. Further, the formulation and shelf life of certain biotechnological therapeutics can also be affected because of protein instability and/or aggregation, even with the presence of excipients intended to increase stability [Kamerzell 2011]. Additionally, protein aggregation is associated with many protein deposition diseases in humans, also known as amyloidoses [Rousseau 2006, Murphy, 2002, Fink 1998]. Therefore, tools to address protein aggregation are needed in the context of both biopharmaceutical production as well as human disease.

Protein aggregation can incorporate many types of molecular interactions, originate from several mechanisms, be categorized as soluble or insoluble, reversible or irreversible, covalent or non-covalent, and adopt native or denatured conformations [Cromwell 2006]. This dissertation will focus specifically on non-native aggregation, defined as the irreversible formation (barring extreme conditions, i.e. elevated temperature, pressure, denaturant concentrations, etc.) [Weiss 2009, Chi 2003, Meersman 2006, Lefebvre 2004, Foguel 1999] of high molecular weight species caused by the protein adopting non-native contacts that ultimately stabilize an aggregated state [Sahin 2011].

As mentioned, aggregation can proceed through a variety of mechanisms and pathways such as via unfolded intermediates, self-association of native, folded species,

or even through chemical linkages or degradations [Wang 2010]. Several of these mechanisms are based upon a general model developed by Lumry and Eyring [Lumry & Eyring 1954], yet subsequent models dealing specifically with non-native aggregation [Roberts 2007] have evolved from the Lumry-Eyring framework. The proteins studied throughout this dissertation are hypothesized to aggregate via a partially unfolded, intermediate species formed along the unfolding pathway as seen in Figure 1.1. Here *N*, *I*, *U*, and *A* represent the native species, the aggregation-prone intermediate species, the fully unfolded or denatured species, and finally the aggregated species, respectively. The conformational changes between *N, I*, and *U* are considered reversible as represented by the double arrows in Figure 1.1. It should also be noted the single arrow shown between *I* and *A* in Figure 1.1 can be a combination of multiple steps, except in the limiting case where unfolding is itself rate-limiting [Andrews 2007].

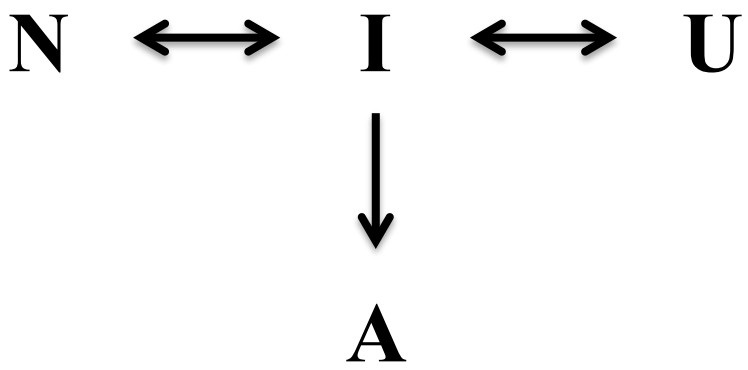


Figure 1.1: Depiction of an aggregation mechanism occurring via a partially unfolded intermediate species that appears along the unfolding pathway. Here, *N*, *I*, *U*, and *A* represent the native species, partially unfolded intermediate species, denatured species, and aggregated species, respectively. Also, the double and single arrows infer reversibility and irreversibility, respectively.

Proteins having undergone non-native aggregation (hereafter referred to as aggregation) often are observed to have increased levels of β-sheet, secondary structures [Weiss 2009], such as the cross-β amyloid structure commonly associated with many protein deposition diseases in humans, also known as amyloidoses

[Rousseau 2006, Murphy, 2002, Fink 1998]. A familiar example is the fibrillation of the β-amyloid peptide (Aβ) within neuron cells in the human brain, largely considered to result in Alzheimer's disease (AD) [Sanchez de Groot 2006]. Another recognized amyloidogenic protein is the extracellular, homotetrameric protein known as transthyretin (TTR). The misfolding and misassembly of TTR is commonly associated with familial amyloid polyneuropathy (FAP) [Du 2010]. Finally, another well-known neurodegenerative disorder is amyotrophic lateral sclerosis (ALS or Lou Gehrig's disease). The majority of ALS cases appear sporadically, however familial forms of ALS (fALS) also appear and are caused by genetic mutations occurring in a number of proteins; but the best studied protein is the homodimeric, enzyme human copper-zinc superoxide dismutase-1 (hSOD1) [Chattopadhyay 2009]. Many consider a cause of fALS to be the misfolding of hSOD1 and subsequent accumulation of hSOD1-rich proteinaceous deposits in the brain and spinal cord of humans [Chattopadhyay 2009, Stathopulos 2003, Cardoso 2002]. These non-native interactions are caused by a variety of hSOD1 variants, but the most frequently occurring fALS variant, also causing the most rapid disease progression, is the alanine to valine substitution at the fourth residue (A4V) [Cardoso 2002].

Additionally, protein denaturation or aggregation can also induce undesired immune responses in patients [Wang 2012, De Groot 2007]. The role of protein aggregates eliciting an immune response to administered protein therapeutic products has been well documented for some time, particularly for immunoglobulins and human growth hormone (hGH) [De Groot 2007]. Further, the reduction of protein aggregation has been linked to decreased immunogenicity [Sauerborn 2010]. For instance, early studies involving doses of human γ-globulin containing aggregates in mice observed enhanced immunogenicity relative to the aggregate-free dosages [Gamble 1966]. Additionally, clinical studies in humans showed the removal of γ-globulin aggregates

also reduced immunogenicity [Weksler 1970]. Another early clinical study involving hGH aggregates also showed an induced immune response in children as well [Moore 1980].

Thus, these detriments of protein aggregation as related to biotechnological manufacturing processes, human disease, and/or undesired immunogenicity in patients have recently piqued interest within biopharmaceutical research to better understand the cause(s) of protein aggregation and limit aggregate formation [Wang 2010]. Two general approaches to limiting protein aggregation are 1) to live with the molecule, and adjust formulation and/or process conditions, and 2) to modify the molecule itself via protein engineering and design. Following the first approach avoids making changes to the target molecule that could alter the biofunctionality of the therapeutic, require further clinical trials, and approval from federal health associations. On the other hand, applying the second approach, particularly only single point mutations, may prevent aggregate formation within certain process steps and avoid additional, costly downstream processing operations intended to remove aggregates generated from the process itself.

In regards to the first approach, it has been well established that biotechnological manufacturing can utilize general strategies to limit aggregation ranging from upstream cell culture processes through downstream recovery, purification, formulation, and storage [Remmele 1999, Webb 2001, Cromwell 2006]. For instance, during cell culture and protein expression, improper protein folding and subsequent intracellular aggregation can occur via incorrect disulfide formation of free, unpaired thiols [Frand 2000]. As such, processes can add oxidizing agents to the growth medium to encourage correct disulfide formation [Zhang 2002]; nonetheless certain constituents can still expose the protein to conditions that favor protein instability [Chi 2003]. Additionally, subsequent purification techniques employed during biotechnological manufacturing require a wide range of solution conditions (e.g. pH, ionic strength, protein concentration, etc.) to maximize purification yields, all of which can affect the extent of aggregate

formation [Gagnon 2012, Cromwell 2006]. As such, this often leads to incorporating additional purification steps to remove aggregates, such as ion exchange (IEC), size exclusion chromatography (SEC), or filtration. However, these additional unit operations can also be costly, inefficient, and subject the protein to harmful, undesired mechanical stresses promoting further aggregation [Cromwell 2006, Aldington 2007, van Reis 2001, Wan 2005]. Finally, protein aggregation can be limited in formulation by the use of excipients [Kamerzell 2011], but also in the final filling or storage steps by optimizing the effects of solution viscosity as well as the interactions of the final therapeutic product with the storage or filling materials [Kamerzell 2011, Cromwell 2006].

Protein engineering: An alternative approach to reducing aggregation

On the other hand, using protein engineering to modify the molecule itself is an orthogonal approach to increasing the biofunctionality or conformational stability of a protein as well as reducing protein aggregation. A familiar example of utilizing protein engineering is the widely manufactured therapeutic human insulin analog, Lyspro. Results have shown the transposition of a lysine (Lys) and proline (Pro) residue within the native insulin sequence enhances biofunctionality of the molecule, observed by faster absorption of the therapeutic into the bloodstream relative to wild type insulin [Trautmann 1994]. Notably however, certain *in vitro* studies have also shown the transposition of the Lys and Pro residues increased aggregation of the insulin analog relative to wild type [Ludwig 2011]. Therefore, this example illustrates protein engineering techniques have been used successfully for manufactured therapeutic products; but also the inherent difficulty associated with finding variants that will deter protein aggregation and also maintain the physiological function of the molecule as well.

Investigators trying to improve protein stability and deter aggregation via mutational strategies have used several techniques to help choose candidate variants

such as rational or structure-based methods [Eijsink 2004, Lehmann 2001, Wetzel 1994, Chrunyk 1993], directed evolution techniques [Lehmann 2001, Eijsink 2005], and consensus design [Lehmann 2001, Forrer 2004]. However, directed evolution involves a substantial amount of time and experimental resources, while consensus design requires the sequence identification of many homologous proteins as well as criteria for selecting a specific variant when a clearly favorable option at a specific location is lacking [Bannen 2008].

On the other hand, there are many examples in literature where rational or structural-based mutational designs resulted in enhanced protein stability [Eijsink 2004, Fernandez-Lafuente 2009, Goihberg 2008, Berezovsky 2005, Gerk 2000, Strickler 2006, Schwehm 2003, Williams 1999, Makhatadze 2003, Melnik 2012]. Further, it has also been shown that some of these design strategies can stabilize the protein against irreversible processes such as aggregation [Logan 2010, Ray 2004, Sekijima 2006]. Nevertheless, at least two significant challenges to rational design remain: (1) the tertiary structure of the protein must be known, and (2) the lack of a reliable process for identifying the specific variant(s) out of vast possibilities that will best improve the desired property of the protein [Eijsink 2004, Bannen 2008].

Computational design: A promising new approach

However, implementing computational design methods [Das 2008, Cellmer 2007, Bratko 2007, Saven 2010] in conjunction with rational design may address the latter of these challenges. There are numerous investigations and reviews reported in literature where successes in computational protein re-engineering or *de novo* design have been established to improve conformational stability [Das 2008, Bratko 2007, Saven 2010, Dantas 2003, Shah 2007, Hu 2008, Schueler-Furman 2005, Tian 2010, Lu 2009]. However, the application of such computational algorithms to minimizing protein

aggregation has been extremely limited thus far [Sahin 2011, Miklos 2012]. Furthermore, the aforementioned studies selected variants using directed evolution [Worn 1998] or were based on knowledge gathered from previous investigations [Chrunyk 1993]. As such, selecting variants that would limit protein aggregation using computational design tools *a priori* would be valuable.

Throughout this dissertation, one such computational tool used was RosettaDesign, the design module of the molecular modeling program Rosetta [Rohl 2004]. It has been used to redesign globular proteins [Miklos, 2012, Sahin 2011, Dantas 2003, Kuhlman 2003], enzymes [Jiang 2008], as well as protein-protein interfaces [Sammond 2007] with reasonable success rates. RosettaDesign utilizes an energy function to estimate a final score relatable to the free energy of folding of the molecule. It does not evaluate the free energy of unfolding, but rather the difference in free energy of folding for mutants relative to wild type. The energy function (Eq. 1.1),

$$\Delta\Delta G_f = \Sigma E_{LJ} + \Sigma E_{HB} + \Sigma E_{SS} + \Sigma E_{tors} + E_{sol} + E_{pair} + E_{ref} \qquad \text{(Eq. 1.1)}$$

is comprised of terms including attractive and repulsive Lennard-Jones potentials (E_{LJ}), the Lazaridis-Karplus implicit solvation model (E_{sol}) [Lazaridis 1999], orientation dependent hydrogen bonding potentials (E_{HB}) [Kortemme 2003], intramolecular disulfide bonding terms (E_{SS}), torsion potentials estimated from backbone and side chain realignment (E_{tors}) [Dunbrack 1997], an electrostatic term accounting for interactions of charged residues (E_{pair}), and finally a reference term (E_{ref}) unique to each of the twenty amino acids used to control their abundance and favorability within the primary sequence [Hu 2008]. The Lennard-Jones potential favors tightly packed atoms, but also provides the steric information required to correctly pack the protein core while preventing steric clashes [Dantas 2003]. The implicit solvation model rewards the packing of hydrophobic amino acids into the core of the molecule and hydrophilic

residues on the protein surface; however, the hydrogen bonding potential can offset the solvation model and reward buried polar molecules that are beneficial to the hydrogen-bonding network [Dantas 2003]. Given a specific design command and starting structure, RosettaDesign will estimate the lowest energy score via Monte Carlo optimization with simulated annealing and search for amino acid sequences that pack well while avoiding steric clashes, burying hydrophobic residues, and maintaining the hydrogen bonding network. Amino acid variants or rotamer realignments are then accepted or rejected based on the Metropolis criterion [Liu 2006].

In addition stabilizing protein native states, there is also the idea that proteins ultimately aggregate by attractive molecular interactions at specific (e.g. amyloid forming segments) or nonspecific (e.g. hydrophobic) segments of polypeptide sequence. These interactions could take place between unfolded or folded polypeptides; however, the simplest and most well defined to measure are kinetics of aggregation of completely unfolded polypeptide sequences. We term the aggregation propensity for such peptide sequences, intrinsic aggregation propensity (IAP). However, there is no agreement in the literature on the definition of such a quantity, and questions remain on the correct basis for the quantity (e.g. solubility, kinetics, etc.), as well as, for example, how conditions should be defined if kinetics is the basis. Despite this difficulty, several sequence-based, empirical aggregation calculators have been developed to correlate peptide sequence to peptide solubility or kinetic data on peptide aggregation. Several were implemented throughout this work to identify sub-sequences of proteins prone to aggregate (referred to as “hot spots”) into cross β-sheet, amyloid-like structures [Caflisch 2006, Tartaglia 2005, Fernandez-Escamilla 2004, Conchillo-Sole 2007, Trovato 2007, Thompson 2006]. These calculators are particularly useful because of their public accessibility as well as their ability to qualitatively predict “hot spots” or changes in the IAP as a function of mutation.

The majority of these programs use phenomenological or informatic analysis of structural databases, physicochemical properties of side chains, and/or molecular simulations that statistically rank the probability of short polypeptide sequences to aggregate [Caflisch 2006]. Nearly all are based on a multivariate statistical regression against experimental aggregation rates for small polypeptides, most of which do not form stable secondary or tertiary structure as isolated monomers, but do readily aggregate [Caflisch 2006]. Therefore, these algorithms provide can a statistical or structural approach to identify changes in the IAP as a function of mutation. They have been shown with reasonable success to predict relatively short aggregation-prone sub-sequences, but the validation of these algorithms to predict "hot spots" in foldable proteins is more limited. Despite this, some correlations have been observed between the predictions of these calculators for proteins as well as peptides and actual experimental results [Sahin 2011, Zhang 2010, Ivanova 2006, Routledge 2009].

Thus far, no single algorithm has clearly demonstrated superiority when compared to other available algorithms, and the quantitative values outputted by the algorithms are not necessarily comparable. Further, they lack predictive capability for aggregates lacking significant β-sheet structure, and some do not fully account for environmental factors that may induce aggregation [Ebrahim-Habibi 2010]. They also cannot account for through space interactions with other distant parts of sequence, or the effects of tertiary structure or partially folded tertiary structure. As such, in this work predicted "hot spots" were predicted by a consensus agreement between the majority of algorithms to eliminate inherent biases caused by varying derivations or parameterizations of each algorithm. Four aggregation calculators were implemented for this work known as TANGO [Fernandez-Escamilla 2004], AGGRESCAN [Conchillo-Sole 2007], PASTA [Trovato 2007], and the 3D profiling method [Thompson 2006].

1.2. Objectives

A primary goal of this dissertation was to test how well the aforementioned computational tools can be used in tandem to identify aggregation resistant point variants within two multi-domain protein systems using three mutational strategies. To our knowledge, no prior studies have utilized these computational design tools in tandem and these mutational strategies to successfully identify single, point variants exhibiting improved aggregation relative to the wild type for multi-domain proteins, or to consider systematically or quantitatively how well they work. Therefore, this work will not only provide insight into mutational strategies that can be used to stabilize proteins, particularly multi-domain proteins which are important to human disease and biopharmaceutical development; but it will also provide a valuable evaluation of these specific computational tools and help identify the kinds of improvements needed for refining and assessing such tools in the future.

The overall success rates of these computational tools for identifying aggregation-resistant variants were evaluated by comparing the computational predictions to the experimentally observed results. Notably, determining if the success rates for these computational design tools are adequate is somewhat subjective as compelling benchmarks or metrics have not yet been determined regarding what are meaningful predictive values for protein designs or acceptable predictive yields for each design algorithm. Further, highly *quantitative*-based correlations between predicted and experimental results were not expected, but were desired, in this work because 1) the aggregation calculators are currently incapable of predicting a quantitative change in the IAP of a protein, and 2) most of the computational tools only predict changes in conformational stability or IAP at physiological pH, but many of the experimental studies, particularly for γD-crys, were conducted at acidic pH where aggregates remained soluble.

Within Chapter 2, γD crystallin (γD-crys), was used as a model protein system to theoretically assess three mutational strategies intended to deter non-native aggregation, specifically for multi-domain proteins: 1) by stabilizing the less stable domain, 2) by stabilizing the interface between domains and 3) by mutating aggregation-prone sub-sequences while maintaining conformational stability. RosettaDesign and the various aggregation calculators identified nine candidate variants, and each was subsequently examined experimentally to characterize their conformational stability and aggregation behavior relative to wild type γD-crys, and to examine the effectiveness of the computational tools. Further, altered conformational stability and intrinsic aggregation propensity (IAP) of the molecule, as a function of mutation, were both evaluated as potential contributors to protein aggregation. Here, conformational stability is defined as the change in unfolding free energy ($\Delta\Delta G^{\circ}_{unf}$) from a folded monomer to a partially unfolded, aggregation-prone, intermediate state. On the other hand, IAP can be interpreted as the reactivity of specific stretches of sequence within a monomeric chain to form intermolecular contacts (e.g. β-sheet formation) ultimately stabilizing an aggregated state [Sahin 2011].

Next, in Chapter 3 the aggregate conformations of selected γD-crys variants, displaying a diverse range of aggregation behavior, were characterized using the experimental technique known as hydrogen-exchange coupled with enzymatic digestion and mass spectrometry (HX-MS). This technique was utilized to determine changes in a protein structure, perhaps caused by partial unfolding or aggregation, and identify residues or regions of a protein that may possibly serve as aggregation contacts. Aggregation-prone “hot spots” were also identified computationally, using several aggregation calculators and then compared to the experimental results. From these results, the aggregation mechanisms of the different γD-crys species were compared,

and potentially could help substantiate the previously described mutational strategies discussed in Chapter 2, or identify new strategies to better reduce protein aggregation.

Finally, in Chapter 4 of this dissertation the computational design tools were applied to another protein system, human copper-zinc superoxide dismutase-1 (hSOD1), whose aggregation is related to the neurodegenerative disease, amyotrophic lateral sclerosis (ALS). The goals for this work were to investigate the aggregation mechanism involving the severely destabilizing, ALS-associated variant A4V in hSOD1 (A4V-hSOD1); and to test if the aggregation mechanism involved inter-domain or intra-domain conformational destabilization as a result of the variant. To accomplish this, RosettaDesign was utilized to identify candidate second-site variants that may repress the conformationally destabilizing effects of A4V-hSOD1. The term *second-site variant* is used for these variants because when combined with the A4V variant, an hSOD1 double variant, relative to wild type hSOD1 was generated. In addition, the various aggregation calculators were implemented to investigate if the selected second-site variants altered the IAP of A4V-hSOD1, and to identify potential aggregation-prone sub-sequences of A4V-hSOD1 for redesign. Afterwards, certain variants were analyzed experimentally to characterize their conformational stability and aggregation behavior relative to wild type hSOD1, and subsequently compared with the computational predictions.

Chapter 2: A computational design shows both conformational stability and predicted aggregation propensity contribute to non-native protein aggregation

2.1. Introduction

Non-native protein aggregation is a problem observed in numerous biotechnology production processes, and can often compromise the biological activity of the molecule [Vazquez-Rey 2011, Mahler 2009, Weiss 2009, Cromwell, 2006, Wang 2005, Chi 2003]. Non-native aggregation (hereafter referred to as aggregation) can be defined as the irreversible formation (barring extreme conditions, i.e. elevated temperature, pressure, denaturant concentrations, etc.) [Weiss 2009, Chi 2003, Meersman 2006, Lefebvre 2004, Foguel 1999] of high molecular weight species caused from non-native intra(inter) molecular contacts. Proteins having undergone non-native aggregation often are observed to have increased levels of β-sheet, secondary structures [Weiss 2009], such as the cross-beta amyloid structure commonly associated with many protein deposition diseases [Rousseau 2006, Murphy 2002, Fink 1998]. Additionally, aggregation can also induce undesired immune responses in patients [De Groot 2007, Rosenberg 2003]. Thus, understanding the cause(s) of aggregation and developing tools or strategies to prevent aggregation have recently been of interest in the context of biopharmaceutical development [Wang 2010] as well as human disease [Rousseau 2006, Murphy 2002].

As previously mentioned, non-native aggregation can proceed through partially or fully unfolded intermediate species formed along the unfolding pathway such as seen in Figure 2.1

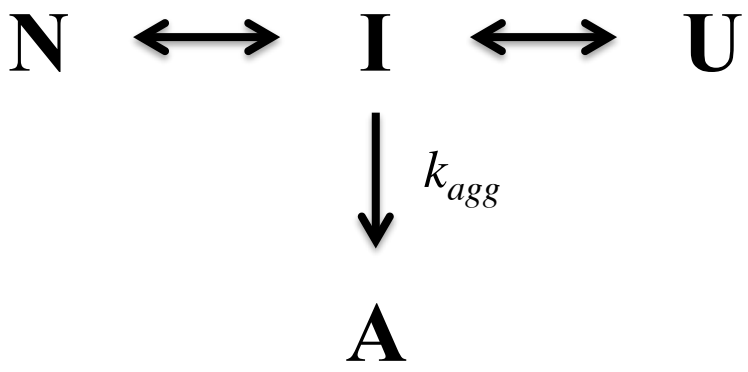


Figure 2.1: The proposed aggregation mechanism of the model, multi-domain protein studied in this work, γD-crys, where aggregation proceeds through a partially folded intermediate species. The mechanism and defined states are identical to that shown in Figure 1.1. However, here the variable k_{agg} is defined, representing the observed, initial aggregate rate coefficient estimated from the native species to the aggregate state.

This aggregation scheme and all of the defined states (e.g. *N*, *I*, *U*, and *A*) are identical to the mechanism and states shown in Figure 1.1, however, here an observed, initial aggregation rate coefficient, k_{agg}, is also defined, and estimated from the native species to the aggregate state. The aggregation rate coefficient specifically describing aggregation from *I* to *A* cannot be directly estimated on its own because the intermediate species is incapable of being observed experimentally. Also notable for Figure 2.1, the single arrow shown between *I* and *A* can be a combination of multiple steps, except in the limiting case where unfolding is itself rate-limiting [Andrews 2007]. Consequently, incorporating mutational design strategies intending to improve the conformational stability of the native species, *N,* or reduce the intrinsic aggregation propensity (IAP) of the aggregation-prone species, *I*, could both conceivably reduce k_{agg} [Weiss 2009]. Here, conformational stability is defined as the unfolding free energy ($\Delta\Delta G^{\circ}_{unf}$) from a folded monomer to a partially unfolded, aggregation-prone, intermediate state. On the other hand, IAP can be interpreted as the reactivity of specific stretches of sequence within a monomeric chain to form intermolecular contacts (e.g. *β*-sheet formation) ultimately stabilizing an aggregated state [Sahin 2011].

Investigators striving to improve protein conformational stability and deter protein aggregation via mutational strategies have used rational or structural-based methods [Eijsink 2004, Lehmann 2001, Wetzel 1994, Chrunyk 1993], directed evolution techniques [Lehmann 2001, Eijsink 2005], as well as consensus design [Lehmann 2001, Forrer 2004]. However, directed evolution involves a substantial amount of time and experimental resources, while consensus design requires identification of many homologous proteins as well as criteria for selecting a specific variant when a clearly favorable option at a specific location is lacking [Bannen 2008].

On the other hand, there are many examples in literature where rational or structural-based mutational designs resulted in enhanced protein stability [Eijsink 2004, Fernandez-Lafuente 2009, Goihberg 2008, Berezovsky 2005, Gerk 2000, Strickler 2006, Schwehm 2003, Williams 1999, Makhatadze 2003, Melnik 2012]. Further, it has also been shown that some of these design strategies can stabilize the protein against irreversible processes such as aggregation [Logan 2010, Ray 2004, Sekijima 2006]. Nevertheless, at least two significant challenges to rational design remain: (1) the tertiary structure of the protein must be known, and (2) the lack of a reliable process for identifying the best variant(s) out of vast possibilities to best improve the desired property of the protein [Eijsink 2004, Bannen 2008].

However, implementing computational design methods [Das 2008, Cellmer 2007, Bratko 2007, Saven 2010] in conjunction with rational design methods may address the latter of these challenges. There are numerous investigations and reviews reported in literature where successes in computational protein re-engineering or *de novo* design have been established to improve conformational stability [Das 2008, Bratko 2007, Saven 2010, Dantas 2003, Shah 2007, Hu 2008, Schueler-Furman 2005, Tian 2010, Lu 2009]. However, the application of such computational algorithms to minimizing protein aggregation has been extremely limited thus far [Sahin 2011, Miklos 2012]. Furthermore,

the aforementioned studies where variants were selected using directed evolution [Worn 1998] or based on knowledge gathered from previous investigations [Chrunyk 1993], but for this work, variants were all identified via computational tools *a priori* to any conformational stability or aggregation assays being experimentally conducted.

In this work, a primary objective was to test the utility of the previously described computational design tools to identify single, point variants that could reduce the aggregation of a model two-domain protein. To do so, three mutational strategies intended to reduce aggregation were evaluated in this work, including (1) stabilizing the less stable domain, (2) stabilizing the domain-domain interface, and (3) mutating aggregation-prone sub-sequences while maintaining conformational stability. To our knowledge, no prior studies have utilized these computational design tools in tandem and these mutational strategies to successfully identify point variants that exhibit improved aggregation relative to the wild type for multi-domain proteins, or to consider systematically or quantitatively how well they work. Therefore, this work should provide a valuable evaluation of these specific tools and help identify the kinds of improvements needed for refining and assessing such tools in the future.

Variants from the first two strategies were identified using RosettaDesign, while variants from the third mutational strategy were predicted from the aforementioned aggregation calculators and tested with RosettaDesign to assure they were not also significantly destabilizing to the folded monomer structure. Candidate variants were characterized experimentally to estimate their unfolding free energy and accelerated (high temperature) aggregation rate coefficients. The experimental results were then analyzed to determine if a correlation existed between conformational stability and aggregation rates as was observed in previous studies [Wetzel 1994, Worn 1999]. Further, success rates for the computational design tools were reported on a qualitative basis to determine the effectiveness of computational design tools. Notably, the success

rates for these computational design tools are somewhat subjective as compelling benchmarks or metrics have not yet been determined regarding what are meaningful predictive values or acceptable predictive yields.

The model protein chosen for this work was the 21 kDa (173 residue) human eye lens protein γD crystallin (γD-crys), a multi-domain protein rich in antiparallel β-sheets arranged into four Greek-key motifs as depicted in Figure 2.2.

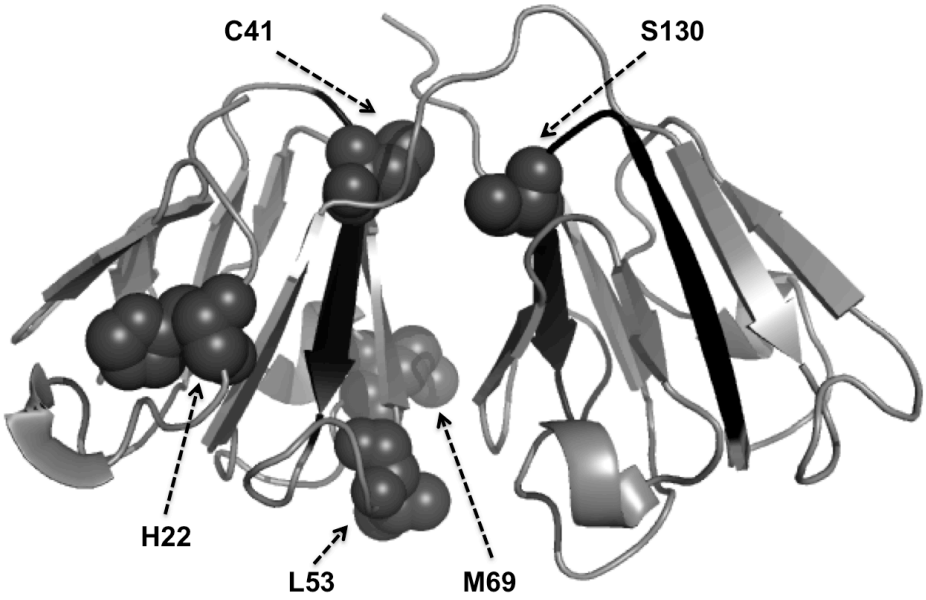


Figure 2.2: Crystal structure of γD-crys illustrating each variant site (spheres) as well as the aggregation-prone "hot spots" identified by a majority of the aggregation calculators (black ribbon).

Crystallins are expressed early in life and must remain soluble and transparent throughout an average human life span; indicating their extreme stability despite their high concentrations in the lens and continued exposure to environmental conditions that may induce aggregation [Kosinski-Collins 2003]. Upon aggregation γD-crys is associated with cataracts, and furthermore, several point variants within the gene

sequence encoding γD-crys have been related to early-onset cataracts [Jung 2009, Pande 2000].

Along with its physiological importance, γD-crys also serves as an appealing model protein for investigating aggregation resistance in reference to multi-domain, initially folded proteins. The structure of the molecule is smaller yet similar to antibody-based pharmaceuticals in development that are also known to experience aggregation problems [Wang 2005, Crabbe 1995]. Furthermore, the aggregation and folding behavior of γD-crys has been extensively studied and is well established at various conditions [Kosinski-Collins 2003, Flaugh 2005a, Flaugh 2005b, Flaugh 2006, Kosinski-Collins 2004, Mills 2007]. Previous studies have observed aggregation of γD-crys during refolding from elevated denaturant concentrations proceeds through a partially folded intermediate species, depicted as *I* in Scheme 1, comprised of an unfolded N-terminal domain (N-td) and a folded C-terminal domain (C-td) [Flaugh 2005a, Flaugh 2005b, Flaugh 2006]. In a previous study involving γD-crys we reported on two aggregation-reducing variants, M69Q and S130P, predicted by computational tools to increase the conformational stability and decrease the IAP, respectively. In this work, we present a broader assessment of RosettaDesign and the aggregation calculators applied using the three design strategies listed above.

2.2. Materials and Methods

Selection of candidate variants via RosettaDesign

The fixed backbone protocol in RosettaDesign 3.0 was utilized for identifying variants of human γD-crys (pdb code: 1HK0) where only the side chains of each amino acid are permitted to move. First, specific regions of the molecule such as the less stable N-td and the interface between domains were globally redesigned to assemble a list of potential variants. Next, point mutation scans were run for each variant identified in the

global redesign runs. Those variants identified to potentially stabilize the N-td or interface region were selected based upon energy scores estimating the change in folding free energies, $\Delta\Delta G_f$, and the change in binding free energies, $\Delta\Delta G_{bind}$, respectively, for each variant relative to wild type. Values for these parameters were estimated using Eq. 2.1 and 2.2, respectively. ΔG_{bind} values were determined by calculating the difference in the folding free energy of the entire molecular complex and the sum of the folding free energies for the isolated N-td and C-td (Eq. 2.3).

$$\Delta\Delta G_f = \Delta G_f^{\text{var}} - \Delta G_f^{wt} \qquad \text{(Eq. 2.1)}$$

$$\Delta\Delta G_{bind} = \Delta G_{bind}^{\text{var}} - \Delta G_{bind}^{wt} \qquad \text{(Eq. 2.2)}$$

$$\Delta G_{bind} = \Delta G_f - \left(\Delta G_f^{Ntd} + \Delta G_f^{Ctd}\right) \qquad \text{(Eq. 2.3)}$$

Candidate variants located within the N-td and identified by RosettaDesign to lower $\Delta\Delta G_f$ relative to wild type were considered (Eq. 2.1). In addition, candidate variants that broke down the hydrogen-bonding network (ΔE_{HB}) or Lennard-Jones (ΔE_{LJ}) contributions within the overall energy score (ΔE_{LJ} and $\Delta E_{HB} \leq 0$) were disqualified from consideration. Furthermore, variants within 4 Å of the interface between domains were required to favorably affect binding ($\Delta\Delta G_{bind} < 0$). Variants passing these filtering metrics were then combined with other variants identified by aggregation calculators to be located within aggregation-prone regions to create a final pool of approximately thirty variants. The final nine variants were chosen to populate the three mutational strategies, with an added criterion requiring each mutated residue to possess partial solvent protection within the tertiary structure in order to better impact conformational stability. Additional information regarding the command line or resfile syntax for RosettaDesign 3.0 can be found in Appendix A.

Expression and purification of γD-crys

Frozen stocks of *E. coli* BL21 (DE3) strains carrying pET-15-γD-crys-WT as well as various other pET-15-γD-crys-variant plasmids were grown overnight on LB media containing 2% (w/v) agar. Individual colonies for WT and each variant were selected and grown overnight as primary cultures in 100 mL of super broth (SB) liquid media in 250 mL baffled flasks at 37 °C and 250 rpm. Primary cultures were diluted into fresh SB liquid media in 2 L flasks to obtain secondary cultures with a final optical density of 0.05 at 600 nm (OD_{600} = 0.05). Secondary cultures were grown at 30 °C and 250 rpm until an OD600 was approximately 0.8, at which point isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM. The IPTG-induced cultures were then further grown at 30 °C and 250 rpm for 24 hours. All solid and liquid media contained 100 μg/ml ampicillin to maintain selectivity.

E. coli cells were harvested by centrifugation at 3500 rpm for 15 minutes. The cell pellets were then weighed and frozen at -80 °C for storage and to aid cell lysis. Frozen pellets were re-suspended and homogenized in 2-5 mL of chilled buffer per gram of cell pellet for cell lysis. The lysis buffer was comprised of 50 mM NaH_2PO_4, 300 mM NaCl, and 10 mM imidazole, adjusted to pH 8.0. The re-suspended cells were then divided into 15 mL Falcon tubes each containing approximately 5-7 mL. The solution in each Falcon tube was sonicated two times each for 60 sec with 10 sec cooling periods at 10% amplitude to ensure adequate lysing but prevent the lysate from reaching high temperatures. The lysate solutions were then recombined and centrifuged for 30 minutes at 10,000 x g and 4 °C. The supernatant was then decanted and 1 mL of 50% Ni-NTA agarose beads (Qiagen) per 4 mL of supernatant was added and rotated end on end overnight at 4 °C.

After affinity adsorption, the Ni-NTA beads were washed four times with chilled wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, adjusted to pH 8.0 with

NaOH) via centrifugation at 3500 rpm for 5 minutes. The Ni-NTA beads were re-suspended in wash buffer and transferred into a Bio-Rad Econo-Pac gravity chromatography column where the final wash step took place. Once the wash buffer entered the column, 3 column volumes (CV) of elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, adjusted to pH 8.0 with NaOH) were added to the gravity column. The elution buffer was allowed to flow through the column and three fractions were collected, the second being the protein-rich fraction which was immediately dialyzed using a 10 kDa dialysis cassette (Thermo) against 2 x 0.5 L of 50 mM citrate buffer, adjusted to pH 3.0 with 5 M NaOH. After dialysis, protein concentration was determined using UV absorbance at 280 nm and samples were aliquoted and frozen at -80 °C for storage.

Equilibrium chemical unfolding

Chemical denaturation experiments were performed using high purity urea (Fisher Scientific) or guanidine hydrochloride (GdnHCl) (MP Biomedical) as denaturants. For experiments conducted at pH 3 and 7, samples of 0.2 mg/ml wild type or γD-crys variants were prepared in 50 mM citrate buffer or 50 mM phosphate buffer, respectively, and then diluted 10-fold into samples containing denaturant concentrations ranging from 0 to 5 M urea or GdnHCl. Samples were incubated for 24 hours at room temperature to ensure equilibrium had been reached. Fluorescence data was collected at room temperature using a Jobin Yvon Horiba FluoroMax-3 Spectrofluorometer. Excitation and emission slit widths of 1.0 and 2.0 mm, respectively, were used along with an excitation wavelength of 295 nm, and emission intensity was recorded for wavelengths between 310 to 450 nm. The ratio of baseline-corrected emission intensities at 360 and 320 nm was used for analysis. To estimate thermodynamic parameters, all equilibrium unfolding

data were globally regressed to a two-state unfolding model using MATLAB software [Pace 1989].

Isothermal aggregation and analysis using size exclusion chromatography

Aggregation of wild type and γD-crys variant species were monitored as a function of incubation time at a constant elevated temperature of 50 °C using samples with an initial protein concentration of approximately 1 mg/ml in 50 mM citrate buffer, pH 3. At these conditions, γD-crys aggregates grow without the formation of insoluble precipitates, the conformational stability of γD-crys is greatly reduced, and thus aggregation occurs much more easily than at physiological pH. Incubation was conducted in upright, 2 mL, glass HPLC vials with PTFE/silicone caps (Fisher Scientific) in a water bath with negligible temperature variability. Individual samples were removed at various incubation times and immediately placed in an ice water bath to quench further aggregation. No additional aggregation or change in monomer fraction was observed between sample quenching and analysis. Aggregated samples were analyzed using size exclusion chromatography (SEC) with a mobile phase comprised of 0.5% phosphoric acid, adjusted to pH 2.7 with 5 M NaOH and operated at 0.8 mL/min. Sample volumes of 100 µL were injected into a TSKgel Guard SWxl column attached in series with a TSKgel G2000SWxl analytical column (TOSOH 7.8 x 30.0 cm, 5 µm) connected to a Waters Alliance 2695 separation module and SpectraSystem UV1000 (ThermoSeparation Products) for separation and detection via UV at 280 nm. Monomer and oligomer peak areas were estimated using Empower software (Waters).

2.3. Results

Selection of variants by computational tools

Guided by the aforementioned computational tools, nine candidate variants populating each of the three defined mutational strategies were chosen for experimental

evaluation. Figure 2.3 depicts each variant grouped into its respective mutational strategy, and Figure 2.2 shows location of each variant site within the tertiary structure of γD-crys. Values for $\Delta\Delta G_f$, and $\Delta\Delta G_{bind}$ estimated by RosettaDesign and calculated using Eqs. 2.1 and 2.2, respectively, for each variant relative to wild type are summarized in Table 2.1.

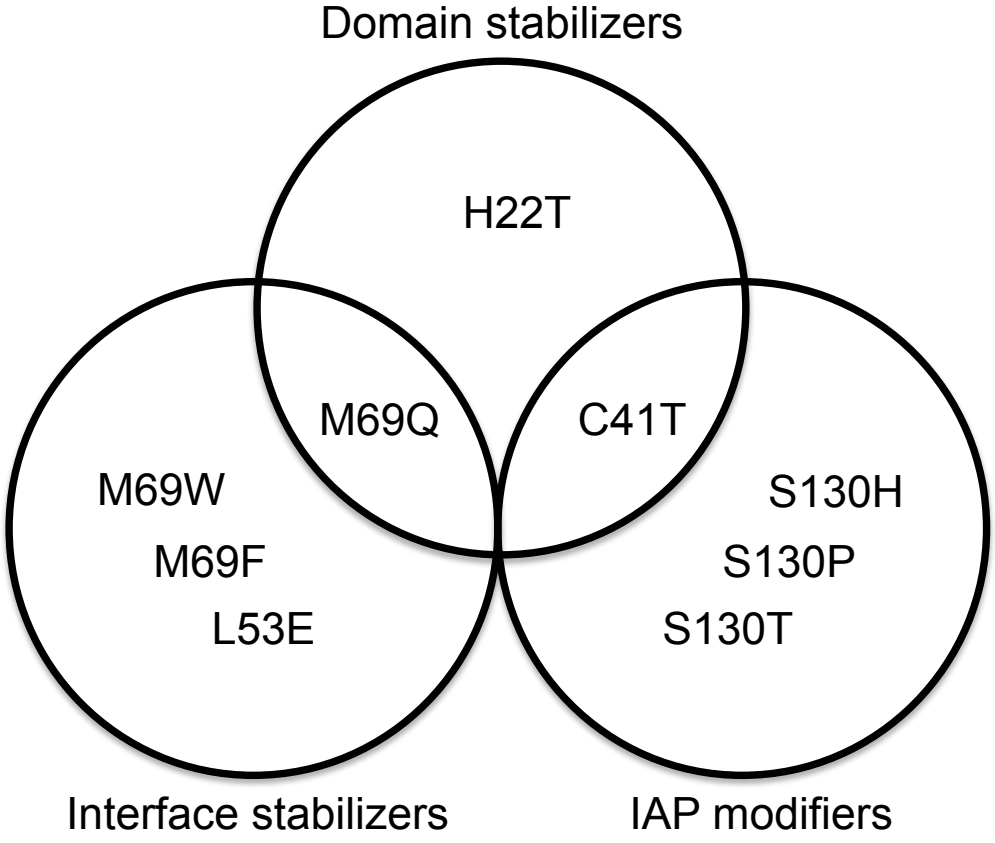


Figure 2.3**:** Candidate γD-crys variants identified by computational design. The variants are divided into their respective mutational strategy as domain stabilizers, interface stabilizers, or IAP modifiers. M69Q and C41T were variants predicted computationally to possibly conform to two of the mutational strategies.

Table 2.1: Summary of $\Delta\Delta G_f$ and $\Delta\Delta G_{bind}$ values calculated from Eqs. 2.1 and 2.2, respectively, based on predictions from RosettaDesign 3.0. $\Delta\Delta G_{assoc}$ values were calculated from Eq. 2.4 using the 3D profiling method. Here negative $\Delta\Delta G_f$ and $\Delta\Delta G_{bind}$ values indicate favorable changes in the folding or binding energies, respectively. Likewise, negative - $\Delta\Delta G_{assoc}$/RT values indicated a favorable decrease in the IAP of the molecule.

Variant	$\Delta\Delta G_f$ (kcal/mol)	$\Delta\Delta G_{bind}$ (kcal/mol)	$-\Delta\Delta G_{assoc}/RT$
H22T	-0.1	--	-0.8
C41T	-0.3	--	1.0
L53E	-0.5	-0.4	-2.0
M69F	-0.3	-0.3	0.8
M69Q	-1.0	-0.5	-1.0
M69W	0.5[a]	-0.3	-1.9
S130H	0.6	--	-1.9
S130P	1.4[b]	--	n/a[c]
S130T	-0.1	--	0.5

[a]RosettaDesign 2.1 predicted a stabilizing energy score of -2.9 kcal/mol for M69W [Sahin 2011].
[b]RosettaDesign 2.1 predicted an energy score of 0.5 kcal/mol for S130P [Sahin 2011].
[c]Energy scores for hexapeptides containing prolines were not calculated by the 3D profiling method, as they were considered β-strand breakers [Thompsen 2006].

The variants H22T, M69Q and C41T were all identified by RosettaDesign to conformationally stabilize the N-td relative to the wild type (denoted by negative $\Delta\Delta G_f$ values in Table 2.1) and thus were selected as N-td stabilizers. M69F, M69Q, M69W, and L53E were all located within 4 Å of the domain-domain interface and were identified to conformationally stabilize the interfacial region relative to wild type (denoted by the negative $\Delta\Delta G_{bind}$ values in Table 2.1). Therefore these four variants were selected as interface stabilizers.

Finally, to select IAP modifiers three of the aggregation calculators were used to identify aggregation-prone sub-sequences. Figure 2.4 shows two of the three aggregation calculators agreed upon two aggregation-prone regions located between residues G40-Y45 and S123-L136.

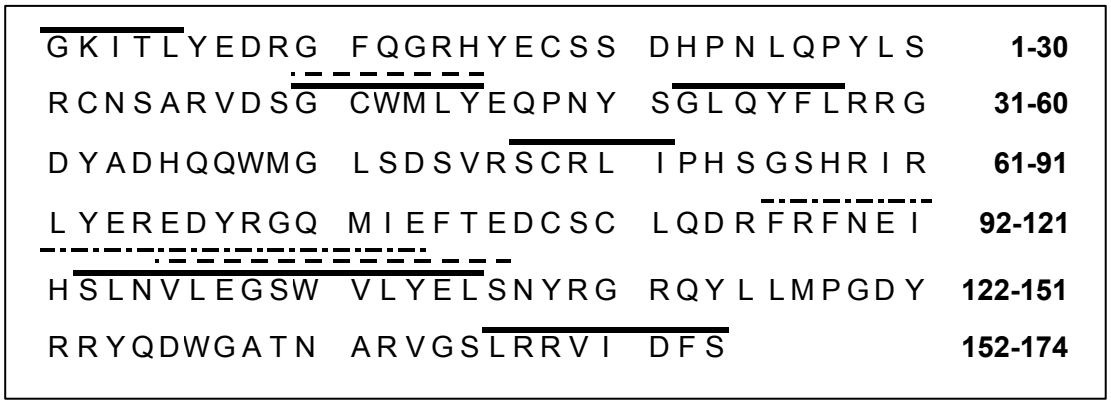


Figure 2.4: Predictions from the three aggregation calculators of aggregation-prone regions of sequence (i.e. "hot spots") within the wild type γD-crys primary sequence. The sequence was split at residue 86; thus the N-td contains residues G1-G85 and the C-td contains residues S87-S174. Lines above the sequences denote predicted "hot spots" for AGGRESCAN (solid line), PASTA (dash-dotted line), and TANGO (dashed line). As seen, two of the three calculators predicted "hot spots" between residues G40-Y45 and S123-L136.

These substitutions were then made in the primary sequence of γD-crys and analyzed by the aggregation calculators again. S130H and S130P were predicted to improve the IAP within the aggregation-prone region on the C-td while C41T was predicted to improve the IAP within the aggregation-prone region on the N-td; thus these three variants were chosen to study as IAP modifiers. Notably, none of these variants were qualitatively predicted by three of the aggregation calculators to increase the IAP of the protein (data not shown). Further, it should be noted RosettaDesign 3.0, used for this study, estimated S130P to be more conformationally destabilizing ($\Delta\Delta G_f$ = 1.4 kcal/mol) than RosettaDesign 2.1 ($\Delta\Delta G_f$ = 0.5 kcal/mol) used in our previous work [Sahin 2011] by nearly 1 kcal/mol. This difference is most likely attributed to changes in the energy function, such as re-optimization of weights for particular energy terms. Thus, based on RosettaDesign 3.0, S130P would not pass the filtering criteria; however, because our group investigated S130P in a previous study [Sahin 2011] it was included among the variants studied for this work as well. S130T was selected to study as a control as it was

predicted to actually increase the IAP within the C-td aggregation-prone region. More detail regarding the categorizing of variants into the three mutational strategies can be found in Section 2.2.

Quantitatively estimating IAP changes using the 3D profiling method

As mentioned previously, the aforementioned aggregation calculators are currently incapable of predicting quantitative changes in the IAP of a protein. Therefore, another aggregation calculator, referred to as the 3D profiling method [Thompson 2006], was utilized to try and quantitatively estimate changes in the IAP for each variant relative to wild type using Eq. 2.4.

$$\Delta\Delta G_{assoc} = \Delta G_{assoc}^{\mathrm{var}} - \Delta G_{assoc}^{wt} \qquad \text{(Eq. 2.4)}$$

Here, $\Delta\Delta G_{assoc}$ represents the free energy of each variant relative to wildtype of pairing two identical hexapeptides of a given sequence (containing the mutation site of each variant), using the known crystal structure of an amyloid peptide as the template. As γD-crys was shown previously to form amyloid fibrils [Papnikolopoulou 2008], and the 3D profile method uses the same energy score function as RosettaDesign, this was considered the most direct way to quantitatively compare folding vs. IAP changes on a similar scale. For this work, $\Delta\Delta G_{assoc}$ was computed for each variant in keeping with the original implementation of the 3D profiling method. Estimated values for $\Delta\Delta G_{assoc}$ are listed in Table 2.1, and more specific details regarding the 3D profiling method are provided in the following references [Thompson 2006, Nelson 2005].

Among the IAP modifiers, Table 2.1 shows the 3D profiling method quantitatively estimated a decreased IAP for S130H and an increased IAP for S130T, thus agreeing with the qualitative predictions from the other three aggregation calculators. On the other hand, the 3D profiling method estimated an increased IAP for C41T, different than the decreased IAP predicted qualitatively by the other three aggregation calculators. Further,

the 3D profiling method does not calculate energy scores for hexapeptides containing proline residues, as they are considered amyloid breakers, so a score for S130P could not be estimated. Table 2.1 also shows that among the variants RosettaDesign identified to be N-td and interface stabilizers, H22T, L53E, M69Q, and M69W were identified quantitatively by the 3D profiling method to also decrease the IAP of the molecule, while M69F was predicted to increase the IAP. These results were different than the other three aggregation calculators that qualitatively predicted no change in the IAP for these N-td and interface stabilizing variants.

Equilibrium chemical denaturation

The conformational stability of wild type and γD-crys variants was quantified experimentally by performing equilibrium chemical denaturation experiments using intrinsic tryptophan fluorescence spectroscopy to monitor protein unfolding as a function of increasing denaturant concentration. Denaturation curves were constructed by plotting the ratio of fluorescence intensities (FI) at 360 and 320 nm versus denaturant concentration at pH 3 (Figures 2.5-2.7) and pH 7 (Figure 2.8).

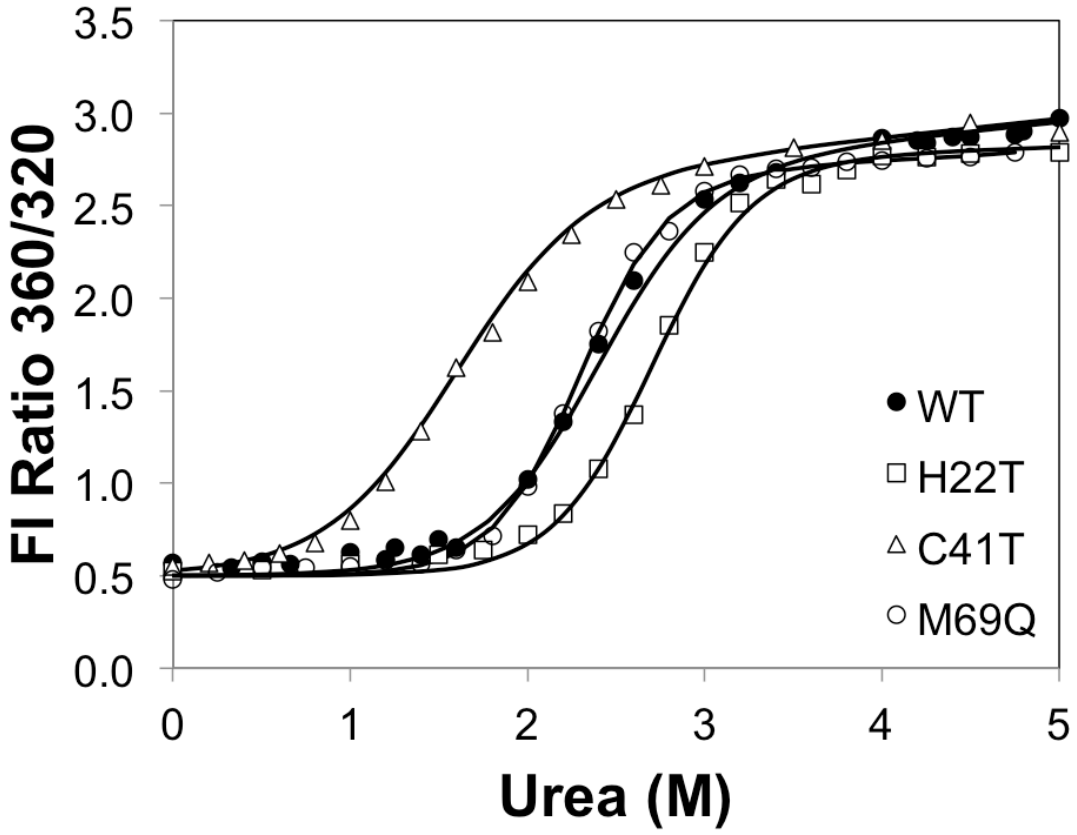


Figure 2.5: Chemical denaturation curves for wild type (WT) and each γD-crys variant identified by RosettaDesign to stabilize the less stable, N-terminal domain. Points represent experimental data, and solid lines represent nonlinear least squares fit of a two state unfolding model.

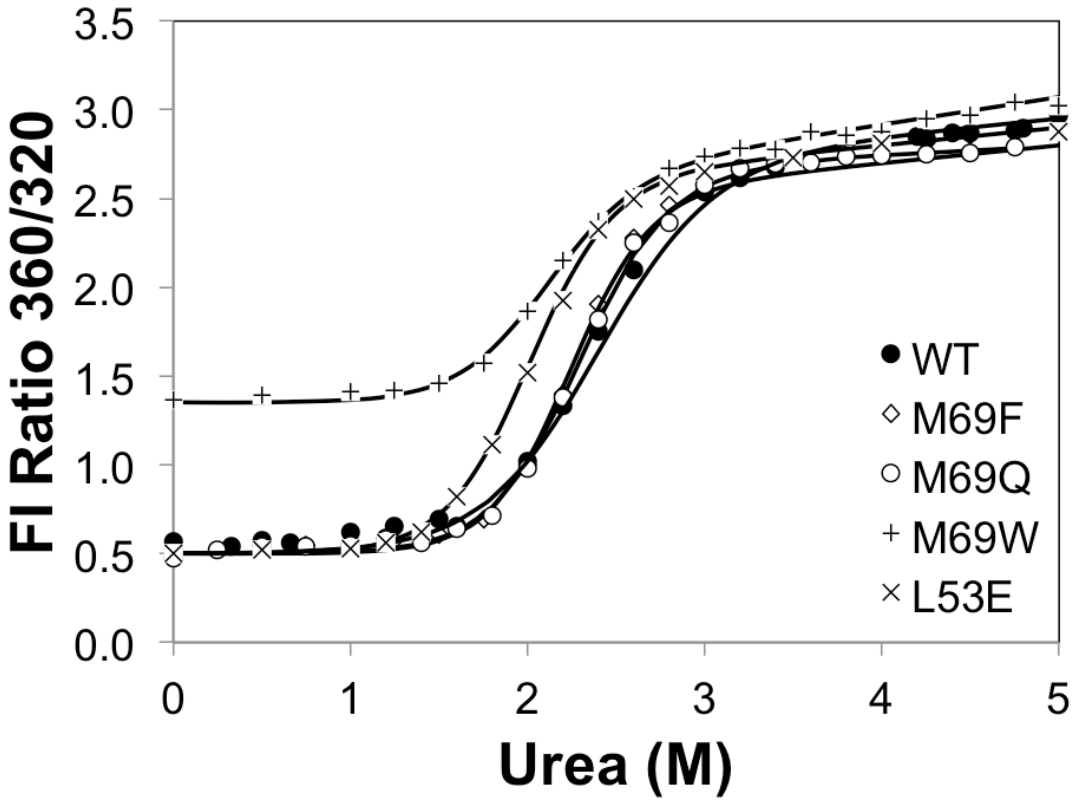


Figure 2.6: Chemical denaturation curves for wild type (WT) and each γD-crys variant identified by RosettaDesign to stabilize the domain-domain interface. Points represent experimental data, and solid lines represent nonlinear least squares fit of a two state unfolding model.

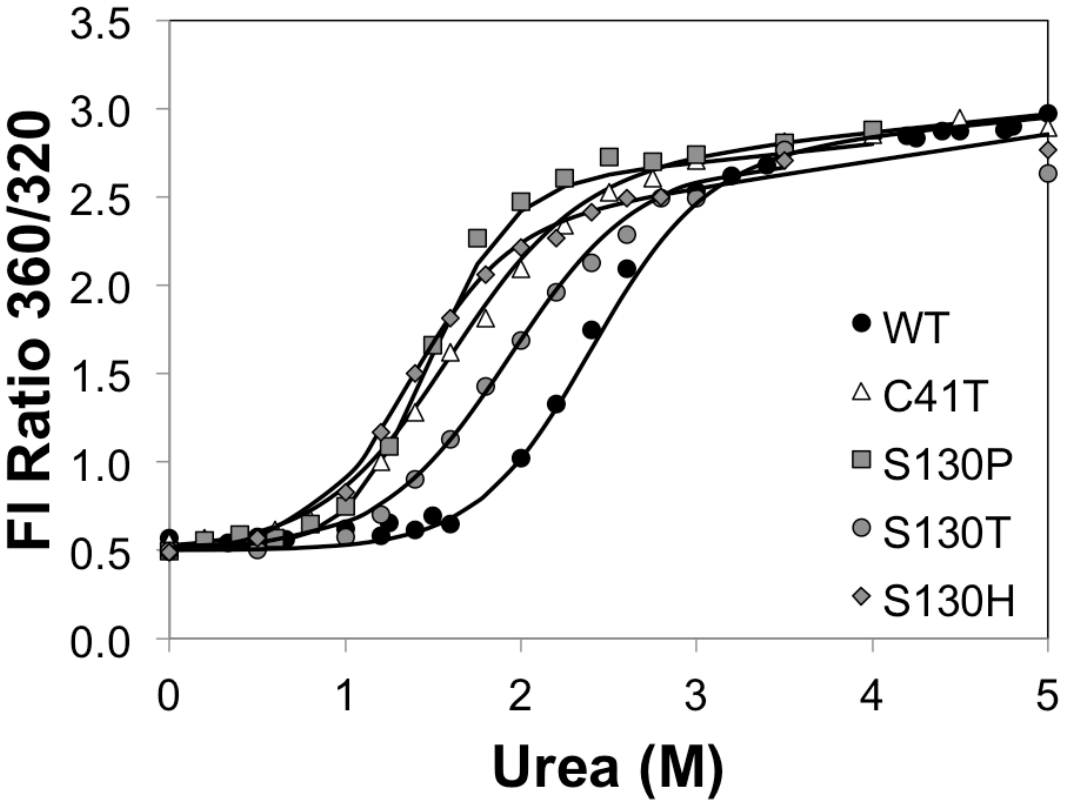


Figure 2.7: Chemical denaturation curves for wild type (WT) and each γD-crys variant identified by the aggregation calculators to modify the IAP, while not significantly destabilizing the protein. Points represent experimental data, and solid lines represent nonlinear least squares fit of a two state unfolding model.

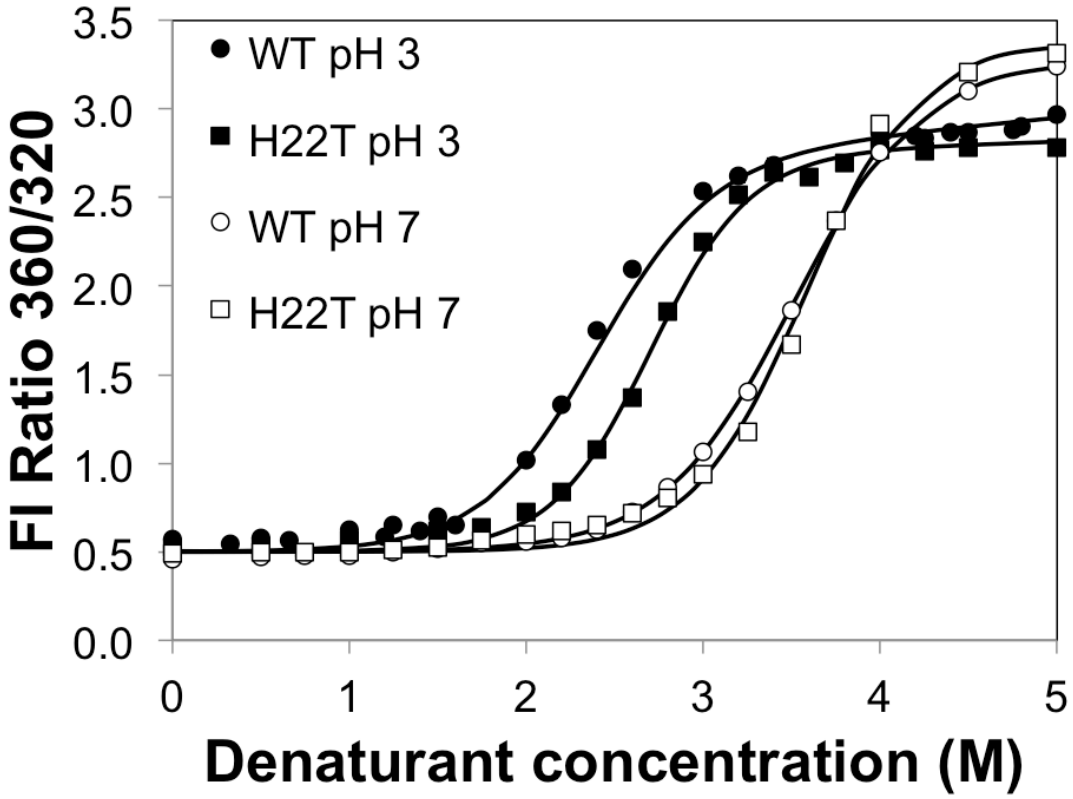


Figure 2.8: Chemical denaturation curves for H22T and wild type (WT) γD-crys at pH 3 and pH 7. Points represent experimental data, and solid lines represent nonlinear least squares fit of a two state unfolding model.

Unfolding curves shown as a function of increasing denaturant concentration indicated a single sigmoidal transition for wild type and all γD-crys variants (Figures 2.5-2.7). Thus the data were fit with a two-state unfolding model to approximate apparent thermodynamic parameters such as the Gibbs free energy of unfolding in the absence of denaturant, $\Delta\Delta G^{\circ}_{unf}$, the corresponding *m*-value, as well as the midpoint unfolding concentration of denaturant, C_{mid}. The two-state unfolding model, fit to the data by adjusting $\Delta\Delta G^{\circ}_{unf}$ and $C_{mid,}$ are also shown in Figures 2.5-2.7 and Figure 2.8 while Table 2.2 lists the apparent thermodynamic parameters including 95% confidence intervals using nonlinear regression.

Table 2.2: Summary of experimentally observed thermodynamic and aggregation parameters for wild type γD-crys and each variant. Thermodynamic parameters were estimated by fitting denaturant induced unfolding data with a two-state unfolding model. Unless specified, thermodynamic parameters were estimated at pH 3. Aggregation data obtained at short incubation times, up until the first ten percent of monomer loss, was used to estimate initial, observed aggregation rate coefficients, k_{agg}, and the corresponding time, t_{10}, for each variant.

Molecule	C_{mid} (M)	$\Delta G°_{unf}$ (kcal/mol)	m-value	k_{agg} (1/min)	t_{10} (min)
Wildtype	2.3 ± 0.1	4.3 ± 0.4	1.8 ± 0.2	0.02	5.8
Wildtype (pH 7)	3.5 ± 0.1	5.7 ± 0.2	1.6 ± 0.1	--	--
H22T	2.7 ± 0.1	5.7 ± 0.6	2.1 ± 0.3	0.003	27.4
H22T (pH 7)	3.6 ± 0.1	6.6 ± 0.9	1.9 ± 0.3	--	--
C41T	1.6 ± 0.1	2.5 ± 0.2	1.6 ± 0.1	0.05	2.2
L53E	2.0 ± 0.0[a]	5.0 ± 0.3	2.5 ± 0.1	0.03	3.0
M69F	2.2 ± 0.0[a]	5.8 ± 0.5	2.6 ± 0.2	0.03	3.4
M69Q	2.3 ± 0.0[a]	5.5 ± 0.4	2.4 ± 0.2	0.01	8.0
M69W	2.1 ± 0.0[a]	4.8 ± 0.5	2.3 ± 0.3	0.15	0.7
S130H	1.4 ± 0.1	2.8 ± 0.4	2.1 ± 0.3	0.45	0.2
S130P	1.4 ± 0.0[a]	3.8 ± 0.4	2.6 ± 0.3	0.01	7.8
S130T	1.9 ± 0.1	3.1 ± 0.4	1.6 ± 0.3	0.04	2.3

[a]The 95% confidence intervals reported for C_{mid} values are not zero, but merely rounded to one significant digit.

Stabilizing or destabilizing effects can be observed visually by shifts in the unfolding transitions to higher or lower denaturant concentrations, respectively. In addition, differences in C_{mid} were more statistically significant than $\Delta\Delta G^{\circ}_{unf}$ values because of the inherent error involved in the extrapolation to zero molar denaturant. Therefore, the C_{mid} values were considered a more reliable metric to infer conformational stability changes relative to wild type, nonetheless, the $\Delta\Delta G^{\circ}_{unf}$ values were used to correlate conformational stability to experimentally determined aggregation rate constants. The relative stability trends in terms of C_{mid} and $\Delta\Delta G^{\circ}_{unf}$ for wild type, M69Q and S130P are similar to those reported in previous studies [Sahin 2011].

We considered the reversibility of the folding in the experiments. Refolding experiments were conducted on wild type, H22T, M69Q, and S130P at select denaturant concentrations along the unfolding curve, and some hysteresis was observed for the

variant species, but not for wild type (data not shown). This suggests slow kinetics of refolding and/or aggregation, which also agrees with previous work where hysteresis observed during protein refolding experiments was less apparent at higher temperatures [Flaugh 2005a]. Thus, while no visible particulates were observed, as in our previous work [Sahin 2011], because we have not been able to show equilibrium can be achieved under all conditions, the estimated free energies of unfolding should be regarded as apparent for this work.

Among the variants identified to be potential domain stabilizers at acidic conditions (Figure 2.5), H22T exhibited significant conformational stabilization relative to wild type, M69Q maintained the conformational stability of wild type, while C41T was found to be destabilizing relative to wild type. Our previous work with M69Q indicated the mutation was conformationally stabilizing relative to wild type, however this behavior was not as evident during this study [Sahin 2011]. Nonetheless, in our earlier study differential scanning calorimetry (DSC) also showed M69Q had a much higher midpoint unfolding transition temperature than wild type at these solution conditions [Sahin 2011]. However, unfolding was convoluted with aggregation during DSC scans, and so it was not possible to quantify changes in $\Delta\Delta G^{\circ}_{unf}$ from such data. Along with M69Q, another variant identified to potentially stabilize the interface, M69F, was found to also maintain the conformational stability relative to wild type. On the other hand, L53E and M69W were visually observed to decrease the conformational stability relative to wild type (Figure 2.6), although the estimated $\Delta\Delta G^{\circ}_{unf}$ values for these variants inferred a stabilizing effect. Finally, each protein variant found within the group of IAP modifiers was observed to be destabilizing relative to wild type (Figure 2.7). Further, the noticeable increase in fluorescence intensity at low denaturant concentrations (folded state) for M69W in Figure 2.7 was most likely attributed to the addition of the tryptophan residue to

the protein sequence, although changes to the native conformation cannot necessarily be ruled out.

Denaturation experiments were also conducted at pH 7 for H22T and wild type γD-crys (Figure 2.8) to probe if the stability related to H22T was charge related. The difference between C_{mid} values for H22T compared to wild type (Table 2.2) evident at acidic conditions was not as pronounced at physiological pH. This indicates a charge change may be responsible for the added stability at acidic conditions that is not present at physiological conditions because of the intermediate pKa value of histidine.

Isothermal aggregation analyzed via size exclusion chromatography (SEC)

The aggregation behavior of wild type and γD-crys variants was compared qualitatively and semi-quantitatively using SEC in terms of the observed aggregation rate coefficient, k_{agg}, estimated from linear interpolation at early time points wherein the first ten percent of monomer is lost (i.e. corresponding to the initial rate regime). Initial monomer loss data was used so the dependence of intermediate aggregate species present for some variants but not for others at longer incubation times would not affect the estimated k_{agg} values. Although we would ultimately find that long time kinetic effects in irreversibility are important, we wanted to use a kinetic experiment to learn about initial time behavior. From a practical point of view, the time when aggregation initially becomes apparent is a critical parameter.

The experiments were conducted at pH 3 where aggregates of wild type and γD-crys variants remained soluble as determined by the lack of particulates visualized after isothermal incubation and a relatively constant area under the combined peaks of the SEC chromatograms. Table 2.2 lists the estimated k_{agg} values along with the corresponding time, t_{10}, defined as the time associated with the loss of the first ten percent of monomer for each variant and wild type. Figures 2.9-2.11 display the

monomer fraction remaining versus longer incubation times also tested for each variant as well as wild type. Values for k_{agg}, listed in Table 2.2 for all variants, were then compared.

For clarity, Table 2.3 summarizes the effect each variant was expected to have on conformational stability and IAP relative to wild type, based on the computational design predictions, as well as whether an improved effect on these properties was actually observed experimentally. Notably, C_{mid} values were used to infer changes in conformational stability instead of the $\Delta\Delta G^{\circ}_{unf}$ values whose inherent errors were larger due to the extrapolation back to zero molar denaturant.

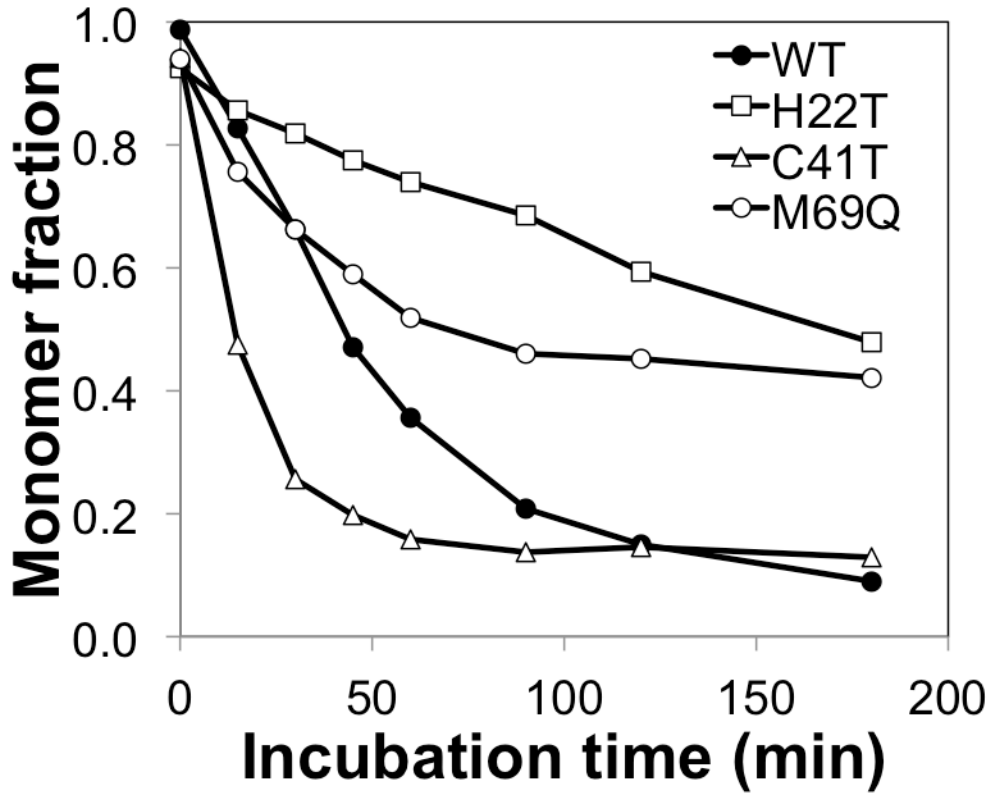


Figure 2.9: Monomer fraction remaining plotted as a function of incubation time at 50 °C for wild type (WT) γD-crys and each variant identified by RosettaDesign to stabilize the less stable, N-terminal domain. Points represent calculated fractions of monomer determined from SEC data, while lines connecting data points for each variant are included as a guide to the eye.

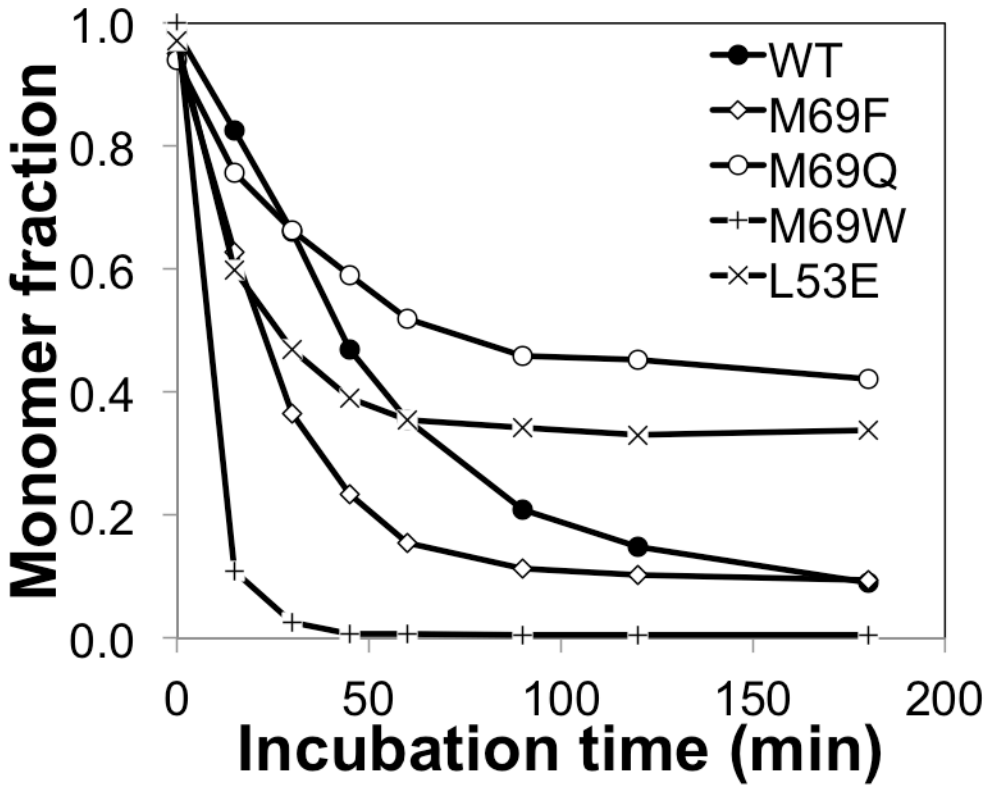


Figure 2.10: Monomer fraction remaining plotted as a function of incubation time at 50 °C for wild type (WT) γD-crys and each variant identified by RosettaDesign to stabilize the interface between domains. Points represent calculated fractions of monomer determined from SEC data, while lines connecting data points for each variant are included as a guide to the eye.

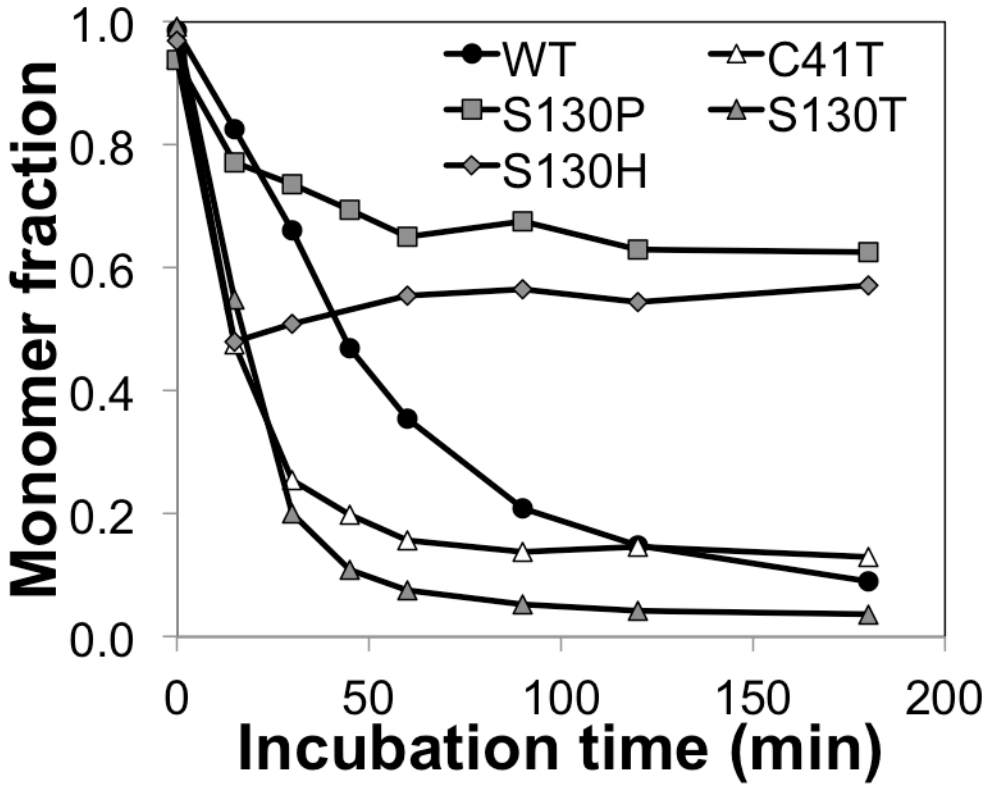


Figure 2.11: Monomer fraction remaining plotted as a function of incubation time at 50 °C for wild type (WT) γD-crys and each variant identified the aggregation calculators to modify the IAP, without significantly destabilizing the molecule. Points represent calculated fractions of monomer determined from SEC data, while lines connecting the data points for each variant are included as a guide to the eye.

Table 2.3: Summary of expected changes *a priori* for conformational stability and IAP estimated by RosettaDesign 3.0 and the aggregation calculators, respectively, as well as the experimentally observed changes in conformational stability (inferred by C_{mid} values) and k_{agg} values for each variant relative to wild type. Here, an expected increase in conformational stability and an expected decrease in IAP was considered potential improvements for each variant relative to wild type.

Variant	Expected change in conformational stability relative to wild type	Expected change in IAP relative to wild type[a]	Experimental change in C_{mid} relative to wild type[b]	Experimental change in k_{agg} relative to wildtype
H22T	increase	minimal change	increase	decrease
C41T	increase	decrease	decrease	increase
L53E	increase	minimal change	decrease	increase
M69F	increase	minimal change	not statistically significant	increase
M69Q	increase	minimal change	not statistically significant	decrease
M69W	decrease	minimal change	decrease	increase
S130H	decrease	decrease	decrease	increase
S130P	decrease	decrease	decrease	decrease
S130T	decrease	increase	decrease	increase

[a]For each variant, expected IAP predictions relative to wild type required consensus predictions between a majority of the aggregation calculators, otherwise a minimal change in IAP was anticipated. Here, a decreased IAP is favorable.
[b]The conformational stability for each variant relative to wild type was determined using the more statistically significant C_{mid} values over the $\Delta\Delta G^{\circ}_{unf}$ values whose inherent errors were larger due to the extrapolation back to zero molar denaturant.

Among the domain stabilizers H22T was observed to deter aggregation nearly five-fold relative to wild type, M69Q was shown to be slightly more resistant to aggregation than wild type, while C41T displayed enhanced aggregation versus wild type. Because M69Q was found by RosettaDesign to stabilize both the N-td and the domain-domain interface, it was also considered to be in the group of interface stabilizing variants. M69Q was the only variant among the interface stabilizers to have a decreased k_{agg} value relative to wild type, while M69F, M69W, and L53E all were observed to have increased aggregation rate coefficients relative to wild type. Finally among the IAP modifiers, S130P was the only variant identified to decrease the aggregation rate coefficient relative to wild type. This was an interesting observation since, unlike the other

aggregation resistant variants H22T and M69Q, S130P was observed to decrease the conformational stability of γD-crys. On the other hand, S130T, S130H, and C41T all augmented aggregation rate coefficients with S130T deliberately designed to do so (Table 2.3).

Notably, some variants, such as L53E and S130H, showed faster initial aggregation rates, but less total monomer loss at longer incubation times relative to wild type. Realistically, the improved colloidal stability observed for these variants at longer incubation times might be desirable in a biotechnological manufacturing setting where formulation and shelf life of vaccines is critical. Nonetheless, the presence of intermediate aggregate species at longer incubation times complicates the estimation of aggregation rates, and additional biophysical techniques that monitor the growth of aggregates along with monomer loss would be needed. Therefore only initial data was used to estimate aggregation rate coefficients in this work.

When comparing the experimental results to the quantitative predictions from the 3D profiling method, H22T, C41T, M69F, M69Q and S130T had net changes in the aggregation rate coefficients that were predicted correctly, while L53E, M69W, and S130H were not. The other three aggregation calculators correctly predicted the aggregation behavior qualitatively for S130P and S130T while C41T and S130H were not.

Observed k_{agg} values vs. conformational stability and predicted IAP

Next, a quantitative analysis was attempted to determine if changes in conformational stability and/or predicted IAP resulting from a point variant directly affected the observed aggregation rate coefficient. To do so, the natural log of k_{agg} for each protein variant was plotted against the experimentally observed change in $\Delta\Delta G^{\circ}_{unf}$ (Figure 2.12) as well as the $\Delta\Delta G_{assoc}$ as predicted by the 3D profiling method (data not

shown) all relative to wild type. Notably, AGGRESCAN, PASTA, and TANGO were incapable of predicting quantitative changes in the IAP, therefore the 3D profiling method was utilized as another aggregation calculator to quantitatively estimate IAP changes. Data points representing each variant shown in Figure 2.12 were divided into two groups; those identified by the aggregation calculators to modify the IAP of γD-crys (C41T, S130H, S130P, and S130T) and those identified by Rosetta to stabilize the N-td or the interfacial region (H22T, L53E, M69F, M69Q, and M69W).

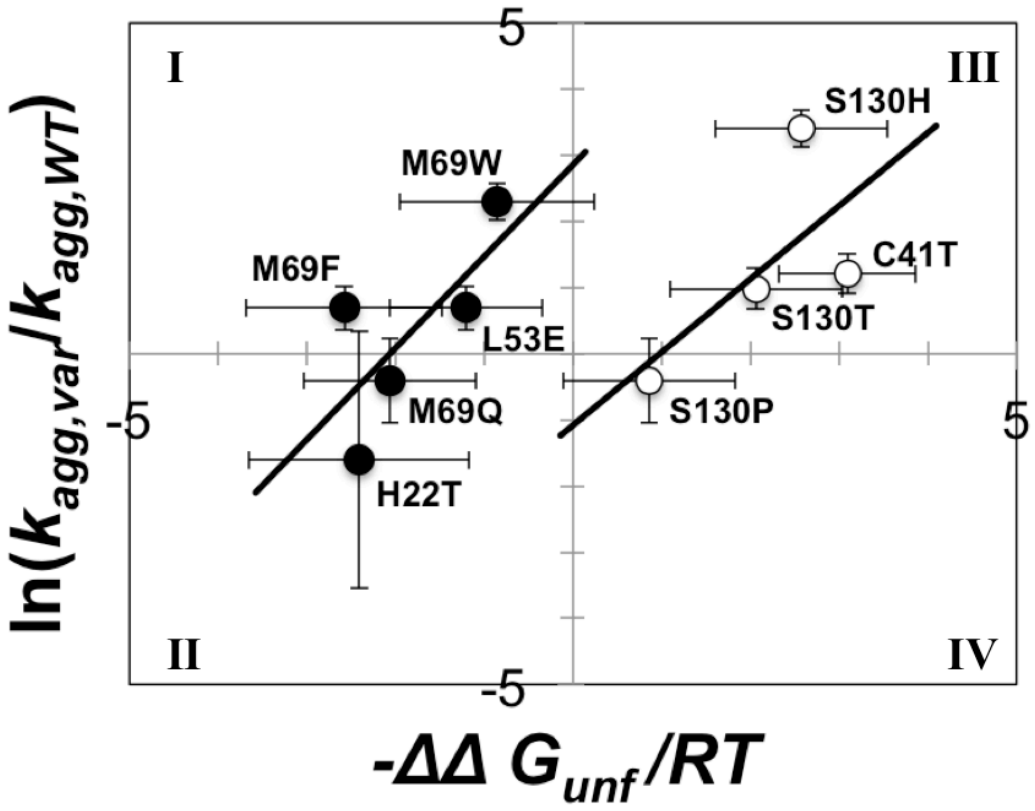


Figure 2.12: The natural log of the ratio of observed aggregation rate coefficients for each variant ($k_{agg,var}$) relative to wild type ($k_{agg,WT}$) plotted against experimentally determined free energies of unfolding ($-\Delta\Delta G_{unf}/RT$) for each variant relative to wild type γD-crys. Data points representing each variant were grouped into two groups, one containing the IAP modifiers (open circles) and the other containing N-td and interface stabilizers (closed circles). Error bars represent the standard error based on the standard deviation. Linear regression of both groups of data produced slopes of 1.1 and 1.4 with corresponding *p*-values of 0.257 and 0.047 for the IAP modifiers and the N-td plus interface stabilizers, respectively. The figure is divided into four quadrants each containing mutations with varying conformational stability and aggregation behavior. Quadrants I and II contain mutations that are conformationally stabilizing but either increase or decrease the k_{agg} value for each variant relative to wild type, respectively. In contrast, quadrants III and IV contain variants that are conformationally *destabilizing* and increase or decrease the k_{agg} value for each variant relative to wild type, respectively.

Linear regression of both data sets produced slopes of 1.1 and 1.4 with corresponding *p*-values of 0.257 and 0.047 for the IAP modifiers and N-td/interface stabilizers, respectively, assuming a α-value of 0.05 (95% confidence interval). This analysis indicates there is a statistically significant correlation between the natural log of the aggregate rate coefficient and the unfolding free energy as expected for aggregation mechanisms that involve monomer unfolding as a key step. On the other hand, the IAP modifiers did not show a statistically significant dependence of conformational stability on aggregation.

As previously noted, S130P decreased aggregate formation relative to wild type while exhibiting decreased conformational stability. This suggests that another molecular property contributes to aggregation. This possibility is further supported by the presence of at least one variant within each of the four defined quadrants shown in Figure 2.12. The four quadrants represent regions of varying conformational stability and aggregation behavior. For instance, quadrants I and II contain mutations that are conformationally stabilizing but increase or decrease the k_{agg} value of the variant, respectively, relative to wild type. In contrast, quadrants III and IV contain mutations that are conformationally *destabilizing* but increase or decrease the k_{agg} value of the variant, respectively, relative to wild type. If changes in conformational stability had the dominant effect on the k_{agg} values, then all variants should lie in either quadrant II or III.

Predicted IAP changes to the protein, based on the 3D profiling method, were also quantitatively compared to the k_{agg} values across the different variants, but no statistically significant, quantitative correlation was found (*p* values of 0.17 and 0.94 for the IAP modifiers and N-td/interface stabilizers, respectively, with α = 0.05). The lack of a statistically significant *p*-value for both groups could be due to the inaccuracies involved with quantifying IAP values using the 3D profiling method, as is the case for the other aggregation calculators implemented in this work. Nonetheless, analyzing the

energy scores from the 3D profiling method in a qualitative manner did identify some differences in IAP changes compared to the other aggregation calculators. For instance, C41T and M69F were predicted to increase the IAP using this method compared to the other aggregation calculators, and thus were qualitatively consistent with aggregation data obtained experimentally. Nonetheless, predictions that disagreed with the experimentally obtained aggregation data were also observed for certain variants (e.g. L53E, M69W, and S130H) using this method as well. Likewise, other investigators have also reported successful qualitative-based correlations between the predictions of these calculators for foldable proteins and actual experimental results; however, incorrect correlations were also found within these studies [Sahin 2011, Zhang 2010, Ivanova 2006, Routledge 2009]. Therefore, more accurate methods or algorithms for accurately measuring and quantifying IAP would be valuable.

2.4. Discussion

For some time, it has been recognized that changes in formulation or process conditions [Remmele 1999, Webb 2001] can reduce protein aggregation. On the other hand, modifying the protein itself is in orthogonal engineering approach to addressing the aggregation problem. Previous investigators have tried to correlate aggregation behavior with conformational stability for variants obtained from large-scale screening [Chrunyk 1993, Worn 1999, Worn 1998]. For instance, the relationship between conformational stability and the temperature at the onset of aggregation for point variants of interleukin-1β showed a strong linear correlation [Chrunyk 1993]. An interesting outlier, K97V, in the same study, possessed increased conformational stability but a lower aggregation temperature relative to wild-type [Chrunyk 1993]. Another study reported a similar trend between conformational stability and aggregation temperature with a polysaccharide-binding antibody single-chain Fv (scFv) fragment

[Worn 1999, Worn 1998]. Notably, another outlying variant that was conformationally stabilizing but exhibited a lower aggregation temperature than wild type was also found in this study [Worn 1999]. However, neither of these studies considered IAP changes as a possible contributor to the aggregation behavior of these outlying variants. K97 is located in an exposed region of in interleukin-1*β* where the replacement of a lysine with a valine might increase the IAP and observed aggregation rates. To strengthen this argument, IAP predictions for the K97V variant of interleukin-1*β* and the N52S variant in the scFv fragment were estimated using the aforementioned aggregation calculators. Results showed a majority of the other three aggregation calculators as well as the 3D profiling method estimated increased IAP for both variant sequences relative to wildtype. Another study reported a correlation between the disease duration of amyotrophic lateral sclerosis (ALS) versus the conformational stability and aggregation propensity of Cu,Zn human superoxide dismutase (hSOD) for a variety of ALS-linked hSOD variants. The aggregation propensity was estimated using a function based on changes in hydrophobicity, secondary structure, and charge resulting from amino acid substitutions [Chiti 2003]. When the conformational stability and aggregation propensity data were combined, a more statistically significant correlation to disease duration was observed compared to the correlation involving only one of the two properties [Wang 2008]. These findings further support the hypothesis that the observed aggregation rate of a protein is dependent upon both the conformational stability and IAP of the molecule.

In these previous studies, the variants were identified by large scale mutagenesis and screening. In this study, we wanted to test the *a priori* ability of RosettaDesign and the sequence-based aggregation calculators to identify variants with reduced aggregation behavior based on three mutational strategies exploiting improved conformational stability or IAP.

Evaluation of mutational strategies

From the observed results it was shown that each predefined mutational strategy produced at least one protein variant that was more aggregation resistant relative to wild type; H22T and M69Q among the domain and interface stabilizers, respectively, and S130P among the IAP modifiers (Table 2.2). However, only two of these variants (H22T and M69Q) were observed to also be conformationally stabilizing, while interestingly S130P was found to be destabilizing (Figures 2.5-2.7, Table 2.2). In particular for M69Q, it was difficult to determine from the experimental results gathered here whether the variant specifically stabilized the N-td, the domain-domain interface, or both. Performing chemical denaturation experiments on the individual N-td, as was done in previous work [Kosinski-Collins 2004] with M69Q and wild type γD-crys would be useful future work to clarify this issue. Nonetheless, this work shows all three mutational strategies are capable of improving the aggregation behavior of a multi-domain protein system, which are prominent in biopharmaceutical development and often associated with human disease.

Evaluation of the success rate of RosettaDesign

Another objective of this study was to evaluate RosettaDesign as a protein engineering design tool for reducing aggregation. Comparing either the C_{mid} or $\Delta\Delta G^{\circ}_{unf}$ values (Table 2.2) obtained experimentally for each protein variant at pH 3 with the computational scores indicated that RosettaDesign (Table 2.1) *qualitatively* predicted destabilizing and stabilizing trends collectively with a two-thirds success rate. Furthermore, RosettaDesign qualitatively predicted stabilizing variants correctly for one third of the variants tested (H22T, M69Q, and M69F). A previous study reported successes in four of nine variants tested when using RosettaDesign to try and improve the stability of redesigned globular proteins [Dantas 2003], and higher success rates (60-

75 percent) were observed when RosettaDesign was used to specifically redesign protein interfaces [Sammond 2007]. Furthermore, previously reported values for the average sequence recovery of protein redesigns compared to wild type sequences was approximately 30-40 percent [Dantas 2003, Dantas 2007].

However, most importantly, utilizing RosettaDesign and these aggregation calculators in tandem qualitatively predicted variants that reduced aggregation relative to wild type correctly with a one-thirds success rate as well. This is promising as no prior studies, to our knowledge, has used these computational design tools together with these mutational strategies to successfully deter protein aggregation using single, point variants. The algorithm also successfully predicted the C-td of wild type γD-crys to be more stabilizing than the N-td, consistent with findings in previous work [Mills 2007].

Notably, the success rates for these computational design tools are somewhat subjective as compelling benchmarks or metrics have not yet been determined regarding what are meaningful predictive values or acceptable predictive yields. Strong quantitative correlations were not observed, and were not expected since the version of RosettaDesign used here was programmed to estimate conformational stability changes only at physiological conditions. As such, differences can be expected at neutral pH compared to the conformational stabilities observed at acidic pH, needed for our experiments.

Furthermore, the success rates observed in this work are higher than those observed using high-throughput random screening. For instance a review by Eijsink et. *al*. reports several studies where the stability of enzymes has been improved by variants selected using random screening; however, the percentage of variants observed to have improved stability during the initial round of screening out of the number of total variants screened was generally shown to be less than 1 percent [Eijsink 2005, Richardson 2002, Palackal 2004].

Furthermore, a fixed-backbone design protocol considering larger rotamer libraries was implemented in this work, and only allowed for the energy minimization of rotamer alignments. Conceivably, the insertion of some point variants could perturb the protein backbone, and previous work has shown implementing a flexible-backbone design protocol can increase the accuracy of predictions [Dantas 2007]. However, decreased accuracy has also been observed using this approach [Sammond 2007], and a significant increase in the expense of computational resources and time is required for such calculations (parallel implementation for multi-day runs on a cluster of 100+ nodes or more). Nonetheless, using other approaches within RosettaDesign that are less computationally expensive than a flexible-backbone protocol, but more involved than a fixed-backbone design could yield more accurate predictions. Two examples include 1) utilizing the soft_rep_design force field that specifies altered scoring weights to dampen Lennard-Jones potentials within a fixed-backbone design [Dantas 2007], or 2) implementing the relaxed mode application that can adjust the protein backbone and torsion angles by small amounts to correct steric clashes and improve intramolecular interactions.

Nonetheless, in this work RosettaDesign exhibited it was capable of identifying a conformationally stabilizing variant (e.g. H22T), as well as a variant that maintained the conformational stability (e.g. M69Q) relative to wild type γD-crys that both reduced aggregation. Further, H22T provided an example where a molecule designed by nature to be extremely stable was further stabilized, thus reducing aggregation by improving the conformational stability of a less stable protein than γD-crys would be expectedly easier.

Molecular analysis of variants with reduced k_{agg} values

The three variants observed to improve γD-crys aggregation were examined more closely to identify possible mechanisms taking place that may result in limited

aggregation. M69Q and S130P were investigated in an earlier study [Sahin 2011] where we discussed in detail proposed mechanisms resulting in decreased aggregate formation for S130P relative to wild type. In that study, we postulated the limited aggregation observed for S130P resulted from decreased IAP; wherein the native fold of the molecule was altered but aggregation-prone “hot spots” were not available in a solvent-exposed manner to the extent of wild type γD-crys [Sahin 2011].

From this analysis it was proposed that H22T reduced aggregation by stabilizing the otherwise unstable N-td of γD-crys [Flaugh 2005a, Flaugh 2005b, Flaugh 2006]. ASA calculations estimated from ASAview [Ahmad 2004] show H22 is only slightly solvent exposed (14.3 percent ASA) with little change upon mutation from histidine to threonine (13.7 percent ASA). Further, predictions from RosettaDesign for H22T showed favorable changes in the solvation term, E_{sol}, and the repulsive contribution to the Lennard Jones energy term, E_{LJ}. Therefore, it is reasonable that substitution of a histidine residue with a more compact threonine residue would relieve steric repulsions; and the removal of a charged His under acidic conditions from a buried cavity containing no neutralizing charges would also be favorable.

For M69Q, it was concluded in our previous work that decreased aggregation rates were a result of improved conformational stability, although a reduction in IAP could not be disregarded [Sahin 2011]. However, we did not extensively propose molecular mechanisms that would support this argument. M69 is located on the N-td adjacent to the domain interface, and a molecular analysis of the variant site structure predicted by RosettaDesign indicated additional hydrogen bonds might form across the interface when the methionine is mutated to glutamine, as seen in Figure 2.13. The estimated length of these hydrogen bonds is 1.90 and 1.97 angstroms, respectively, within the generally accepted limit of distance for a hydrogen bond to form. ASA values estimated M69 to be moderately solvent exposed (42 percent), while the insertion of

glutamine decreased this value (35 percent). Consistent with these observations, RosettaDesign predicted a favorable change for M69Q for the sum of the hydrogen bonding terms, E_{HB}, relative to wild type. Thus, the additional hydrogen bonding suggested, particularly across the interface between domains, could feasibly result in increased conformational stability, but a reduction in IAP cannot be ruled out based on the limitations of available methods.

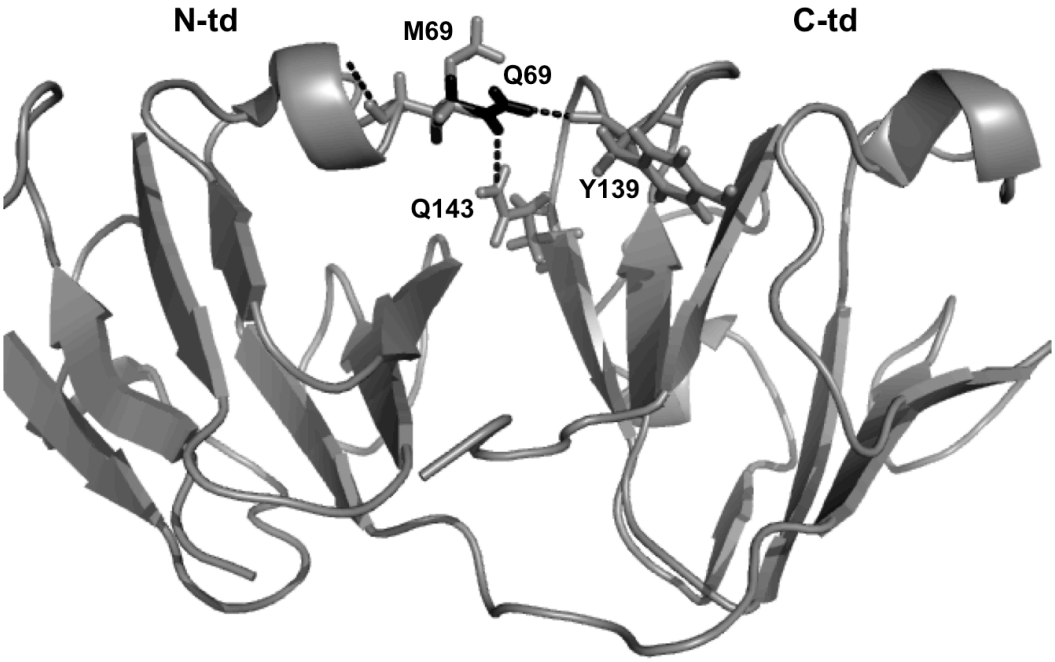


Figure 2.13: *PyMOL* image illustrating the added hydrogen bonds (black dashed lines) spanning the domain interface associated with the M69Q variant (black sticks) compared to wild type γD-crys (gray sticks). The hydrogen bonds cross the interface from Q69 to the main chain of Y139 and the side chain of Q143, also denoted by gray sticks.

Observed k_{agg} values are not solely dependent on conformational stability

Since non-native aggregation is affected by many of the same driving forces as folding, attempting to alter one of these parameters by mutation will likely also affect the other. The covariance of these molecular properties would suggest that optimizing both

simultaneously might be a more worthwhile approach than optimizing them serially or individually. Here, we attempted to demonstrate this concept by correlating the observed aggregation rate coefficients for numerous γD-crys variants against both the $\Delta\Delta G^{\circ}_{unf}$ and the IAP relative to wild type (Figure 2.12). Correlations between protein aggregation and conformational stability have previously been observed [Chrunyk 1993, Worn 1999, Worn 1998, Wang 2008], but changes in IAP were not necessarily considered to also play a role in these studies.

The results demonstrated the observed aggregation rate coefficients were generally more dependent on changes to the conformational stability of the molecule over changes in the IAP. Nonetheless, if changes in conformational stability were the lone contributor to the resulting aggregation rate coefficients, one might not expect the decreased k_{agg} value observed for S130P or even the increased k_{agg} value observed for M69F (Tables 2.2 and 2.3) relative to wild type because of their observed conformational stability. These two variants serve as strong examples indicating another factor besides conformational stability, such changes in the IAP, can also affect the observed aggregation rate coefficient. However, additional work is needed to elucidate more reliable, quantitative correlations.

2.5. Conclusions

In this work the computational design algorithm RosettaDesign and several sequence-based aggregation correlations (e.g. PASTA, AGGRESCAN, TANGO, 3D profiling method) were used to predict point variants that would alter the conformational stability or intrinsic aggregation propensity (IAP) of the model multi-domain protein, human γD-crystallin (γD-crys). Nine variants were identified to test three mutational strategies to reduce the aggregation: (1) stabilizing the less stable domain, (2) increasing the binding interaction between the domains, and (3) decreasing the IAP of

the polypeptide chain. Variants from each strategy were able to significantly reduce aggregation rate coefficients, however, in each case incorrect predictions were also notable. Nonetheless, this demonstrates each mutational strategy is capable of reducing aggregation in multi-domain proteins, which is applicable to many biopharmaceutical products and some human diseases. Further, experimental values of unfolding free energy (conformational stability) correlated well with aggregation rates coefficients, wherein higher unfolding free energy resulted in lower aggregation rate coefficients. However, notable outliers were observed, in particular for those molecules designed to alter the IAP while minimally affecting conformational stability. This highlights a need to consider a balance between altered conformational stability and IAP when using computational design and protein engineering to mitigate protein aggregation.

More importantly, this work demonstrated utilizing these computational tools in tandem could identify proteins variants with improved aggregation relative to wild type. This is particularly noteworthy as, to our knowledge, no prior studies have used these computational design tools together with these mutational strategies to successfully deter aggregation using single, point variants in multi-domain proteins. Nonetheless, the observed success rates were moderate and are subjective as benchmarks for meaningful predictive values or acceptable predictive yields have not yet been established. Therefore, improved design tools are still desired that can incorporate changes to both the IAP and conformational stability to yield higher predictive success rates on a quantitative as well as qualitative basis.

Chapter 3: Investigating the aggregation mechanism of wild type γD crystallin and several point variants using hydrogen-deuterium exchange coupled with mass spectrometry

3.1. Introduction

Non-native protein aggregation (hereafter referred to as just aggregation) is a significant problem observed in many biotechnological manufacturing processes and can often compromise the biological activity of the molecule [Vazquez-Rey 2011, Wang 2005, Mahler 2009, Weiss 2009, Cromwell 2006, Chi 2003]. In addition, aggregation can elicit an undesired immune response in patients after treatment [Wang 2012, De Groot 2007]. Thus, additional downstream purification operations are often required and inserted before formulation steps to remove these aggregates and increase the purity of the target molecule. However, these extra purification steps can result in significant yield losses of the therapeutic product as well as require costly equipment and resources. Furthermore, previous work has suggested protein aggregation is attributed to several human neurodegenerative diseases such as Alzheimer's, Parkinson's, and amyotrophic lateral sclerosis (ALS) [Rousseau 2006, Murphy, 2002, Fink 1998, Sanchez de Groot 2006, Chattopadhyay 2009].

Two strategies to reduce aggregate formation include 1) optimizing process and/or formulation conditions, and 2) modifying the molecule itself via protein engineering techniques. The first approach can only address and attempt to mitigate aggregation within a manufacturing or production process. As a result, aggregate impurity levels observed within a production process or found within a final formulation or vaccine can potentially be reduced using this approach. However, the first approach is incapable of reducing aggregate formation as related to human disease.

The second approach to deterring aggregation is to modify the molecule itself. Here, protein engineering can be used to improve the conformational stability of the molecule, or reduce the aggregation propensity by replacing amino acid(s) that may contribute to the aggregation mechanism to residues that disfavor these undesired interactions. However, before this can be accomplished, candidate variants that may improve the conformational stability or aggregation propensity of the molecule must first be identified, along with potential aggregation contact sites. In this dissertation, Chapter 2 discussed the selection of candidate variants using computational design tools to deter aggregation in a model multi-domain protein. Therefore, in this chapter we will focus on identifying aggregation contact sites within this same molecule using a combination of computational and experimental approaches.

Past work has highlighted the fact that many primary sequence-based aggregation calculators have been developed to identify aggregation-prone segments within a given protein sequence [Caflisch 2006, Tartaglia 2005, Fernandez-Escamilla 2004, Conchillo-Sole 2007, Trovato 2007, Thompson 2006]. However, these calculators were developed from experimental data from short polypeptides, and therefore were not developed to capture nonlocal interactions or global changes in the tertiary structure of larger protein molecules [Sahin 2011]. Furthermore, they lack predictive capability for aggregates lacking significant β-sheet structure, and do not fully account for environmental conditions that may instigate aggregation [Ebrahim-Habibi 2010]. Despite this, some correlation between predicted and experimentally identified aggregation-prone regions has been observed [Sahin 2011, Tartaglia 2008, Routledge 2009].

Additionally, other computational tools have also been developed that do account for the tertiary structure of the protein. For example, a simulation-based technology, referred to as Spatial Aggregation Propensity (SAP), can identify potential aggregation-prone regions of a protein based on the dynamic exposure and spatial proximity of

hydrophobic residues [Chennamsetty 2010]. This computational technique accounts for conformational fluctuations in the protein backbone or side chains in the predictions, and can identify clusters of hydrophobic amino acids that may be aggregation-prone because they are close in space, even if they are distant in the primary sequence [Voynov 2009].

On the other hand, experimental approaches have also shown promise in regards to identifying aggregation contact sites in proteins. For instance, the fluorescent dyes 4,4'-dianilino-1,1,'-binaphthyl-5,5'-disulfonate (Bis-ANS) and Nile Red are capable of binding to hydrophobic protein residues exposed to the solvent. The binding of these dyes to the protein increases fluorescence intensity, and thus they can be used as molecular probes to detect surface hydrophobicity of proteins using fluorescence spectroscopy [Pande 2010]. Pande et. *al.* used this experimental technique to study the less soluble, P23T variant of γD-crys. Results from the dye binding experiments showed the amount of bound Bis-ANS increased with the insertion of the P23T variant relative to wild type, but decreased upon aggregation of P23T [Pande 2010, Banjerjee 2011]. This observation suggested Bis-ANS competed with the aggregation of P23T and potentially identified the mutation site as an aggregation contact [Banerjee 2011]. Nonetheless, these dyes can generally only bind to solvent exposed, hydrophobic amino acids within the protein structure, although electrostatic interactions between Bis-ANS and charged residues can also occur [Pande 2010]. Therefore, the use of these dyes in identifying aggregation contact sites throughout the entire protein structure is limited.

Alternatively, the experimental approach known as hydrogen-deuterium exchange is capable of monitoring changes in solvent exposure throughout the entire structure of a protein. In this technique, amide hydrogens located on the backbone of the protein and corresponding to each amino acid within a protein sequence, are capable of exchanging with a deuterium at a defined rate dependent upon their role in hydrogen bonding, solvent accessibility, and environmental conditions (e.g. pH, temperature). The

solvent accessibility of the amide hydrogen is predominantly dependent upon its location within the tertiary structure of the protein [Hvidt 1966, Englander 1983], and protein unfolding in a highly concentrated deuterium environment can facilitate deuterium, *D*, exchange of an otherwise buried amide hydrogen, *H*. Thus, this process can be modeled by the following reaction scheme [Hvidt 1966]

$$N(H) \underset{k_f}{\overset{k_{unf}}{\rightleftarrows}} U(H) \xrightarrow{k_{int}} N(D) \underset{k_f}{\overset{k_{unf}}{\rightleftarrows}} U(D)$$

where *N* and *U* represent the amide in its native or unfolded state, and k_{unf}, k_f, and k_{intr} are defined as the unfolding, folding, and intrinsic exchange rate constants. The intrinsic exchange rate of a given amide hydrogen is defined as the rate observed if the amide hydrogen is completely exposed to solvent. This rate is dependent upon the environmental conditions (e.g. pH, temperature), as well as the neighboring amino acids to the amide hydrogen of interest [Bai 1993]. Further, this model assumes all of the buried amides are initially hydrogens and the exchange from hydrogen to deuterium is irreversible because deuterium is highly concentrated in the labeling buffer [Hvidt 1966].

Nonetheless, certain molecular interactions or conformational changes that can occur in protein aggregation can also reduce deuterium labeling. For instance, increased hydrogen bonding, such as in the cross-β structure associated with amyloid formation, can slow deuterium labeling of amide hydrogens. Additionally, deuterium labeling of solvent exposed residues in the native state can decrease if those residues become buried in an aggregated state. As such, utilizing hydrogen-deuterium exchange can potentially identify aggregation contacts for a given protein by comparing the deuterium labeling patterns of an aggregated state versus the native monomeric state. This technique has been used extensively to elucidate the structure of aggregates for many different proteins such as the monoclonal antibody Bevacizumab [Zhang 2011], α-

chymotrypsinogen [Zhang 2010], human interferon-γ [Tobler 2002]; as well as peptides including the β-amyloid peptide whose fibrillation is associated with Alzheimer's disease [Qi 2009, Kheterpal 2006].

Changes in mass of a protein or peptide resulting from deuterium labeling can be detected and analyzed using nuclear magnetic resonance (NMR) [Hoshino 2002] or mass spectrometry (MS) [Zhang 2011, Zhang 2010, Qi 2009, Tobler 2002]. However, NMR is difficult to implement for the analysis of aggregate conformations because it is limited to smaller molecular weight species (approximately less than 30 kDa) and requires larger amounts of protein compared to mass spectrometry [Schuster 2008].

Additionally, hydrogen-deuterium exchange coupled with mass spectrometry (HX-MS) can provide an abundant amount of information (e.g. structural, thermodynamic, kinetic), particularly when a protein digestion step is included prior to MS analysis. As examples, previous studies using HX-MS with protein digestion has helped elucidate the aggregate structure α-chymotrypsinogen [Zhang 2010], generate protein unfolding curves to estimate thermodynamic parameters [Dai 2006], and monitored aggregation kinetics of Aβ(1-40) [Qi 2008]. Nonetheless, using protein digestion to obtain peptide-level resolution also results in data processing and analysis that is quite complicated and involved, particularly when analyzing multiple labeling times or solution conditions. To combat this, many useful tools are becoming available to allow for more efficient and faster data processing, such as the software program *HDExaminer* developed by Sierra Analytics. Additionally, new graphical formats as well as a detailed, statistical analysis have recently been developed that aid in rapidly observing both qualitative and quantitative changes in deuterium labeling amongst different protein conformations or variants [Houde 2011]. These tools were implemented within this work to help identify potential aggregation contacts in γD-crys variants and wild type. More specific details describing their use can be found in Section 3.2.

Using experimental techniques such as HX-MS to identify potential aggregation contacts typically requires a significant amount of time and resources. Therefore, implementing alternative strategies that would incorporate faster data accumulation and analysis is desired. One such strategy is to use computational approaches, however, only moderate success rates have been observed thus far for various computational tools in regards to specifically deterring protein aggregation [Sahin 2011, Dantas 2003, Sammond 2007, Miklos 2012, Tartaglia 2008, Routledge 2009] or identifying intermolecular contacts in proteins [Zhang 2010]. As such, their reliability in identifying aggregation contact sites may not be as certain compared to experimental techniques, such as HX-MS. Therefore, until higher predictive success rates for computational approaches can be achieved, the combination of computational and experimental techniques may provide the most robust route to identifying aggregation contact sites.

Thus for this work, both experimental and computational approaches were implemented in an attempt to elucidate the structural, aggregation mechanism of notable γD-crys variants and wild type. The γD-crys variants chosen for this study were identified *a priori* by computational design tools to alter the conformational stability and/or aggregation propensity of the protein, and were subsequently observed experimentally to display diverse aggregation behavior relative to wild type. HX-MS, incorporated with an inline enzymatic step to digest deuterium-labeled protein and achieve peptide-level resolution, was utilized as the experimental approach to compare the labeling patterns of aggregated versus monomeric γD-crys conformations. Afterwards, qualitative and quantitative comparisons were made, including a detailed, statistical analysis to identify statistically significant differences in deuterium labeling. Lastly, the utility of the previously described aggregation calculators was assessed by comparing the predicted, aggregation-prone segments of sequence for each γD-crys species to the experimental measurements. HX-MS has been used to study the structural changes of other

crystallins, such unfolding of the destabilizing F9S variant in γS-crystallin [Mahler 2010, Lee 2010] and deamidation at the dimer interface of βB2-crystallin [Takata 2010], but to our knowledge has not been applied to wild type γD-crys and the variants studied in this work.

Therefore, the primary objectives of this work were 1) to experimentally identify, via HX-MS, peptides within the wild type γD-crys structure that may contain aggregation contacts, 2) to compare and contrast the aggregation-prone peptides suggested by HX-MS for wild type γD-crys to those suggested for three γD-crys variants, and 3) to determine if correlations exist between the computational predictions of several aggregation calculators and the experimentally obtained HX-MS results.

3.2. Materials and Methods

γD-crys monomer and aggregate preparation

Purified protein aliquots of selected γD-crys variants and wildtype, stored at -80 °C, were first thawed and then placed on ice. Next, protein solutions were prepared in 50 mM citrate buffer, adjusted to pH 3 in deionized and distilled H_2O (ddH_2O), to an initial concentration of 1 mg/ml. γD-crys aggregates were prepared for variants as well as wildtype by incubating 250 µL of protein solution at 1 mg/ml in a water bath at 50 °C for 180 minutes. Incubation was conducted in upright, 2 mL, glass HPLC vials with PTFE/silicone caps (Fisher Scientific) in a water bath with negligible temperature variability. Samples were removed after 180 minutes and immediately placed in an ice water bath to quench further aggregation. No additional aggregation or change in monomer fraction was observed, between sample quenching and SEC analysis (data not shown). The composition of aggregate and monomer included in each protein sample was estimated using SEC with a mobile phase comprised of 0.5% phosphoric acid, adjusted to pH 2.7 with 5 M NaOH and operated at 0.8 mL/min. Sample volumes of

75 µL were injected into a TSKgel Guard SWxl column attached in series with a TSKgel G2000SWxl analytical column (TOSOH 7.8 x 30.0 cm, 5 µm) connected to a Waters Alliance 2695 separation module and SpectraSystem UV1000 (ThermoSeparation Products) for separation and detection via UV at 280 nm. Monomer and oligomer peak areas were estimated using Empower software (Waters).

Hydrogen-deuterium exchange

Solutions containing monomer and aggregates of γD-crys variants and wildtype were analyzed using hydrogen-deuterium exchange. Approximately 15 µL of protein solution in 50 mM citrate buffer, pH 3.0, was diluted nine-fold into a deuterium-rich labeling buffer also containing 50 mM citrate, pD 3.0. The samples were incubated at room temperature for a defined labeling time, and four total labeling times of 0, 12, 120, and 1200 minutes were tested. Fully labeled protein samples were also prepared by incubating the same amount of protein for 24 hours in a deuterium-rich solution containing 8 M guanidine hydrochloride (GdnHCl), 100 mM tris(2-carboxyethyl)phosphine) (TCEP), and 20 mM ethylenediaminetetraacetic acid (EDTA). Three replicates were prepared and analyzed for all four labeling times tested. After labeling, each sample was adjusted to pH 2.6, where the hydrogen-deuterium exchange reaction rate is at a minimum, by first adding ice cold quench buffer (150 mM phosphate, pH 1.5) to halt deuterium labeling, followed by dissociation buffer (8 M GdnHCl, 100 mM TCEP, 20 mM EDTA, adjusted to pH 2.6) to a GdnHCl concentration of 2.5 M to aid proteolytic digestion. The sample was then immediately placed on ice for 2 minutes followed by the addition of sample pump solution (95% ddH_2O, 5% acetonitrile (ACN), 0.1% formic acid, 0.01% trifluoroacetic acid (TFA)) to dilute the GdnHCl concentration to approximately 1.5 M to promote efficient protein digestion.

HPLC-MS analysis of deuterium-labeled protein samples

Each sample was then injected into a 500 µL stainless steel sample loop using a 500 µL glass syringe. A sample pump (LabAlliance Series I) then sent the protein solution through the sample loop at a flow rate of 150 µL/min to an immobilized pepsin column (2.1 mm ID, 30 mm length) to initiate proteolytic digestion. Pepsin was used because it cleaves at numerous residues and is active at acidic conditions. Peptides eluting from the pepsin column were then trapped, desalted, and concentrated on a Peptide Microtrap C_8-desalting column (1 mm ID, 8 mm length, Michrom Bioresources) for a desalting time of 9 minutes. After the desalting step, flow was switched from the sample pump to an HPLC-MS Surveyor system pump (Thermo Scientific) to initiate a gradient of increasing ACN to elute the protein from the C_8-desalting column. Downstream from the C_8-desalting column, an XBridge C_{18} column (2.1 mm ID, 50 mm length, 3.5 *µ*m pore size, Waters) was used to improve peptide resolution. To minimize back exchange, the sample loop as well as all columns and lines were placed in a refrigerated cooler maintained at 0-1°C. Back exchange refers to the deuterium exchanging back to protons when the labeled protein sample is diluted back into protonated buffers used for sample quenching and HPLC-MS solvents. Thus, minimizing back exchange is important to retain discernable differences between unlabeled and labeled protein samples.

An optimized gradient method was also used to minimize back exchange, yet still effectively resolve peptides. The method consisted of a 17 minute gradient from 70% solvent A (ddH_2O, 0.1% formic acid, 0.01% TFA) and 30% solvent B (ACN, 0.8% formic acid) to 60% solvent B, followed by a 2 min gradient from 60% solvent B to 95% solvent B, followed by 2 additional minutes at 95% solvent B. Peptides were eluted from the C_{18} column directly into a LTQ electrospray ionization linear ion trap mass spectrometer (Thermo Finnigan). Data were collected in a positive ion, full scan type, and normal scan

rate profile mode with an ESI voltage of 4.3 kV, a capillary temperature of 275 °C and a sheath gas flow rate of 25 units.

Estimating extent of deuterium labeling for reporter peptides

All of the reporter peptides used for this analysis were assigned by performing MS/MS mass spectrometry, followed by analysis with Turbo SEQUEST software. The extent to which a peptide was labeled with deuterium was estimated using the data processing software program, *HDExaminer,* developed by Sierra Analytics, and is based on Eq. 3.1. The corresponding percentage of deuterium labeling that was observed for each reporter peptide is estimated from Eq. 3.2.

$$\frac{D}{N} = \frac{m_t - m_0}{m_{100} - m_0} \qquad \text{(Eq. 3.1)}$$

$$\frac{D}{N}\% = \frac{m_t - m_0}{m_{100} - m_0} \times 100\% \qquad \text{(Eq. 3.2)}$$

Here, D represents the number of deuterated amide hydrogens and N is the total number of exchange competent residues within a given peptide. Further, m_t is defined as the mass of a peptide after labeling time t, m_0 is the non-deuterated mass of that peptide, and m_{100} is the fully deuterated mass of that peptide. Exchange-competent residues refer to all amino acids within each reporter peptide except proline, which does not possess an amide hydrogen because of its unique side chain structure, and the N-terminal residue of each reporter peptide, because it does not contain a backbone amide. In addition, the residue located directly after the amino acid on the N-terminal end of each reporter peptide is also not considered as the back-exchange for this residue is high.

Construction of butterfly plots to qualitatively assess HX-MS data

Houde et *al*. developed a graphical format to aid in rapidly observing qualitative changes in deuterium labeling when comparing reporter peptides between two protein molecules (e.g. wild type vs. variant), two conformational states (e.g. monomer vs. aggregate), or two environmental conditions [Houde 2011]. An example of this graphical format is illustrated in Figure 3.1, and compares the labeling patterns of H22T monomer to wild type γD-crys monomer.

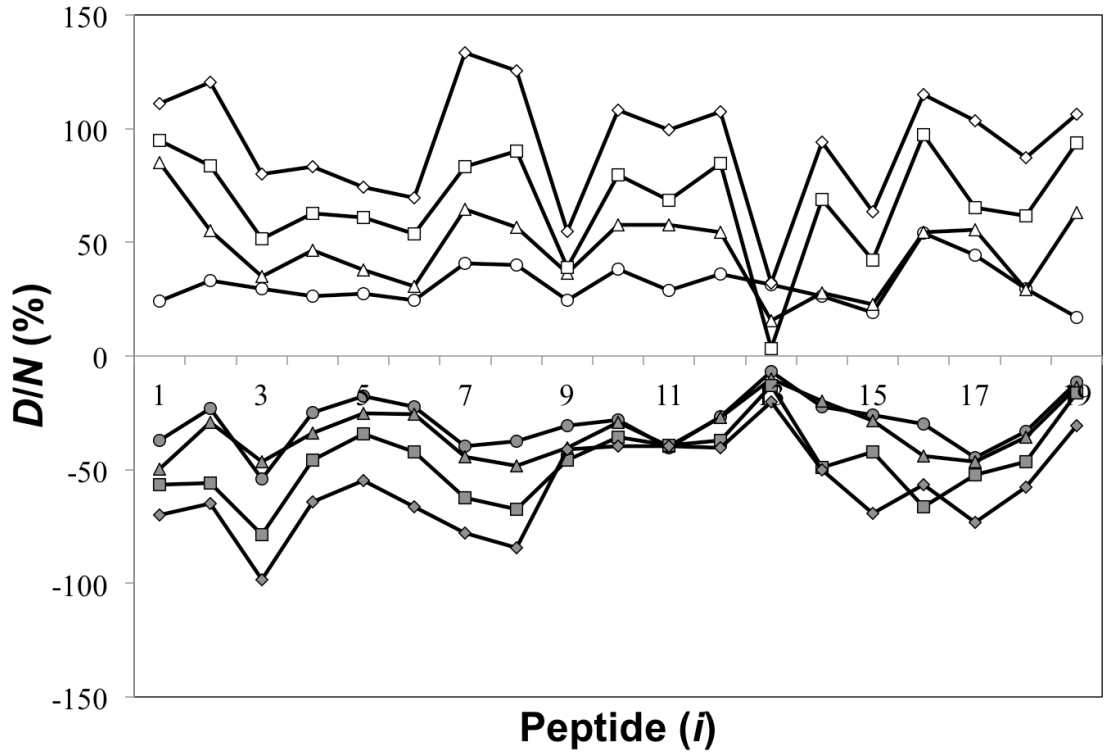


Figure 3.1: Butterfly plot showing the extent of labeling for each reporter peptide, *i*, to visually compare the monomeric structure of H22T (filled symbols) versus wild type (open symbols) γD-crys. Labeling times of 0 min (circles), 12 min (triangles), 120 min (squares), and 1200 min (diamonds) are shown for both proteins.

Here, the extent of deuterium labeling, shown as a percentage estimated using Eq. 3.2, for both the reference and experimental sample is plotted on the *y*-axis. These data represent a mean *D*/*N*% value observed for each reporter peptide estimated by taking the average of the three replicate samples at each labeling time. Data for the

reference samples (e.g. wild type or monomeric protein) are plotted as positive numbers, and data for the experimental samples (e.g. protein variant or aggregated protein) are multiplied by -1 and plotted as negative numbers. Using this mirrored format, exchange patterns for two environmental conditions, two protein states, or two protein variants can be visually compared on one graph quickly. The *x*-axis of the butterfly plots represents the sequential order of each reporter peptides analyzed. Here, each reporter peptide was numbered and ordered based on its sequence midpoint. For example, the reporter peptide with the lowest sequence midpoint was assigned a value of 1, and the peptide with the next lowest sequence midpoint was assigned a value of 2, etc. For situations were multiple peptides had the same sequence midpoint, the peptide containing the earliest sequence position at the N-terminal end of the fragment was assigned the lower value. The organization of the reporter peptides in this manner allows for the relative location of the peptide within the protein sequence to be determined quickly and simply.

Using a statistical analysis to examine HX-MS data

A detailed, statistical analysis has also been developed by Houde et. *al*. that allows for quantitative comparisons to be made between two protein species (e.g. wildtype vs. variant), two protein conformational states (e.g. monomer vs. aggregate), or two environmental conditions [Houde 2011]. To perform this analysis, two sample arrays containing raw data were first created for a reference $S_{ref}(M_{i,t})$ and experimental sample $S_{exp}(M_{i,t})$. The raw data contained in each array was the absolute deuterium uptake, $M_{i,t}$ measured in Daltons for a given peptide, *i*, and labeling time, *t*. Next, two difference arrays were assembled. The first difference array was assembled by calculating the difference in $M_{i,t}$ between the experimental and reference sample for a specific peptide and labeling time, as shown in Eq. 3.3.

$$D(\Delta M_{i,t}) = S_{\exp}(M_{i,t}) - S_{ref}(M_{i,t}) \qquad \text{(Eq. 3.3)}$$

A second difference array, denoted D_s, was also assembled by calculating the sum of $D(\Delta M_{i,t})$ from all labeling times tested for each reporter peptide, as indicated in Eq. 3.4.

$$D_s(i) = \sum_{t=1}^{4} D(\Delta M_{i,t}) \qquad \text{(Eq. 3.4)}$$

Both $D(\Delta M_{i,t})$ and $D_s(i)$ for a given peptide are measured in Daltons, where positive values indicate more deuterium labeling occurs for a peptide within the experimental sample relative to the reference sample, and vice versa for negative values. For this work, a positive value indicates a peptide has a more open or flexible structure within the experimental sample compared to the reference sample, and thus is prone to experience more deuterium labeling. On the other hand, a negative value indicates a more flexible and open structure was observed for a peptide in the reference sample compared to the experimental sample.

The statistical uncertainty for labeling differences of reporter peptides was assessed using a method very similar to that described by Houde et. *al*. [Houde 2011]. Briefly, three replicate samples were analyzed for each γD-crys species or conformational state tested. The corresponding uncertainty for each $D(\Delta M_{i,t})$ data point was estimated with the standard deviation (SD). An average of these SD values, denoted SD_x, was then calculated using Eq. 3.5 where i_{tot} and t_{tot}, represent the total number of reporter peptides and labeling times tested, respectively. For this work, 19 reporter peptides and 4 labeling times were examined resulting in 76 data points being averaged. This average was then used to calculate a standard error of the mean (SEM_x) for three replicates, using Eq. 3.6, that was representative for all average $D(\Delta M_{i,t})$ values. Next, a 98% confidence interval was estimated for any mean value of $D(\Delta M_{i,t})$, denoted 98% CI_x, using Eq. 3.7. Here, the appropriate value from the Student's t table

for a 98% confidence interval using two degrees of freedom was used (corresponding to 6.965).

$$SD_x = \frac{\sum_{i=1}^{i_{tot}} \sum_{t=1}^{t_{tot}} SD(i,t)}{i_{tot} t_{tot}} \qquad \text{(Eq. 3.5)}$$

$$SEM_x = \frac{SD_x}{\sqrt{3}} \qquad \text{(Eq. 3.6)}$$

$$98\%\, CI_x = SEM_x * 6.965 \qquad \text{(Eq. 3.7)}$$

A new average SD value, denoted SD_y, was also estimated and used for any $D_s(i)$ value, as shown in Eq. 3.8. Here, the value for SD_x and the standard propagation of error equation for a simple sum of variables was used to estimate SD_y. This value was used to estimate SEM_y, defined as the average for any $D_s(i)$ value from the three replicates tested, shown in Eq. 3.9. Another 98% confidence interval was estimated for any mean $D_s(i)$ value, denoted 98% CI_y, using Eq. 3.10. As before, the same value from the Student's t table for a 98% confidence interval using two degrees of freedom was used (corresponding to 6.965). The dashed lines shown in the following bar plots illustrate the confidence intervals calculated by Eq. 3.10.

$$SD_y = \sqrt{4(SD_x)^2} \qquad \text{(Eq. 3.8)}$$

$$SEM_y = \frac{SD_y}{\sqrt{3}} \qquad \text{(Eq. 3.9)}$$

$$98\%\, CI_y = SEM_y * 6.965 \qquad \text{(Eq. 3.10)}$$

Thus, differences in deuterium labeling between the reference and experimental sample exceeding the 98% CI_y, were considered statistically different. For example, Figure 3.2 shows the bar plot comparing the monomeric structure of H22T to wild type γD-crys. Here, the 98% CI_y was calculated to be ± 2.2 Da, and peptides 1, 2, 3, 7, 10,

12, and 19 exceeded this interval. Thus these peptides were considered to exhibit a statistically significant difference in labeling in H22T monomer relative to wild type monomer.

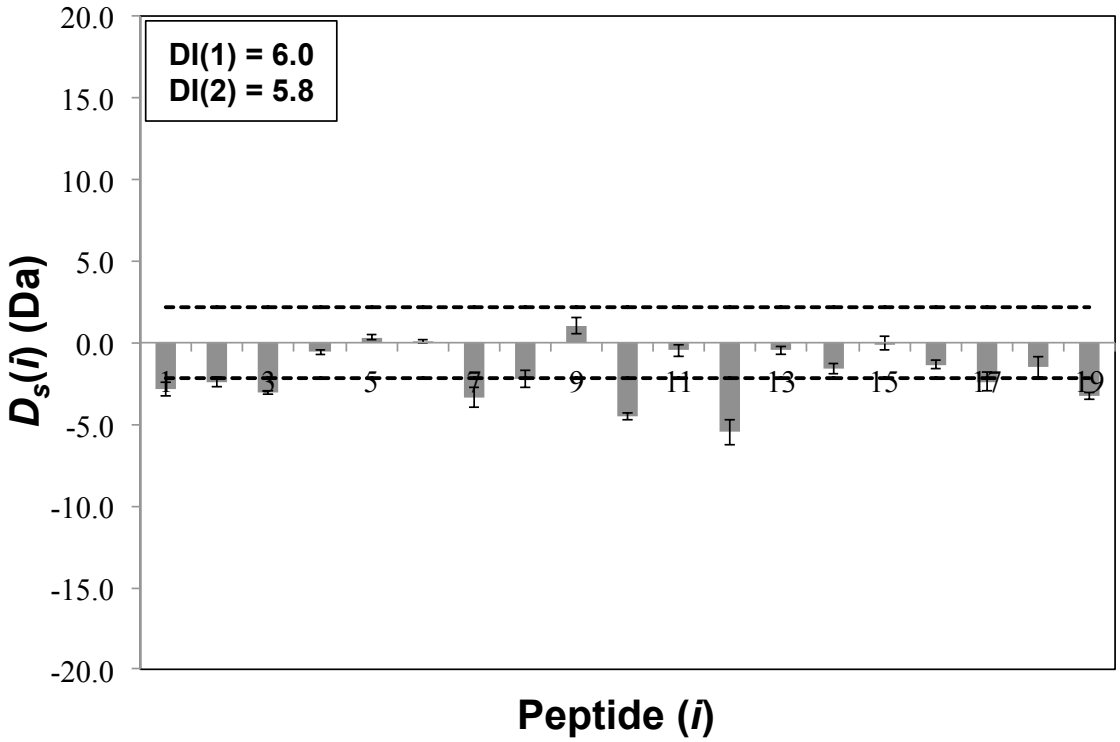


Figure 3.2: Values of $D_s(i)$ plotted for each reporter peptide, i, to compare the monomeric structures of H22T and wildtype γD-crys. Negative values indicate less deuterium labeling for H22T relative to wild type, and vice versa for positive values. Error bars represent the standard error in $D_s(i)$ (estimated by one standard deviation). Dashed lines represent a 98% confidence interval of ± 2.2 Da, calculated using Eq. 3.10. Those peptides having $D_s(i)$ values that exceeding this limit were considered statistically different. DI(1) and DI(2) values are calculated using Eq. 3.12 and 3.13, respectively.

Far-UVCircular Dichroism (CD)

Far-UV CD measurements were performed in 50 mM citrate buffer, adjusted to pH 3.0, for both monomeric and aggregated species of γD-crys variants and wild type. Optimal signal intensity was produced using a protein concentration of 0.3 mg/ml (14.6 μM). All measurements were collected at room temperature using a Jasco J-710 spectrapolarimeter, under nitrogen purge. Spectra were recorded from 200-250 nm with a scan rate of 50 nm/min in a circular, quartz cuvette (Helma) with a 1 mm path length. Three spectra were recorded and an average ellipticity value was estimated for each sample with the baseline subtracted to account for the buffer solution. The contribution of the aggregated state(s) of each γD-crys species to the overall CD spectra was estimated by treating the average ellipticity at a given wavelength as a linear combination of monomer and aggregate contributions. The fractions of monomeric and aggregated state(s) for each γD-crys species were determined by SEC analysis.

3.3. Results

Determining monomer and aggregate compositions via SEC

Prior to HX-MS analysis, γD-crys aggregates were prepared for each variant and wild type via isothermal incubation of the protein sample at 50 °C for 180 minutes. Following this, the composition of aggregate and monomer remaining in the protein sample was determined via HPLC-SEC. The corresponding chromatograms for S130T and wild type γD-crys are shown in Figure 3.3, while Figure 3.4 contains the chromatograms for H22T and S130P. Before isothermal incubation, all γD-crys samples contained greater than 92% monomer. After isothermal incubation, integration of the chromatographic peaks estimated that S130T and wild type γD-crys samples contained less than 5% monomer. On the other hand, the more aggregation-resistant variants, H22T and S130P, were observed to have approximately 40% monomer species

remaining; as well as more intermediate aggregated state(s) relative to wild type and S130T. For S130P, in particular, the formation of larger aggregate species was not observed to the extent of the other variants. Further a shift to a later elution volume for the thermally stressed, monomeric species of S130P was also observed (Figure 3.4). Similar results for S130P were also observed in our previous work [Sahin 2011]. Further, variations in the total peak were less than 10 percent indicating a relatively conserved mass balance for all protein samples. These samples were then analyzed by HX-MS. Notably, HX-MS data obtained for incubated S130P and H22T samples were corrected using Eq. 3.11 to approximate the extent of deuterium labeling attributed to the aggregated state(s) for each peptide, *i*, and labeling time, *t*,. This was conducted because of the significant amount of monomer remaining after incubation for these samples, and was done assuming a linear combination of the contributions from monomer and aggregate, as shown in Eq. 3.11.

$$D_{agg} = \frac{D_{tot} - x_{mon} D_{mon}}{x_{agg}} \qquad \text{(Eq. 3.11)}$$

Here, D_{agg} represents the deuterium uptake attributed to the aggregates in the protein sample, D_{tot} represents the total deuterium uptake observed experimentally for incubated samples containing both aggregates and monoemrs, and D_{mon} represents the deuterium uptake observed experimentally fornon-incubated samples containing mostly monomer. Further, x_{mon} and x_{agg} represent the composition of monomer and aggregate included in each sample as determined by SEC.

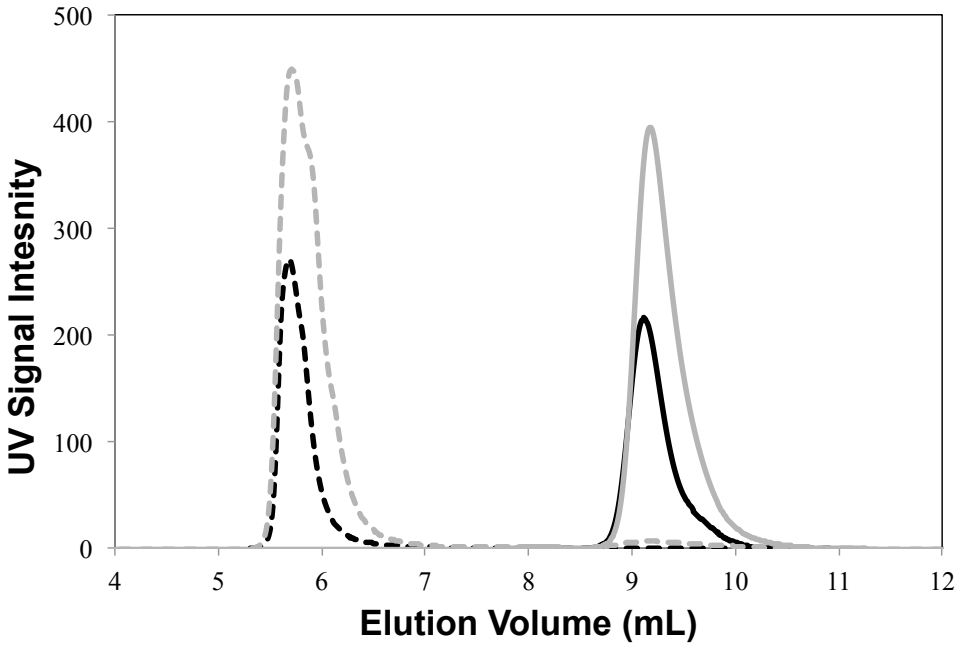


Figure 3.3: HPLC-SEC chromatograms of S130T (gray) and wild type (black) γD-crys for protein samples that were incubated at 50 °C for 0 minutes (solid lines) and 180 minutes (dashed lines).

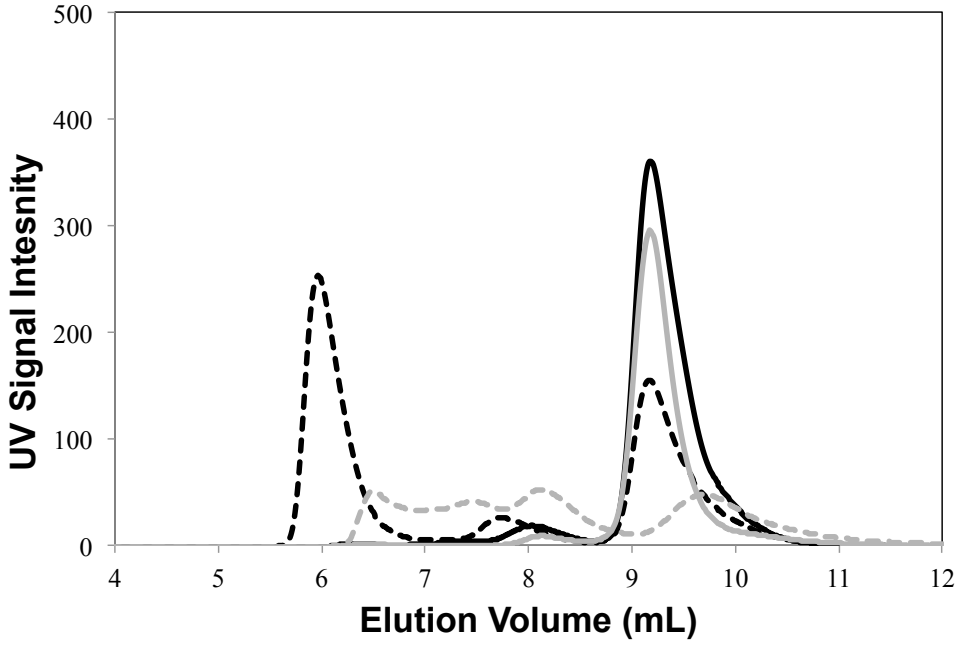


Figure 3.4: HPLC-SEC chromatograms of S130P (gray) and H22T (black) γD-crys for protein samples that were incubated at 50 °C for 0 minutes (solid lines) and 180 minutes (dashed lines).

Far-UV circular dichroism (CD) of monomeric and aggregated states

The secondary structure of each γD-crys species was analyzed using far-UV CD. Figure 3.5A shows the resulting CD spectra for each monomeric γD-crys variant and wild type species, while Figure 3.5B shows the CD spectra of each γD-crys variant and wild type having undergone isothermal incubation, and thus contained a mixture of aggregates and monomer. Comparing the CD spectra for in Figure 3.5A showed few differences, and a negative, minimum ellipticity value at 218 nm for each monomeric species. This CD spectra is characteristic of β-sheet structure, and consistent with the predominant secondary structure of native γD-crys as discussed in Chapter 2.

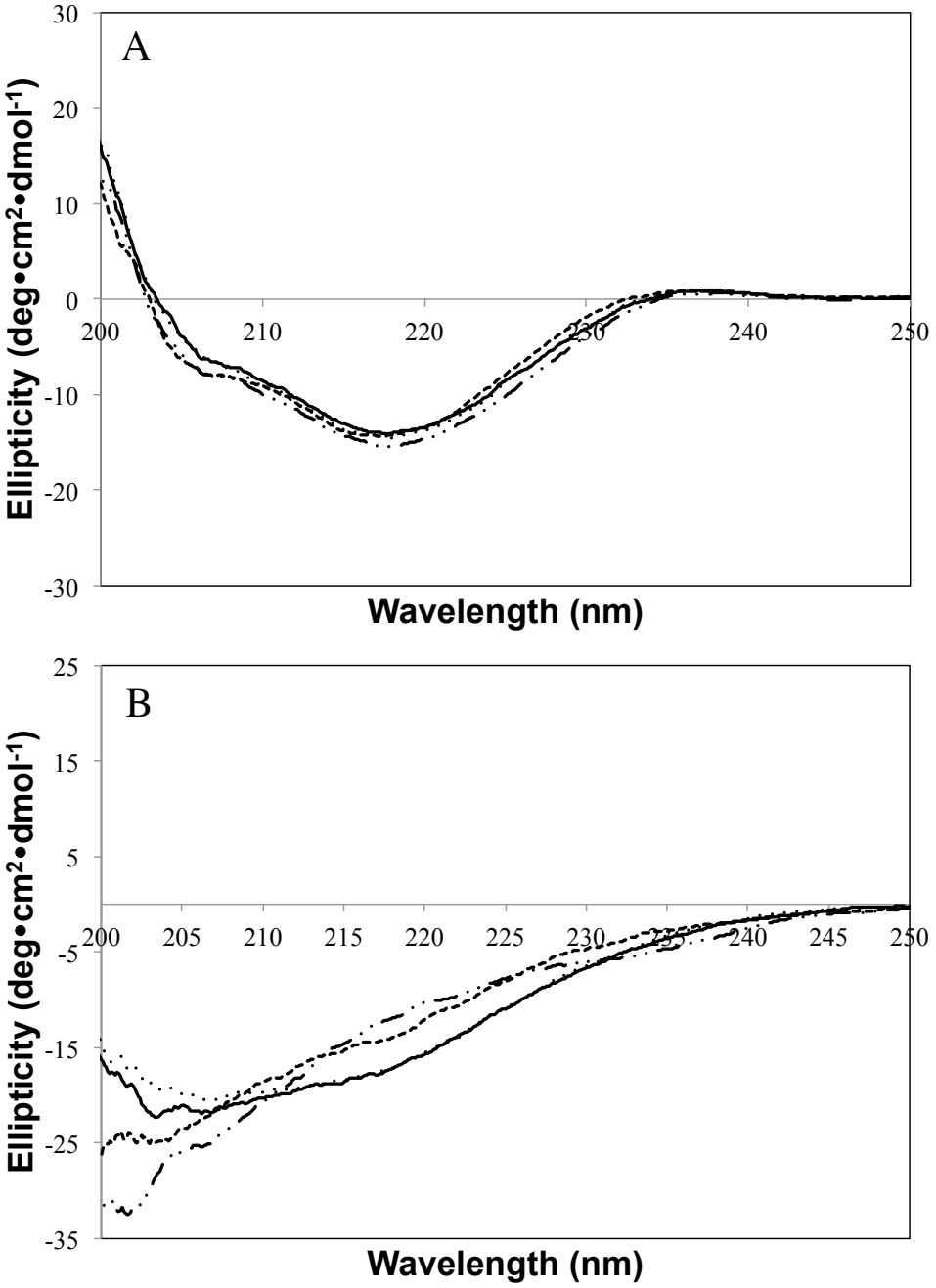


Figure 3.5: Far-UV circular dichroism data obtained at room temperature in 50 mM citrate, pH 3 for wild type (solid line), H22T (dash-dot-dot line), S130P (dashed line), and S130T (dotted line). Here, A) shows the CD spectra for the monomeric states of each γD-crys species and B) shows the CD spectra for the aggregated state(s) of each γD-crys species, with the monomer contribution subtracted and the resulting spectrum normalized to account for the fraction of aggregate present. All spectra were also corrected by subtracting the contribution of buffer solution.

On the other hand, Figure 3.5B shows the γD-crys species possessed an altered secondary structure in the aggregated state(s) after isothermal incubation relative to monomer. These alterations in structure produced a shift in the CD spectra where the minimum ellipticity occurred at lower wavelengths for each γD-crys species. Further, focusing on 218 nm in Figure 3.5B, the aggregated state(s) of wild type and S130T displayed more negative ellipticity values than H22T and S130P. This may suggest increased, antiparallel β-sheet formation for both S130T and wild type γD-crys compared to H22T and S130P; however, this is still somewhat inconclusive as the changes are not dramatic.

Analyzing deuterium labeling of peptides as a function of solvent exposure

Overall, the extent of deuterium labeling observed for reporter peptides via HX-MS was consistent with the solvent exposure of each peptide in the native protein structure. For instance, upon visually examining the labeling patterns of each γD-crys monomeric species (Figures 3.1, 3.7, and 3.9), the reporter peptides located near both terminals of each protein sequence (e.g. peptides 1-3, and 17-19), as well as peptides 7, 8, and 10-12 were observed to label faster than other peptides analyzed. On the other hand, peptides 4-6, 9, and 13-16 were observed to label slower compared to other reporter peptides analyzed. Notably, however, some discrepancies were also observed such as with peptides 10-12 and 19 in H22T (Figure 3.1), highlighting these remarks as general observations amongst the species tested.

After examining the tertiary structure of γD-crys, many of the residues located within peptides exhibiting faster labeling were generally more surface exposed or located on the periphery of the molecule, as shown in Figure 3.6. Further, these peptides were estimated to have higher accessible surface areas (ASA) (~30-50%), approximated by ASAview [Ahmad 2004], when compared to other reporter peptides analyzed. On the

other hand, the residues located within peptides exhibiting slower labeling (e.g. 4-6, 9, 13-16) were generally more buried, such as in the core of the domain-domain interface, as also seen in Figure 3.6. Additionally, lower ASA values (~10-30%), as approximated by ASAview, were estimated for these peptides.

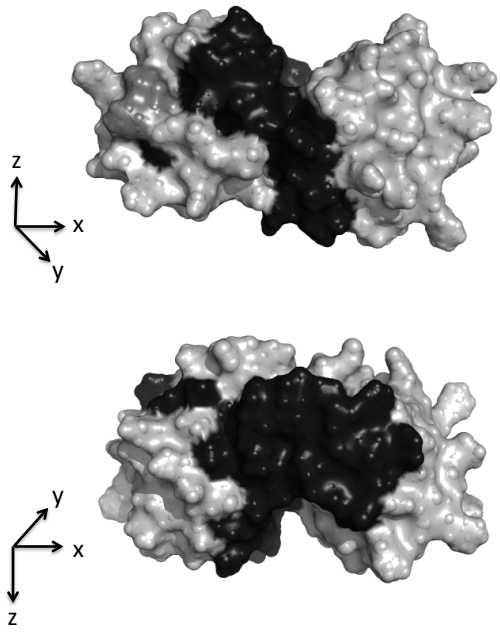


Figure 3.6: Surface exposure of residues within the native, tertiary structure of wild type γD-crys. Reporter peptides 1-3, 7, 8, 10-12, and 17-19 exhibited faster labeling, were generally more surface-exposed, and located on the periphery of the molecule (gray surfaces). Alternatively, residues 4-6, 9, and 13-16 exhibited slower labeling, were generally more buried, and located within the domain cores or near the domain-domain interface (black surfaces).

Analyzing monomeric structures of each γD-crys species using HX-MS

The monomeric structures of the variants H22T, S130P, and S130T were qualitatively assessed relative to wild type γD-crys by visually comparing the labeling trends in several butterfly plots shown in Figures 3.1, 3.7, and 3.9), respectively. Further, the labeling trends of these variants were also quantitatively evaluated using a detailed, statistical analysis shown in Figures 3.2, 3.8, and 3.10, respectively. First, nineteen reporter peptides were identified from peptide mapping, corresponding to 98% coverage

of the γD-crys sequence, and are listed in Table 3.1. These peptides were then used to construct the corresponding butterfly and bar plots.

Table 3.1: Reporter peptides identified from HX-MS analysis for each γD-crys variant and wildtype species.

Peptide (*i*)	Sequence	Position	Exchangeable amides
1	GSSHHHHHHSSGLVPRGSHMGKITL	HisTag-5	22
2	YEDRGFQGRHYECSSDH(T)PNLQPYL[a]	6-29	20
3	SDH(T)PNLQPYL[a]	20-29	6
4	NSARVDSGCWM	33-43	9
5	LYEQPNYSGLQ	44-54	8
6	LYEQPNYSGLQY	44-55	9
7	FLRRGDYADHQQWMGL	56-71	14
8	FLRRGDYADHQQWMGLSD	56-73	16
9	SDSVRSCRLIPHSGSHRIRL	72-92	17
10	YEREDYRGQMIE	93-104	10
11	TEDCSCL	106-112	5
12	LQDRFRFNEIHSL	112-124	11
13	NEIHSLNVL	119-127	7
14	NVLEGS(P/T)WVL[b]	125-133	7(6)[c]
15	YELSNYRGRQ	134-143	8
16	YLLMPGD	144-150	4
17	YRRYQDWGATNA	151-162	10
18	YRRYQDWGATNARVGSL	151-167	15
19	RRVIDFS	168-174	5

[a]H22T mutation site shown in parentheses within the peptide sequence
[b]S130P or S130T mutation site shown in parentheses within the peptide sequence
[c]Number of exchangeable amides for the S130P variant is shown in parentheses

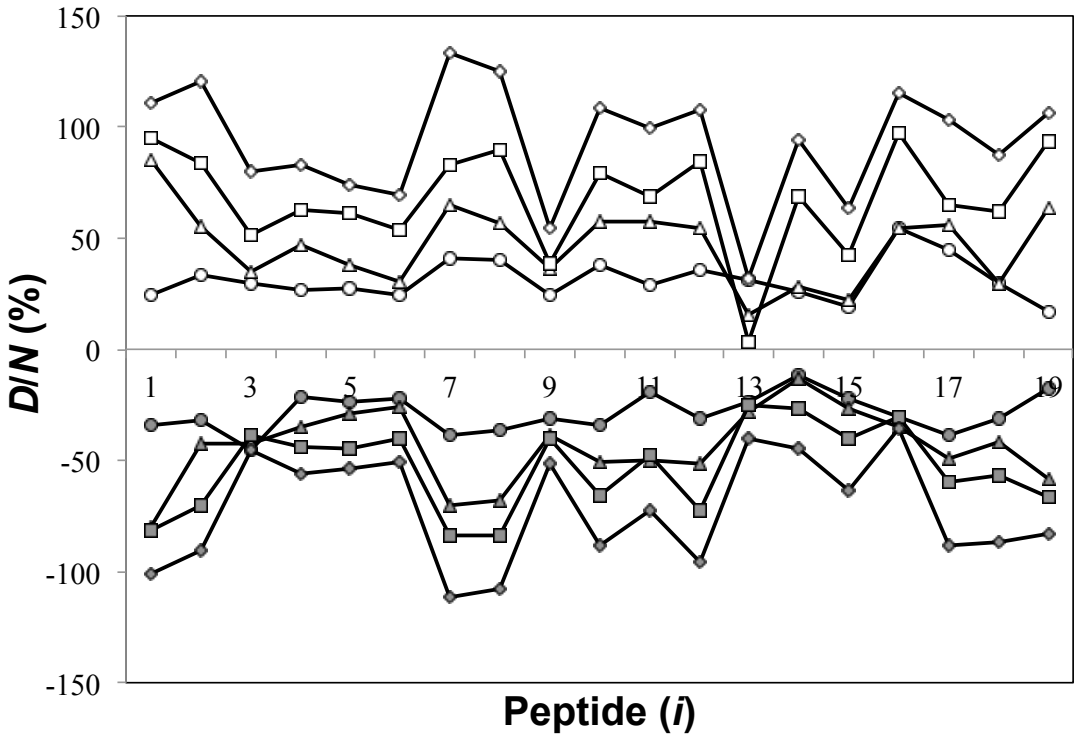


Figure 3.7: Butterfly plot showing the extent of labeling for each reporter peptide, *i*, to visually compare the monomeric structure of S130P (filled symbols) versus wild type (open symbols) γD-crys. Labeling times of 0 min (circles), 12 min (triangles), 120 min (squares), and 1200 min (diamonds) are shown for both proteins.

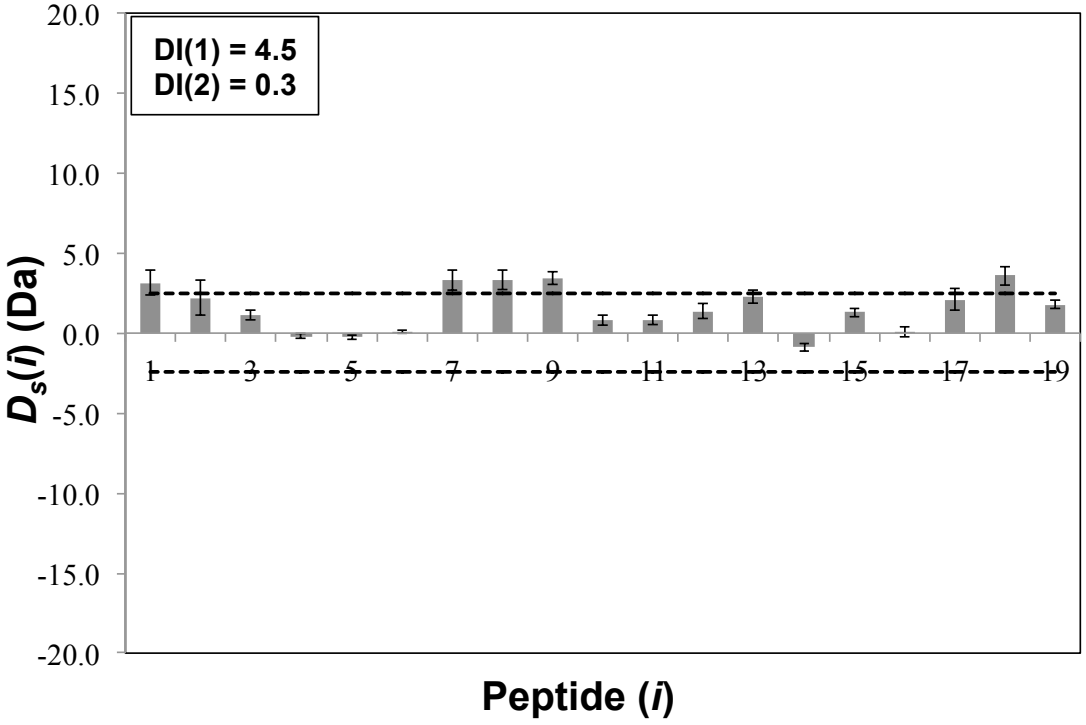


Figure 3.8: Values of $D_s(i)$ plotted for each reporter peptide, *i*, to compare the monomeric structures of S130P and wildtype γD-crys. Negative values indicate less deuterium labeling for S130P relative to wild type, and vice versa for positive values. Error bars represent the standard error in $D_s(i)$ (estimated by one standard deviation). Dashed lines represent a 98% confidence interval of ± 2.5 Da, calculated using Eq. 3.10. Those peptides having $D_s(i)$ values that exceeding this limit were considered statistically different. DI(1) and DI(2) values are calculated using Eq. 3.12 and 3.13, respectively.

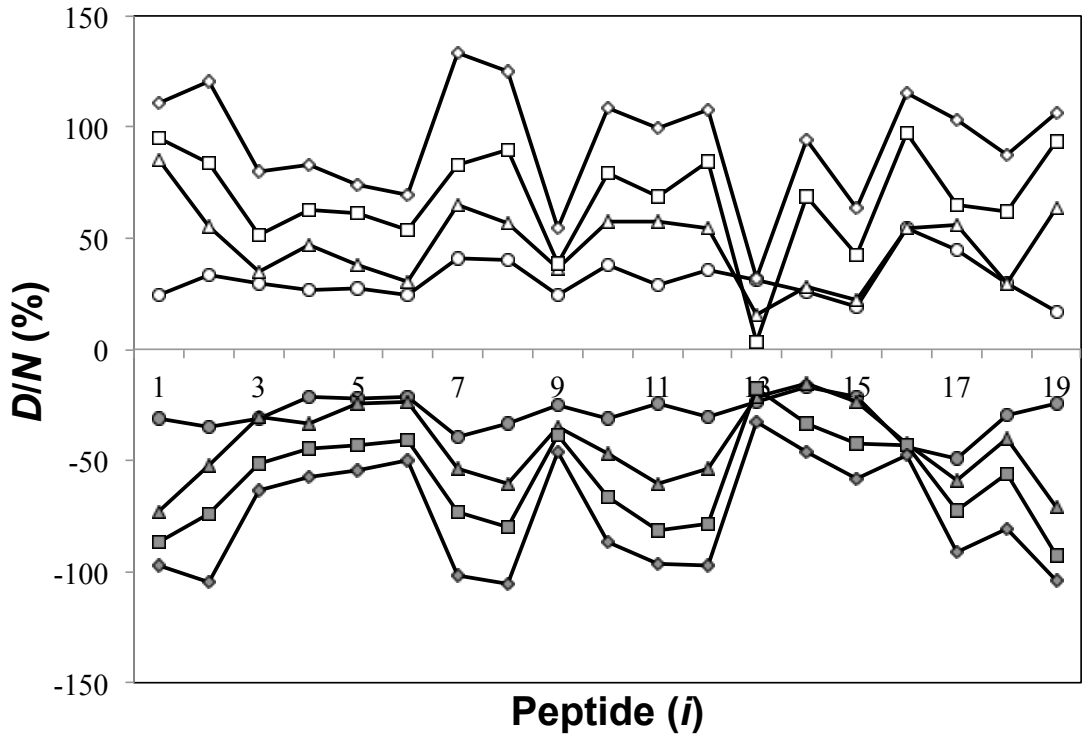


Figure 3.9: Butterfly plot showing the extent of labeling for each reporter peptide, *i*, to visually compare the monomeric structure of S130T (filled symbols) versus wild type (open symbols) γD-crys. Labeling times of 0 min (circles), 12 min (triangles), 120 min (squares), and 1200 min (diamonds) are shown for both proteins.

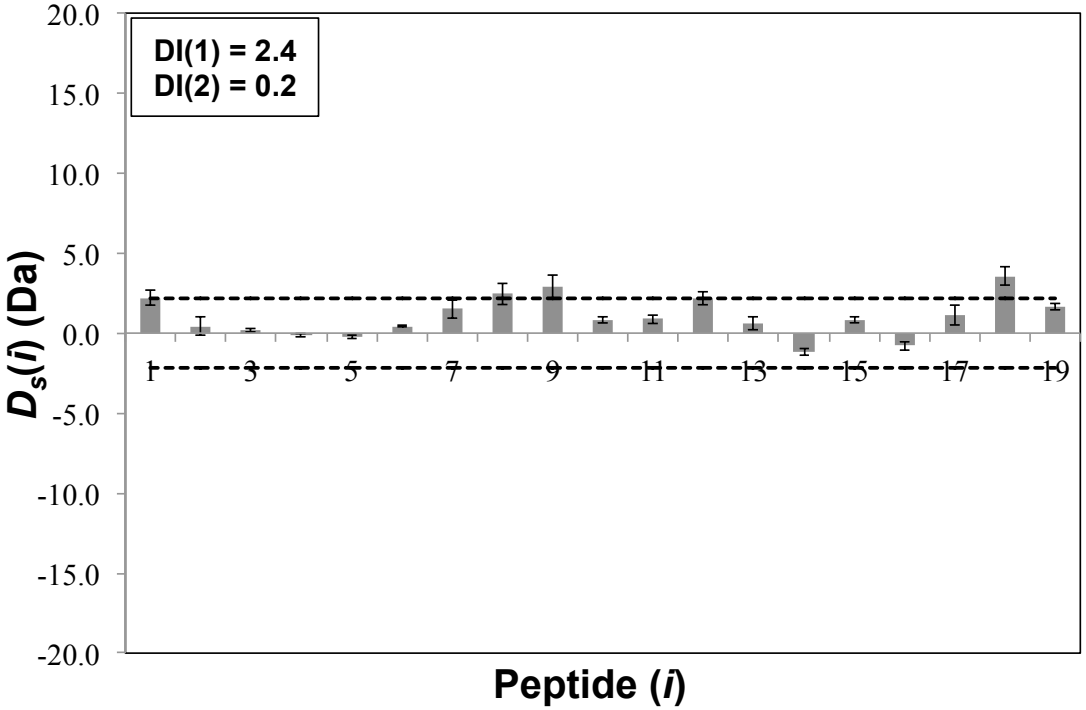


Figure 3.10: Values of $D_s(i)$ plotted for each reporter peptide, *i*, to compare the monomeric structures of S130T and wildtype γD-crys. Negative values indicate less deuterium labeling for S130T relative to wild type, and vice versa for positive values. Error bars represent the standard error in $D_s(i)$ (estimated by one standard deviation). Dashed lines represent a 98% confidence interval of ± 2.2 Da calculated using Eq. 3.10. Those peptides having $D_s(i)$ values that exceeding this limit were considered statistically different. DI(1) and DI(2) values are calculated using Eq. 3.12 and 3.13, respectively.

After visually comparing the butterfly plots in Figures 3.1, 3.7, and 3.9, both similarities and differences in labeling were observed for peptides in each γD-crys variant monomer versus wild type monomer.

For instance, similar trends were shown for peptides 7-8, which displayed more labeling for all γD-crys monomeric species relative to the other reporter peptides. On the other hand, peptides 4-6, 9, and 13 generally exhibited slower deuterium labeling for all γD-crys monomer species. As mentioned, these labeling trends were generally consistent with the accessible surface area of these peptides. However, some differences were also observed. For example, peptides located at the sequence

terminals, such as peptides 1-2 and 17-19, as well as peptides 10-12 generally displayed less labeling in H22T (Figure 3.1) compared to the same peptides in wild type, S130P (Figure 3.7), and S130T (Figure 3.9). Further, slower labeling was displayed for peptide 16 and peptide 3 in H22T and S130P, respectively, compared to the other γD-crys monomer species. Nonetheless, few of these differences were observed to be statistically significant as shown by the bar graphs in Figures 3.2, 3.8, and 3.10. Here, each bar graph indicated relatively small differences in labeling for nearly all reporter peptides compared to the typical 98% CI, and few peptides exceeded the confidence limit.

In addition, the overall labeling differences between two protein species or conformational states was also quantitatively compared by calculating difference indices, DI(1) and DI(2). These are defined in Eq. 3.12 and 3.13, and calculated using the standard deviations for $D_s(i)$ and $D(\Delta M_{i,t})$, respectively.

$$DI(1) = \sum_{i=1}^{19} [abs(D_s(i)) - SD_y] \qquad \text{(Eq. 3.12)}$$

$$DI(2) = \sum_{i=1}^{19} \sum_{t=1}^{4} [abs(D(\Delta M_{i,t})) - SD_x] \qquad \text{(Eq. 3.13)}$$

The indices essentially represent the sum of any differences observed for all reporter peptides analyzed between two protein species or two protein conformations being compared. Larger values observed for these difference indices indicated a larger difference in deuterium uptake existed between the two samples. In previous work by Houde et. *al.,* DI(1) and DI(2) values near zero were reported for structurally comparable samples [Houde 2011]. In this work, the values calculated for DI(1) and DI(2) to compare the monomeric structure of each variant to wild type were all observed to be equal to or less than 6, as shown in Figures 3.2, 3.8, and 3.10. These values were larger than those observed by Houde et. *al.*, but much smaller than the difference indices calculated when

comparing γD-crys aggregate to monomer structures. Thus, they indicate relatively small, overall differences between each monomeric structure. Further, these figures also show the values for the 98% confidence intervals were calculated to be 2.2, 2.5, and 2.2 Da for H22T, S130P, and S130T monomer, respectively, when compared to wild type monomer. This suggested similar levels of uncertainty were also observed for each comparison, however; these uncertainty limits are also slightly larger than the values estimated by Houde et *al*. (1.1 Da) [Houde 2011], and most likely are attributed to testing less sample replicates and differences in instrumentation.

On the other hand, some peptides for each γD-crys variant were observed to have small, yet statistically significant differences in deuterium labeling relative to wild type. For instance, Figure 3.2 showed peptides 1, 3, 7, 10, 12, and 19 were statistically significant between H22T and wild type monomer structures. In this case, the differences were all negative, indicating regions of H22T may be less flexible compared to wild type. These results are consistent with the increased conformational stability observed for H22T relative to wild type as discussed in Chapter 2. Notably, peptide 3 showed a negative difference in labeling and contained the H22T variant site. This may indicate localized conformational stabilization within this region as a result of the variant.

For the S130 variants peptides 1, 7, 8, 9, and 18 showed statistically significant labeling differences between S130P and wild type monomer (Figure 3.8), while peptides 9 and 18 exhibited statistically significant labeling differences between S130T and wild type monomer (Figure 3.10). However, none of these peptides contained the S130 variant site. Nonetheless, the peptides that exhibited statistically significant labeling differences relative to wild type were all positive differences, suggesting both variants contained regions that were more flexible, or solvent exposed, compared to wild type. This observation is also consistent with the decreased conformational stability observed for S130P and S130T relative to wild type as discussed in Chapter 2.

Analyzing aggregate structures of each γD-crys species using HX-MS

The conformational structures of aggregated and monomeric states of each γD-crys species were also compared qualitatively and quantitatively using experimentally obtained HX-MS data.

Qualitative differences in deuterium labeling were identified by visually comparing the butterfly plots shown in Figures 3.11, 3.13, 3.15, and 3.17 for wild type γD-crys, H22T, S130P, and S130T, respectively. Upon examination of the overall deuterium labeling patterns, where all reporter peptides were considered, the aggregates of wild type and S130T experienced slower deuterium labeling, even at long labeling times, relative to monomer. In fact, no notable increases in deuterium labeling for any peptides were visually observed within the butterfly plots for wild type and S130T aggregates compared to the monomer. This observation may suggest the exchange competent, amide hydrogens are located in buried regions of the aggregates, or are involved in hydrogen bonding resulting in significantly slower deuterium exchange.

Alternatively, the aggregates of H22T (Figure 3.13) and S130P (Figure 3.15) showed generally faster labeling compared to S130T and wild type aggregates. Additionally, H22T and S130P aggregates experienced more comparable labeling patterns when compared to the respective monomer states. Notably, the extent of deuterium labeling attributed to the aggregates of H22T and S130P accounted for the presence of monomer in the sample using Eq. 3.11. These observations suggest faster labeling correlates with more monomer remaining in these samples after isothermal incubation, but may also be a result of intermediate, aggregated states (e.g. dimers, trimers, etc.) present in H22T and S130P samples that were not observed in S130T and wild type, as shown by SEC chromatograms (Figures 3.3 and 3.4).

However, notable labeling differences were also visually observed between the aggregates of H22T and S130P as well. For instance, the labeling observed for S130P

aggregates reached a maximum extent after 12 minutes for the majority of reporter peptides. This observation was not as prominent for as many peptides in H22T aggregates. As a result, S130P aggregates may adopt a more flexible structure that permits faster deuterium labeling kinetics compared to H22T aggregates. On the other hand, S130P aggregates also showed some peptides that did not experience fast deuterium labeling relative to monomer. Peptides 2, 3, 11, and 14 exhibited this behavior, even for long labeling times, while more labeling was observed for the same peptides in S130P monomer. Therefore, these peptides were considered potential aggregation contacts.

Further, the butterfly plots also showed some peptides did not exhibit a monotonically increasing labeling pattern with increasing labeling time. This was most commonly observed for H22T and S130P aggregates. For example, peptides 3-9 and in H22T aggregates and peptides 7-9 and 17-19 in S130P aggregates showed less labeling at the 1200 min labeling time compared to the shorter labeling times. This could conceivably result from additional aggregation taking place during longer deuterium labeling times. However, only small changes in the aggregate and monomer composition of each γD-crys species were observed via SEC analysis, even after incubation at room temperature for the longest labeling time (data not shown). Further, during data analysis, low signal to noise ratios were observed for H22T and S130P samples, and a correction was conducted (Eq. 3.11) to account for the monomer remaining in the samples after isotherm aggregation. Both of these matters could conceivably increase the experimental uncertainty of the data, and result in irregular labeling trends.

A more quantitative assessment, including a detailed, statistical analysis, was also conducted for each γD-crys species to distinguish statistically significant labeling differences between aggregated and monomeric states; and as a result identify potential aggregation contacts. The bar graphs in Figures 3.12, 3.14, 3.16, and 3.18, for wild type

γD-crys, H22T, S130P, and S130T, respectively, show these analyses, and are discussed in the following paragraphs.

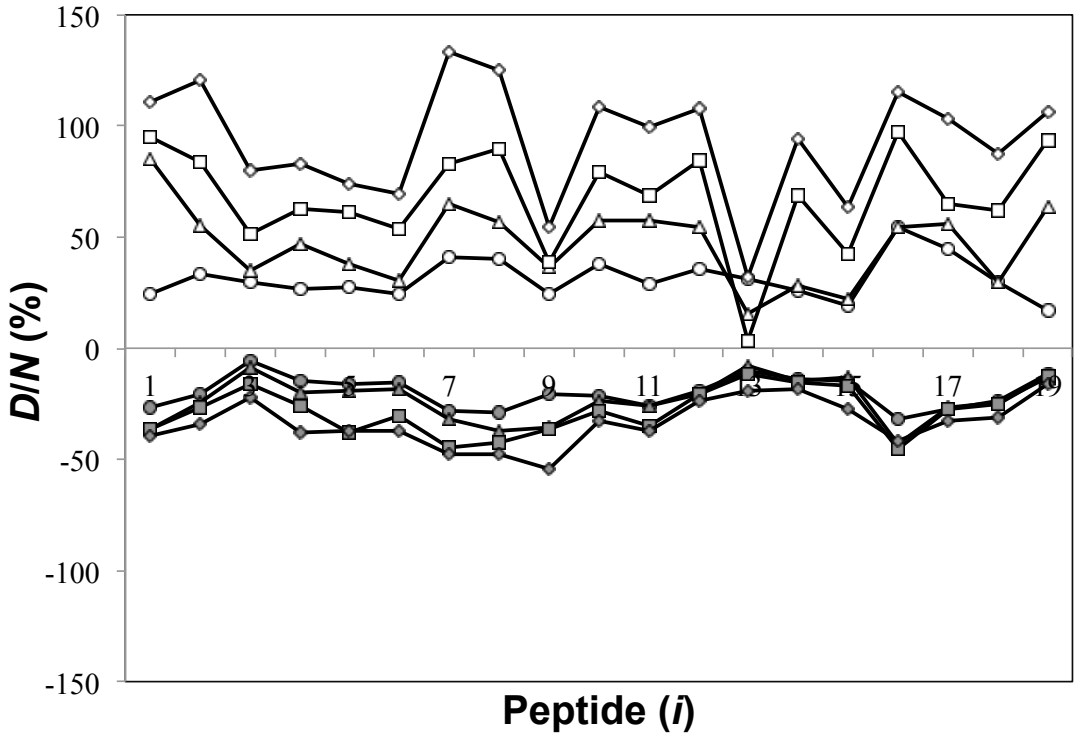


Figure 3.11: Butterfly plot showing the extent of labeling for each reporter peptide, *i*, to visually compare the aggregated structure (filled symbols) vs. monomeric structure (open symbols) of wild type γD-crys. Labeling times of 0 min (circles), 12 min (triangles), 120 min (squares), and 1200 min (diamonds) are shown for both conformational states.

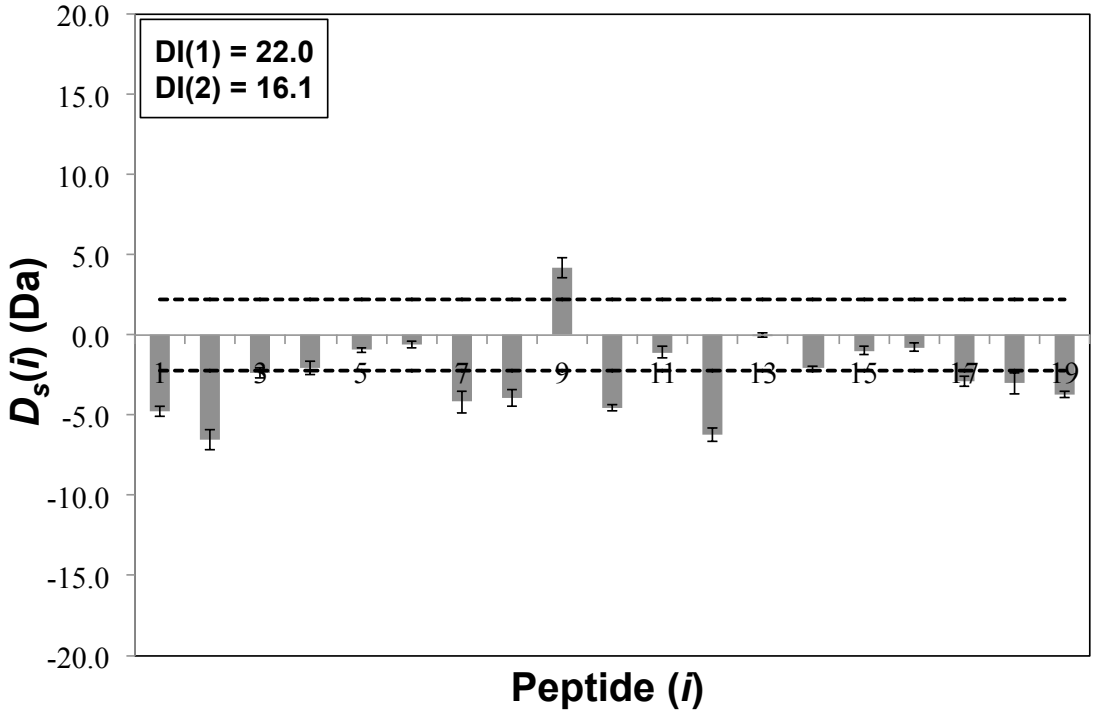


Figure 3.12: Values of $D_s(i)$ plotted for each reporter peptide, *i*, to compare the aggregated and monomeric structure of wild type γD-crys. Error bars represent the standard error in $D_s(i)$ (estimated by one standard deviation). Dashed lines represent a 98% confidence interval of ± 2.2 Da, calculated using Eq. 3.10. Those peptides having $D_s(i)$ values that exceeding this limit were considered statistically different. DI(1) and DI(2) values are calculated using Eq. 3.12 and 3.13, respectively.

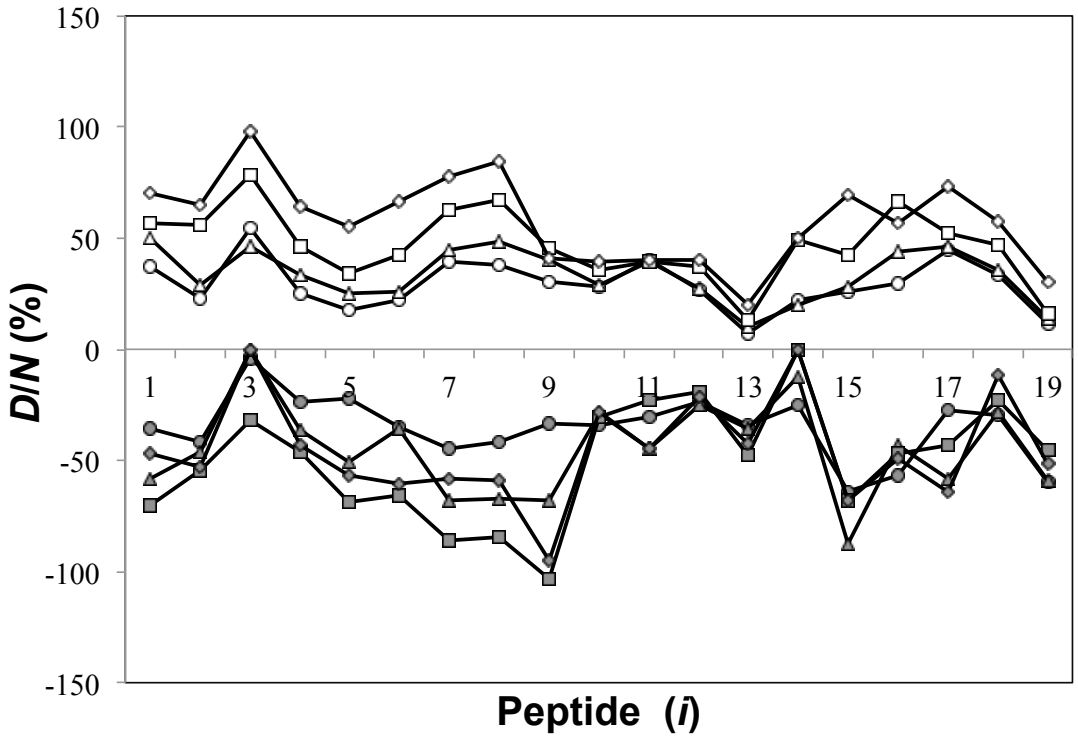


Figure 3.13: Butterfly plot showing the extent of labeling for each reporter peptide, *i*, to visually compare the aggregated structure (filled symbols) vs. monomeric structure (open symbols) of the γD-crys variant H22T. Labeling times of 0 min (circles), 12 min (triangles), 120 min (squares), and 1200 min (diamonds) are shown for both conformational states.

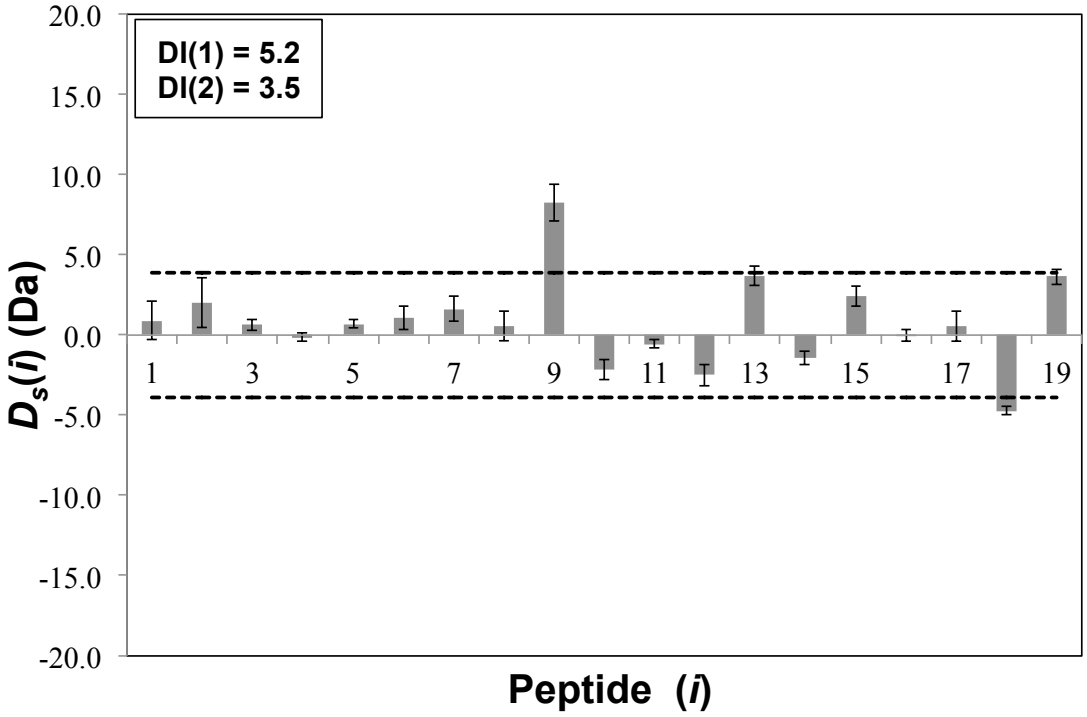


Figure 3.14: Values of $D_s(i)$ plotted for each reporter peptide, *i*, to compare the aggregated and monomeric structure of the γD-crys variant H22T. Error bars represent the standard error in $D_s(i)$ (estimated by one standard deviation). Dashed lines represent a 98% confidence interval of ± 3.9 Da, calculated using Eq. 3.10. Those peptides having $D_s(i)$ values that exceeding this limit were considered statistically different. DI(1) and DI(2) values are calculated using Eq. 3.12 and 3.13, respectively.

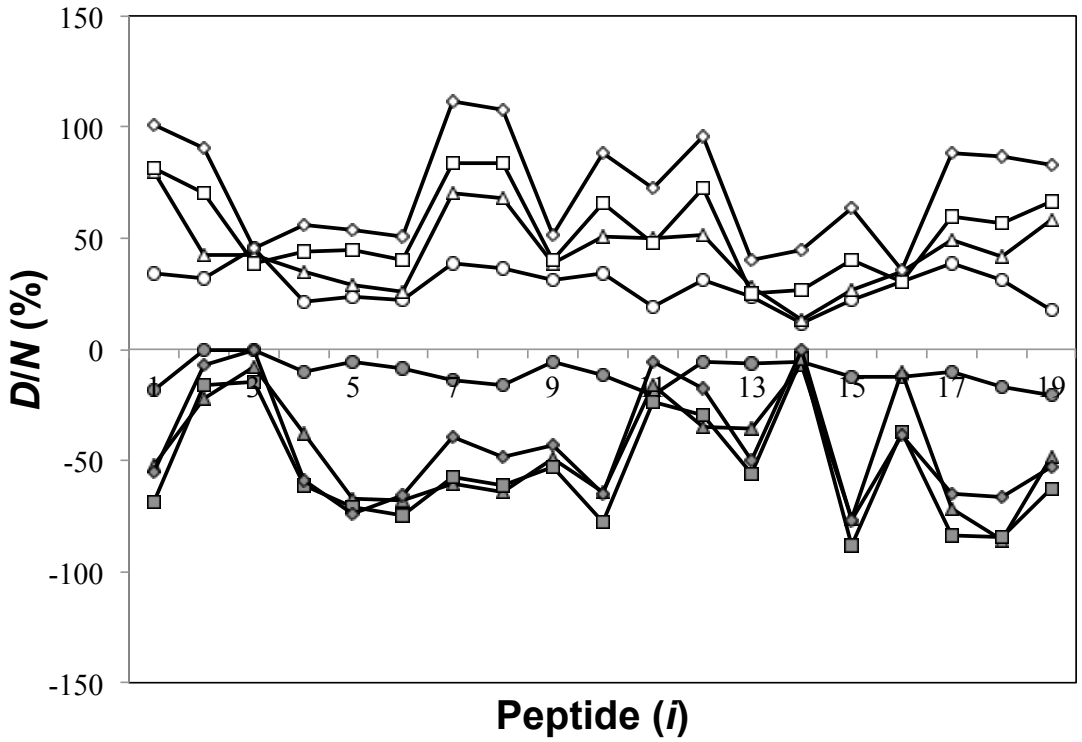


Figure 3.15: Butterfly plot showing the extent of labeling for each reporter peptide, *i*, to visually compare the aggregated structure (filled symbols) vs. monomeric structure (open symbols) of the γD-crys variant S130P. Labeling times of 0 min (circles), 12 min (triangles), 120 min (squares), and 1200 min (diamonds) are shown for both conformational states.

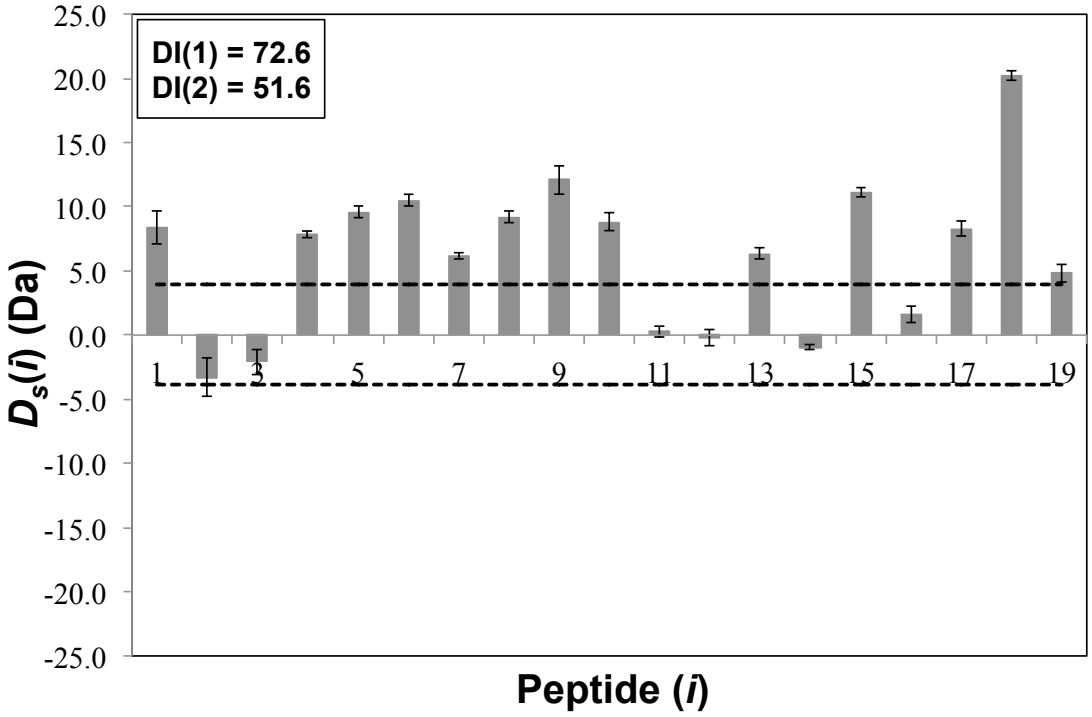


Figure 3.16: Values of $D_s(i)$ plotted for each reporter peptide, *i*, to compare the aggregated and monomeric structure of the γD-crys variant S130P. Error bars represent the standard error in $D_s(i)$ (estimated by one standard deviation). Dashed lines represent a 98% confidence interval of ± 3.9 Da, calculated using Eq. 3.10. Those peptides having $D_s(i)$ values that exceeding this limit were considered statistically different. DI(1) and DI(2) values are calculated using Eq. 3.12 and 3.13, respectively.

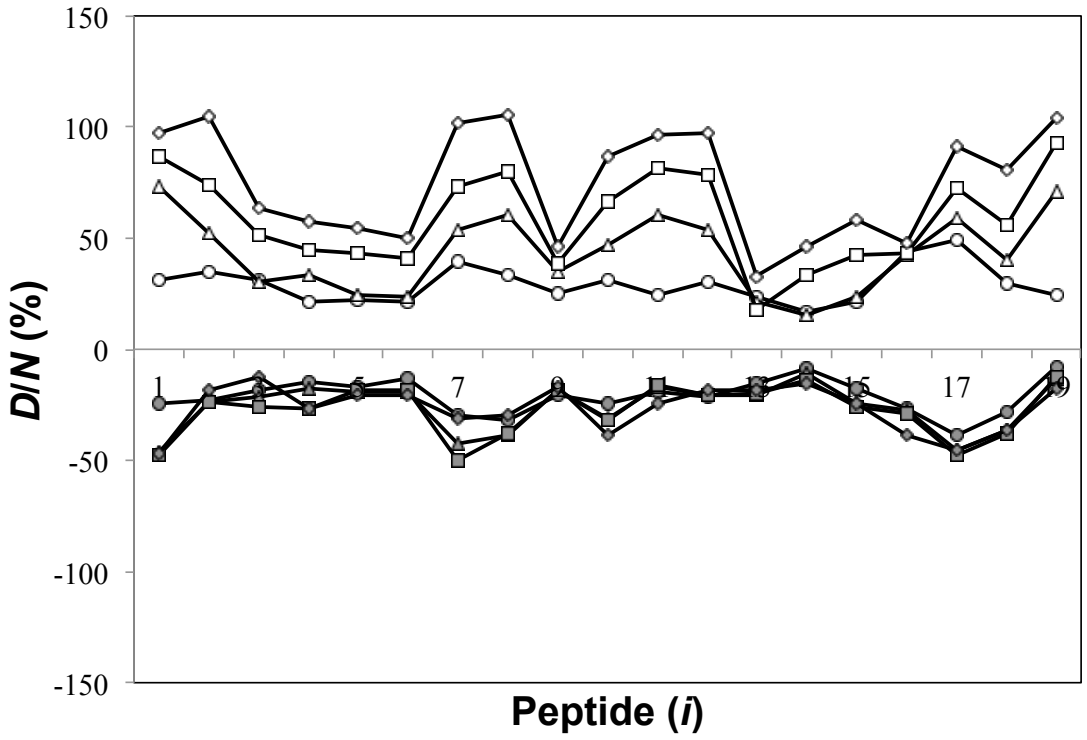


Figure 3.17: Butterfly plot showing the extent of labeling for each reporter peptide, *i*, to visually compare the aggregated structure (filled symbols) vs. monomeric structure (open symbols) of the γD-crys variant S130T. Labeling times of 0 min (circles), 12 min (triangles), 120 min (squares), and 1200 min (diamonds) are shown for both conformational states.

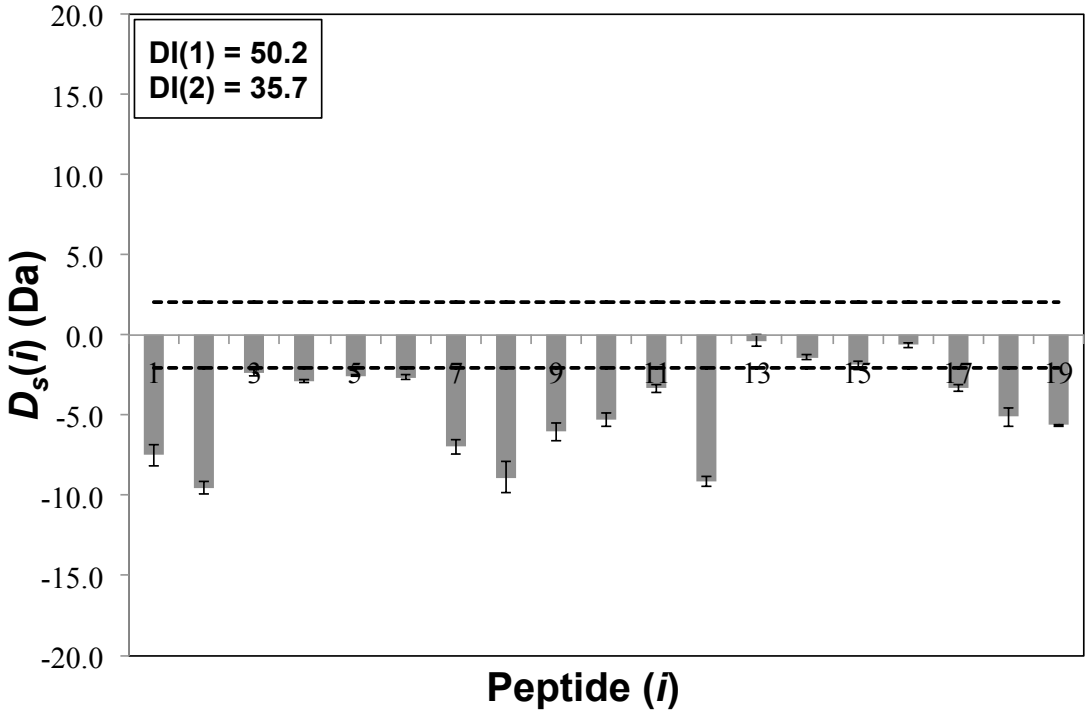


Figure 3.18: Values of $D_s(i)$ plotted for each reporter peptide, *i*, to compare the aggregated and monomeric structure of the γD-crys variant S130T. Error bars represent the standard error in $D_s(i)$ (estimated by one standard deviation). Dashed lines represent a 98% confidence interval of ± 2.1 Da, calculated using Eq. 3.10. Those peptides having $D_s(i)$ values that exceeding this limit were considered statistically different. DI(1) and DI(2) values are calculated using Eq. 3.12 and 3.13, respectively.

For instance, Figure 3.12 showed several reporter peptides for wild type γD-crys (peptides 1, 2, 7, 8, 9, 10, 12, 17, 18, and 19) exceeded the 98% confidence interval of ± 2.2 Da, and thus showed statistically significant differences in deuterium labeling. Notably, most of these statistically significant labeling differences were negative, indicating the aggregate structure is predominantly less flexible and less solvent exposed, except for peptide 9, which exhibited a positive difference. Additionally, the difference indices, DI(1) and DI(2), were 22.0 and 16.1, respectively. These values are significantly larger than the 98% CI as well as the difference indices calculated for monomer comparisons of variant versus wild type γD-crys. Thus, the changes in structure for the aggregates appear by HX-MS to be significantly larger than the

differences between the monomeric variant structures. This is consistent with the CD results (Figure 3.5A & 3.5B). Similar results were observed for S130T, and thus may correlate with the lack of monomer or intermediate aggregated state(s) present in these protein samples as determined by SEC (Figure 3.3).

Less deuterium labeling observed for a particular peptide in an aggregated state relative to the monomeric state indicates that peptide as a potential aggregation contact. Thus, according to these data, all reporter peptides in the wild type aggregates, besides peptide 9, may serve as potential aggregation contacts, although some do not exhibit a statistically significant difference in labeling.

On the other hand for H22T, only peptides 9 and 18 exhibited statistically significant labeling differences between the aggregated and monomeric states (Figure 3.14). Here, peptide 18 displayed a negative difference in labeling and peptide 9 showed a positive differential. The 98% confidence interval estimated for H22T was ± 3.9 Da, larger than the uncertainty limit observed for wild type, and may have reduced the number of peptides exhibiting a statistically significant difference. Further, the peptide containing the H22T variant site, peptide 3, showed only a small positive difference in labeling, and was not statistically significant. The calculated DI(1) and DI(2) values were smaller, 5.2 and 3.5, respectively, compared to indices calculated for aggregate versus monomer comparisons for the other γD-crys species. This may indicate less conformational differences occurred between the aggregates of H22T relative to monomer, and may be consistent with the conformational stabilization observed for this variant as discussed in Chapter 2. Less conformational differences between H22T aggregates and monomer may be a direct result from larger fractions of monomer or intermediate aggregated state(s) present during HX-MS analysis of these protein samples as determined by SEC (Figure 3.4). However, as mentioned, the 98% CI interval was also larger here, and most likely reduced the number of peptides that

displayed a statistically significant difference in labeling. The larger CI observed for H22T was most likely caused by lower signal to noise ratios during data analysis, that generated larger experimental errors.

Peptide 18 was the only peptide that exceeded the 98% CI, and thus was the only potential aggregation contact (with statistical significance) observed in the H22T aggregates. Other peptides in the C-td also displayed negative labeling differences (peptides 10, 11, 12, and 14), but were not statistically significant. Notably, peptide 14 contained the S130 variant site. On the other hand, positive (though statistically insignificant) labeling differences were observed for all peptides located in the N-td (peptides 1-9). This may suggest the aggregation contacts for H22T are limited to the C-td, although additional labeling experiments, involving more replicates and more reporter peptides, would be valuable to more confidently support this observation.

For S130P, the statistical analysis showed most of the reporter peptides (peptides 1, 4-10, 13, 15, and 17-19) exhibited positive, statistically significant differences in deuterium labeling (Figure 3.16). The 98% confidence interval for S130P was estimated to be ± 3.9 Da, similar to H22T, and larger than the value estimated for wild type. Further, the calculated DI(1) and DI(2) values were 72.6 and 51.6, respectively, the largest of all the γD-crys species tested. These results indicate large increases in conformational flexibility or solvent exposure for S130P aggregates relative to monomer. Further, the positive differences in labeling may also have resulted from larger fractions of monomer or intermediate aggregated state(s) present during HX-MS analysis of S130P aggregated samples as determined by SEC (Figure 3.4), and may be consistent to the aggregation resistance observed for this variant as discussed in Chapter 2.

On the other hand, peptides 2, 3, and 14 showed negative labeling differences in S130P aggregates, but were not statistically significant. Notably, peptide 14 contained

the S130P variant site. Negative labeling differences could indicate these peptides are aggregation contacts for S130P, but additional labeling experiments may be needed to produce statistically significant results.

Finally for S130T, the statistical analysis showed all reporter peptides exhibited a negative difference in deuterium labeling between the aggregated state(s) and monomer (Figure 3.18), with most of these being statistically significant (peptides 1-12 and 17-19). Therefore, conceivably all of these peptides could act as an intermolecular contact. The 98% confidence interval estimated for S130T was ± 2.1 Da, similar to the value calculated for wild type. Further, the difference indices, DI(1) and DI(2), were quite large, 50.2 and 35.7, respectively. These results suggested large conformational differences also occurred between the aggregated state(s) of S130T and the monomeric state. However, unlike S130P, these differences highlighted an aggregate conformation(s) that was less solvent exposed, or more structurally rigid, than the monomeric conformation.

Computationally predicting aggregation "hot spots" and comparing them to experimental HX-MS results

Next, potential, aggregation-prone segments ("hot spots") of primary sequence for wild type γD-crys and each variant were computationally predicted using the aforementioned aggregation calculators. These predictions were then compared to the aggregation contacts observed experimentally via HX-MS. Figure 3.19A-C show the predicted, consensus "hot spots" for wild type γD-crys, S130P, and S130T sequences, respectively. A consensus "hot spot" was determined if a majority of the aggregation calculators predicted it to be aggregation-prone. The sequence for H22T was not shown as the variant did not alter any of the aggregation-prone "hot spots" relative to wild type, and therefore was identical to Figure 3.19A.

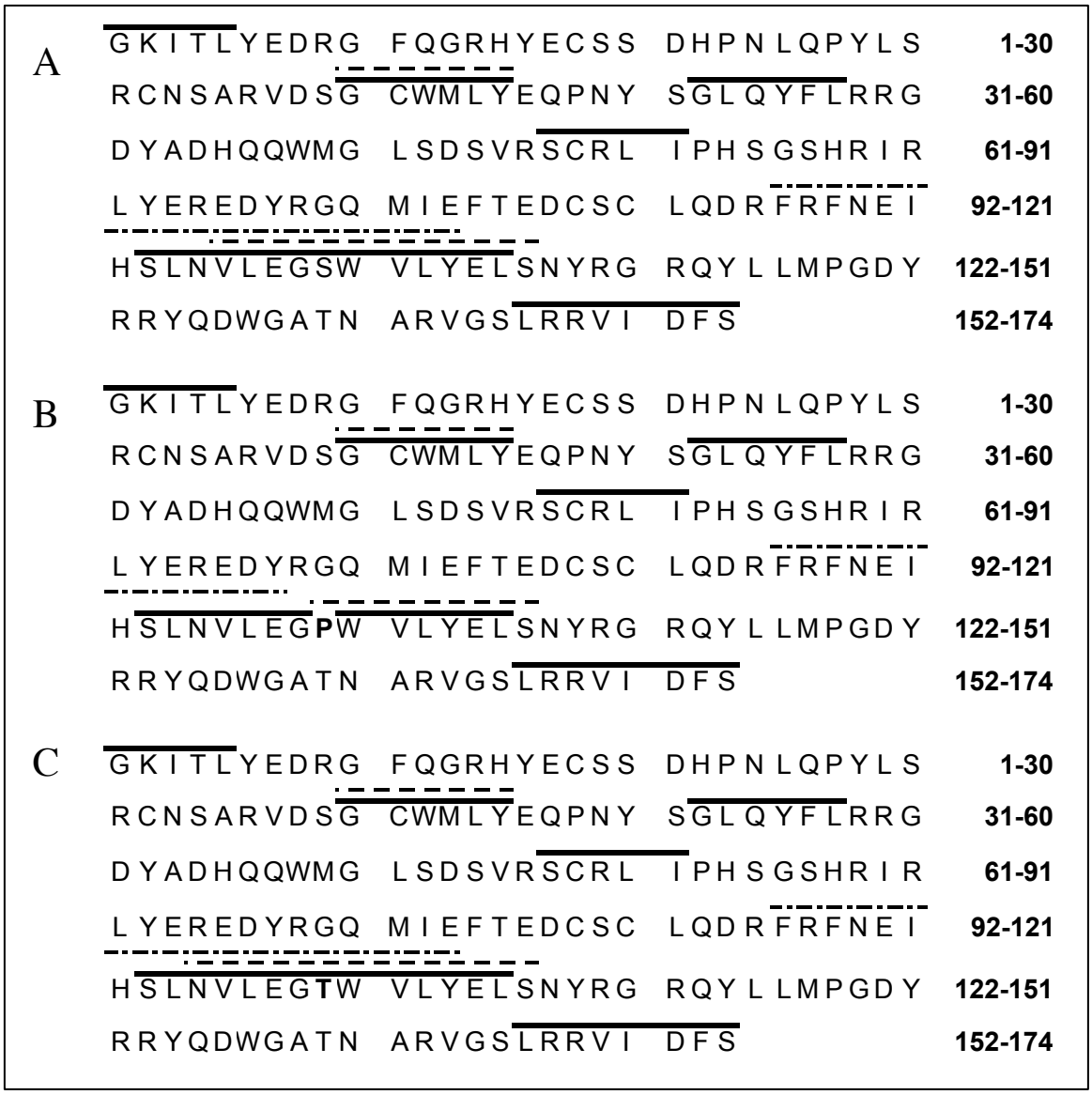


Figure 3.19: Potential aggregation-prone segments, "hot spots", of γD-crys sequence predicted by the three aggregation calculators for A) wild type γD-crys, B) S130P, and C) S130T. Predicted "hot "spots" are denoted by lines above the sequence for AGGRESCAN (solid line), PASTA (dash-dotted line), and TANGO (dashed line). The variant S130 sites are bolded, and the sequence for H22T was not shown because the predicted "hot spots" were identical to wild type.

For wild type γD-crys, Figure 3.19A show potential, aggregation-prone "hot spots" were predicted for residues G40-Y45 on the N-td and S123-L136 on the C-td. In fact, all three calculators agreed residues V126-Y134 were prone to aggregate. However, for S130P, the "hot spot" located on the C-td was broke into two regions,

S123-E128 and W131-L136, that were only agreed upon by 2 of the 3 calculators and did not contain the S130P variant site. S130T was identified by TANGO to increase the aggregation-prone “hot spot” on the C-td by one amino acid, resulting in a larger region consensus “hot spot” containing residues N125-Y134. These results may qualitatively indicate a reduction in IAP for S130P, but an increase for S130T relative to wild type.

The predicted “hot spots” for each γD-crys species were then compared to the HX-MS results. First, only reporter peptides that exhibited statistically significant differences in deuterium labeling between aggregates and monomer for each γD-crys species were considered. Of these, only peptides 4, 5, and 6 (corresponding to residues N33-Y55) in the S130T aggregates included residues that the aggregation calculators also predicted to be a potential, aggregation-prone “hot spot”. However, no other peptides displaying a statistically significant difference in labeling between the aggregated state(s) and monomer state correlated with the computational predictions for any of the other γD-crys species.

Peptides exhibiting negative differences in deuterium labeling for the majority of the γD-crys species tested were also compared to the computational predictions. Using this criterion, peptides 2, 3, and 14 were the only peptides found to having negative differences in deuterium labeling between the aggregates and monomer for 3 of the 4 γD-crys species tested. Thus, these three peptides were flagged as potential aggregation contacts that may be important to the aggregation mechanism of multiple γD-crys species. Specifically, peptide 14 (corresponding to residues N125-L133) displayed negative differences in deuterium labeling for all γD-crys species tested. However, the difference was smaller for S130P (-0.9 ± 0.2 Da in Figure 3.16) compared to wild type (-2.1 ± 0.2 Da in Figure 3.12) and S130T (-1.4 ± 0.2 Da in Figure 3.18). This may indicate reduced intermolecular contacts are formed at this region as a result of the S130P variant, although more certain statistical differences are needed to more

confidently validate this observation. Further, the aggregation calculators predicted a potential “hot spot” for residues included in peptide 14 for each γD-crys species; however, a reduction in IAP at this “hot spot” was suggested for S130P relative to wild type and S130T, potentially correlating with the experimental results.

On the other hand, peptides 2 and 3 (corresponding to residues Y6-L29 and S20-L29, respectively) were only observed to have negative differences in deuterium labeling for wild type, S130P, and S130T, and notably, both peptides contained the H22T variant. This observation may suggest the insertion of H22T reduces the aggregation propensity for this region, although the aggregation calculators did not predict an aggregation-prone “hot spot” corresponding to peptides 2 and 3.

3.4.Discussion

The aggregate conformations of three γD-crys variants (H22T, S130P, and S130T) were analyzed and compared to their respective monomer conformations as well as to the monomer and aggregate conformations of wild type γD-crys using hydrogen-deuterium exchange coupled with mass spectrometry (HX-MS). This technique has been shown in previous studies to be successful in analyzing aggregates of several different proteins [Zhang 2011, Zhang 2010, Tobler 2002, Qi 2009, Kheterpal 2006]. Utilizing HX-MS to analyze the tertiary structure of proteins in addition to spectroscopic methods, for instance, is valuable as HX-MS can examine the entirety of the molecule with peptide-level resolution. On the other hand, spectroscopic methods, such as fluorescence spectroscopy and circular dichroism (CD) are limited to the location of intrinsic or extrinsic probes within the molecule.

The reliability and validity of these HX-MS data obtained for these studies was analyzed by comparing the extent of deuterium labeling observed experimentally for each reporter peptide to the average area of solvent accessibility (ASA), estimated by

ASAview [Ahmad 2004], for each peptide. Generally speaking, those reporter peptides located on the surface or periphery of each γD-crys species tested were observed to have higher ASA values and extents of deuterium labeling compared to more buried peptides located within the domain cores or within the domain interface (Figure 3.6). Thus, a general correlation did seem to exist between the extent of deuterium labeling and solvent exposure for a given reporter peptide. This exemplifies the reliability and usefulness for utilizing HX-MS to identify aggregation contacts within γD-crys by monitoring changes in the solvent exposure of peptides upon aggregation, as has also been done successfully for other proteins in previous work [Zhang 2011, Tobler 2002, Qi 2009, Kheterpal 2006].

Similar monomeric conformations observed for all γD-crys species tested

The monomeric structures of each γD-crys variant were first analyzed and compared relative to the wild type to identify any potential structural differences. Large perturbations within the monomeric structures were not expected for any of the variants tested because only single, point mutations were inserted into the sequence; rather than multiple mutations that may affect the protein conformation more dramatically. Far-UV circular dichroism (CD) spectra showed the secondary structure of each variant was similar to the wild type (Figure 3.5A). Additionally, the far-UV CD spectra exhibited by each monomeric, γD-crys species indicated β-sheets were the predominant form of secondary structure populating the protein conformation at these solution conditions. Similar results for γD-crys were also shown in previous work at these conditions [Sahin 2011] and at more neutral conditions [Flaugh 2005a, Flaugh 2005b, Kosinski-Collins 2003, Kosinski-Collins 2004, Mills 2007, Acosta-Sampson 2010]. Further, examination of fluorescence intensity as a function of denaturant for these γD-crys species (Figure 2.5 and 2.7) showed the starting spectra, at low denaturant concentrations, were all similar.

Together, this suggested the secondary and tertiary structure of monomeric H22T, S130P, and S130T was similar to monomeric, wild type γD-crys.

The HX-MS experimental results also indicated similar monomeric conformations for all γD-crys species tested. This was observed qualitatively by visually comparing the deuterium labeling patterns for each γD-crys variant relative to wild type shown in the butterfly plots (Figures 3.1, 3.7, and 3.9), developed by Houde et *al*., [Houde 2011]. Similarities were also observed quantitatively using a statistical analysis, also developed by Houde et *al*. [Houde 2011]. Few reporter peptides displayed statistically significant differences in deuterium labeling when comparing the monomeric states of the variant to wild type (Figure 3.2, 3.8, and 3.10), and those differences that were statistically significant were small in magnitude. Furthermore, the values calculated for DI(1) and DI(2) (Eq. 3.12 and 3.13, respectively) to compare the monomeric states of each γD-crys species were less than 6.0, indicating a relatively small degree of difference compared to other indices calculated when comparing aggregates to monomers. Houde et *al*. observed smaller values for difference indices when comparing protein samples that they concluded to be highly comparable, but notably in their work they tested more sample replicates, thus difference indices observed here for γD-crys monomers would likely decrease if additional replicates were tested.

Altered aggregated conformations observed for each γD-crys species tested

Conformational differences were also evaluated between the aggregates of each γD-crys species, as well as between the aggregates and monomers of each γD-crys species. Far-UV CD and HX-MS were used to assess changes in the secondary and tertiary structure of the aggregated states, respectively, for each protein variant and wild type.

The far-UV CD spectra indicated notable changes occurred in the secondary structure of the aggregated state(s), relative to each monomeric state for all γD-crys species tested (Figure 3.5A & 3.5B). For example, the CD spectra for the aggregated state(s) of each γD-crys species showed the minimum ellipticity shifted to a lower wavelength compared to the respective monomeric states. Similar changes in far-UV CD spectra were observed in previous work for aggregates of γD-crys variants relative to monomer at these conditions [Sahin 2011]. Additionally, the far-UV CD spectra for aggregates of the different γD-crys species were also compared. A more negative minimum in ellipticity at 218 nm was observed for wild type and S130T aggregated state(s) compared to those of H22T and S130P (Figure 3.5B). This may suggest increased β–sheet structure is present in S130T and wild type γD-crys aggregated state(s) compared to H22T and S130P. Together, these data suggests differences do exist between γD-crys aggregated state(s) versus each respective monomer state, but also that differences may be apparent between the aggregated state(s) of each γD-crys species as well.

HX-MS was used as another technique to identify conformational changes taking place between the aggregated state(s) of each γD-crys species, as well as conformational changes occurring relative to the respective monomeric states. These differences were observed qualitatively by visually comparing the differences in deuterium labeling within the butterfly plots (Figures 3.11, 3.13, 3.15 and 3.17); and also quantitatively by conducting a detailed, statistical analysis (Figures 3.12, 3.14, 3.16, and 3.18).

Upon visual examination of the overall deuterium labeling patterns for each γD-crys species, slower labeling was observed for peptides in wild type and S130T aggregates, even for long labeling times, compared to monomer, and compared to the H22T and S130P aggregates. Thus, these statistical analyses indicated most peptides in

wild type and S130T aggregates exhibited more statistically significant, negative labeling differences relative to monomer; while the aggregates of H22T and S130P displayed more statistically significant, *positive* labeling differences relative to their respective monomeric state (Figure 3.14 and 3.16, respectively). Thus, the patterns of deuterium labeling for the aggregated state(s) seem to be qualitiatively consistent with the aggregation behavior shown by each γD-crys species as discussed in Chapter 2.

Further, difference indices, DI(1) and DI(2) were calculated within the statistical analysis developed by Houde et *al.* [Houde 2011]. Values for DI(1) and DI(2) were used to quantitatively assess the overall differences in deuterium labeling between aggregates and monomers of each γD-crys species. Larger values of DI(1) and DI(2) indicate larger conformational differences between the two samples being compared, and vice versa [Houde 2011]. DI(1) and DI(2) values calculated to compare the monomeric structures were less than 6.0 for each γD-crys species (Figures 3.1, 3.7, and 3.9), indicating a small degree of difference. However, DI(1) and DI(2) values calculated to compare the aggregates to monomer were much larger for wild type, S130P, and S130T, indicating a more significant degree of difference. However, DI(1) and DI(2) values calculated to compare the H22T aggregates to the monomer were also less than 6.0, suggesting less significant differences occurred between the two conformations (Figure 3.14). Thus, quantitatively the aggregates also appear to be significantly different than monomer, particularly for wild type, S130P, and S130T. Fewer differences observed quantitatively for H22T may be consistent with the improved conformational stability for this variant as discussed in Chapter 2.

The globally reduced deuterium labeling observed for wild type and S130T aggregated state(s), even at long labeling times, may suggest the formation of well-structured, amyloid-like aggregates for these two γD-crys species. Previous studies of amyloid-forming cases have also reported globally reduced solvent exposure during

hydrogen-deuterium exchange for amyloid-β [Qi 2008], α-synuclein [Del Mar 2005], insulin [Dzwolak 2006], and β2-microglobulin [Hoshino 2002]. A single, high molecular weight species was predominantly present in the wild type and S130T HX-MS samples, as determined by SEC analysis that showed less than 5% monomer remaining (Figure 3.3). Further, previous work has shown γD-crys is capable of forming amyloid-like aggregates [Papnikolopoulou 2008], and amyloid formation has been associated with increased hydrogen bonding involving the protein backbone [Tsemekhman 2007, Zheng 2006]. Thus, the slower deuterium labeling kinetics observed for wild type and S130T peptides in aggregates may be caused by significantly decreased solvent exposure and increased hydrogen bonding, particularly involving the backbone amides, that ultimately contribute to well-structured, amyloid-like aggregates.

On the other hand, faster deuterium labeling kinetics observed for H22T and S130P may be attributed to the lack of well-structured aggregates forming for these two species. SEC analysis showed approximately 40% monomer was remaining in incubated protein samples used for HX-MS, and the presence of intermediate aggregated state(s) for both species as well (Figure 3.4). The aggregation resistance for H22T was attributed to a conformational stabilization of the N-td, and decreased aggregation propensity for S130P as discussed in Chapter 2. Thus, increased resistance for H22T to unfold or S130P to aggregate may result in more flexible, or more solvent exposed, aggregate conformations that exhibit faster exchange kinetics compared to the more structured aggregates assumed to form for wild type and S130T.

These HX-MS data for S130P aggregates shows evidence for the formation of more flexibles, less-structured conformations. Here, peptides from the S130P aggregates displayed faster deuterium labeling that reached a maximum after approximately 12 minutes (Figure 3.15). This was significantly different than the labeling kinetics observed for wild type and S130T aggregates. Additionally, peptides in the H22T

aggregates also exhibited faster deuterium labeling compared to wild type and S130T; however, certain peptides displayed slower exchange kinetics compared to S130P (Figure 3.13). Notably, these peptides exhibiting slower exchange kinetics were predominantly located in the N-td (peptides 5 through 9), which may be explained by the conformationally stabilizing behavior of the H22T variant to the N-td as also discussed in Chapter 2.

Identification of aggregation contacts using HX-MS

A primary objective of this work was to experimentally identify the regions or residues responsible for forming intermolecular contacts that contribute to the aggregation of the γD-crys species analyzed. The ability of HX-MS to detect changes in solvent exposure of a peptide upon protein aggregation allows potential aggregation contacts to be identified. For instance, a surface exposed peptide in a native, folded protein state may experience more deuterium labeling than if that same peptide served as an intermolecular contact in an aggregated state. On the other hand, a peptide normally buried in the native, folded state would experience less labeling than if that same peptide experienced local unfolding, became exposed, and was not involved in intermolecular contacts contributing to aggregation.

Of course, reporter peptides contain a mixture of residues, some of which could be involved in intermolecular contacts within an aggregate, and some of which could be disordered. Thus, while increases in labeling of a reporter peptide would indicate a substantial portion of the reporter residues have become more solvent exposed, a minority of the residues could be involved in intermolecular contacts. Further, while intermolecular contacts have been shown to reduce deuterium exchange rates in amyloids (Kheterpal 2006, Wang 2003) and protein-ligand binding interactions (Hopper 2009; Paterson 1990), such interactions involve both a high degree of shape

complementarity and highly favorable equilibrium constants. Taken together, sites of reduced labeling determined by HX-MS would appear safe to ascribe to intermolecular contacts. Nonetheless, some peptides could display similar deuterium labeling patterns in both the native and aggregated state, and thus would be difficult to identify as aggregation contacts using this method.

The HX-MS experiments conducted here showed several reporter peptides in wild type and S130T aggregates exhibited significantly less deuterium labeling than monomer (Figure 3.11-3.12 and 3.17-3.18, respectively). These data suggests much of the molecule is involved in the aggregation mechanism, and is not particularly consistent with the hypothesis that only a relatively short sub-sequence of residues with increased IAP contributes to the aggregation process. The results also do not indicate a strong preference for either less protection or more protection of the N-td or C-td. However, it is also conceivable that non-native intramolecular interactions could result in decreased deuterium labeling of certain reporter peptides, as discussed in previous work [Routledge 2009]. These results and possibilities make it difficult to single out individual peptides or even larger regions that could potentially serve as aggregation contacts for wild type and S130T.

On the other hand, results from H22T showed several peptides located in the C-td displayed less deuterium labeling in the aggregates relative to the monomer, however, only peptide 18 showed a statistical significance. This may suggest aggregation contacts for H22T are limited to the C-td, but this remains inconclusive because of the lack of statistical significance for many of these peptides. Further, results for S130P showed only peptides 2, 3, and 14 displayed less deuterium labeling in aggregates relative to the monomer, but none were statistically significant. Here, the lack of any aggregation contacts may highlight the resistance of S130P to aggregate.

Also noteworthy is the fact that some peptides were observed to have negative differences (although statistically insignificant) in deuterium labeling for the majority of γD-crys species tested. This was the case for peptides 2, 3, and 14. Specifically, peptide 14 displayed this behavior across all γD-crys species tested, while peptides 2 and 3 exhibited this behavior for wild type, S130P, and S130T. These results indicated peptides 2, 3, and 14 could conceivably contribute to aggregate formation and are important to the aggregation mechanism of many γD-crys species. Nonetheless, results remain somewhat inconclusive due the lack of statistical significance observed. As such, further HX-MS studies analyzing more sample replicates would be valuable to reduce uncertainties. Further, using MS or NMR in future studies to analyze many more reporter peptides would also be valuable to try and achieve higher peptide-level or even single residue-level resolution.

Correlating computationally predictions to experimental results

Another promising avenue for determining aggregation-prone regions ("hot spots") within a protein sequence is via computational prediction. Several primary-sequence-based aggregation calculators were implemented here to predict regions of γD-crys sequence that have the potential to aggregate, specifically into amyloid-like structures. These predictions were then compared to experimental results obtained via HX-MS for reach variant and wild type. For wild type γD-crys, two "hot spots" were predicted, one in the N-td from G40-Y45 and one in the C-td ranging from S123-L136. When the calculators were applied to the H22T variant, no differences relative to wild type were observed, suggesting no alteration in the IAP. For S130T, the length of "hot spot" segment in the C-td was increased, suggesting an increased IAP. However, for S130P the same "hot spot" was broken into two "hot spots" at the variant site, suggesting a possible decreased IAP. For S130P, this was not particularly surprising as

proline is largely considered to be an amyloid breaker in protein mutagenesis [Williams 2004].

These two computationally predicted “hot spots” were first compared to the potential aggregation contacts identified via HX-MS. Here, those peptides that exhibited a statistically significant difference in deuterium labeling in the aggregated state versus the monomer state were considered potential aggregation contacts for each γD-crys species (Figure 3.12, 3.14, 3.16, and 3.18). Table 3.1 shows 6 reporter peptides contained residues predicted by the aggregation calculators to be potential “hot spots”. Of these, the reporter peptides 4 (N33-M43), 5 (L44-Q54), and 6 (L44-Y55) contained residues that were included in the “hot spot”, G40-Y45, located in the N-td, and peptides 13 (N119-L127), 14 (N125-L133), and 15 (Y134-Q143) contained residues that were included in the “hot spot”, S123-L136, located in the C-td. However, the only correlation between these predicted “hot spots” and a statistically significant aggregation contact observed experimentally was for peptides 4, 5, and 6 in S130T.

If the criteria used to determine an aggregation contact disregarded statistical certainties, and only required a negative difference in deuterium labeling between aggregates and monomers, then more correlations were observed for S130T and the other γD-crys species. For instance, peptides 4-6 and 13-15 all displayed negative labeling differences in S130T, and thus were consistent with both computationally predicted “hot spots”. Likewise, the peptides 4-6, 14, and 15 in wild type γD-crys also displayed negative labeling differences and correlated with both “hot spots”. However, only peptide 14 exhibited a negative difference in deuterium labeling for H22T and S130P, which only correlated with the “hot spot” on the C-td.

Although the correlations with these peptides are promising, the results are somewhat inconclusive because of the lack of statistical significance, based on the 98% confidence intervals, used in the statistical analysis of HX-MS data. The uncertainties

calculated in this work were slightly larger than those reported by Houde et *al*. and are likely attributed to fewer replicates being tested and instrumentation differences [Houde 2011]. Thus, it may be valuable to test more than three sample replicates to reduce these statistical uncertainty limits; or rather, implement a statistical analysis using a lower α-value than was used by Houde et *al*. as well as in this work (α = 0.05) to try and increase the number of peptides that exceed the confidence interval [Houde 2011].

Nonetheless, peptide 14, located on the C-td, was observed to have a negative difference in labeling across all γD-crys species tested, and notably, contained the S130 variant site. Although previous work has suggested the aggregation-prone intermediate species of γD-crys contains an unfolded N-td and a folded C-td, conceivably aggregation could occur via residues contained in peptide 14 that are exposed when the interface between N-td and C-td is not intact. It was also noteworthy that the difference in deuterium labeling for peptide 14 was less negative for S130P compared to wild type and S130T. This conceivably could indicate S130P was less prone to form intermolecular contacts at this region, compared to S130T and wild type. If so, a successful correlation was observed between experiment and the reduced IAP predicted computationally for S130P. Nonetheless, the differences were not dramatic, and thus further work should be conducted to improve the statistical uncertainties regarding these data.

Peptides 2 (Y6-L29) and 3 (S20-L29) exhibited negative differences in deuterium labeling as well, but only for wild type, S130P, and S130T. As mentioned, the fact that these results are observed across three species suggests they may contribute to aggregate formation and are important to the aggregation mechanism of γD-crys. Conceivably, aggregation-prone residues included within these peptides could become exposed upon unfolding of the N-td, and subsequently allowing aggregation to take place. However, the aggregation calculators did not predict a “hot spot” corresponding to

any of the residues included in peptide 2 and 3. Thus, correct and incorrect predictions were observed here indicating a moderate success rate for these aggregation calculators in this work. This moderate predictive yield is similar to success rates shown in previous work with γD-crys as discussed in Chapter 2, and further indicates the need for improved and optimized computational algorithms.

3.5. Conclusions

The aggregation mechanism of three γD-crys variants, all displaying a range of aggregation behavior relative to wild type, were examined using hydrogen-deuterium exchange coupled with mass spectrometry (HX-MS). Visual comparisons and a statistical analysis were conducted to compare and contrast the conformation of aggregated and monomeric states for each γD-crys species. Results showed the monomeric structures of all three γD-crys species tested were similar, however; the aggregate conformations showed significant differences.

For instance, HX-MS measurements suggested wild type and S130T might form well-structured amyloid-like aggregates. This correlates with the faster aggregation kinetics observed for wild type and S130T compared to H22T and S130P. On the other hand, the HX-MS results suggested H22T and S130P formed more flexible, less-structured aggregates, which may correlate with their greater resistance to aggregate relative to wild type and S130T. The results for H22T and S130P also highlighted the promise of the mutational strategies to which they were affiliated as discussed in Chapter 2. Many potential aggregation contacts were identified experimentally for wild type and S130T, and relatively few for H22T and S130P. Nonetheless, a potential aggregation contact at N125-L133 was identified among all γD-crys species tested, although less so for S130P, suggesting this region may be important to the aggregation mechanism of γD-crys. Further, various aggregation calculators predicted residues

located within N125-L133 to be an aggregation-prone, “hot spot”, and also indicated a potential decrease in IAP for S130P.

Therefore, this work reports some promise for using HX-MS and these aggregation calculators as experimental and computational tools, respectively, to identify aggregation contacts. Notably, however, incorrect predictions were also observed, and the lack of statistical significance for several reporter peptides using HX-MS was apparent. Thus, additional experimentation is needed to test more sample replicates and decrease statistical uncertainties; and more reporter peptides are needed for analysis to achieve smaller peptide-level or single-residue resolution. Finally, further improvements and optimization of the computational algorithms would also be valuable to yield higher predictive success rates.

Chapter 4: Using RosettaDesign to investigate the aggregation mechanism of the ALS-associated variant A4V in human Cu, Zn superoxide dismutase

4.1. Introduction

Non-native protein aggregation can cause problems in many biotechnological manufacturing processes [Vazquez-Rey 2011, Wang 2005, Mahler 2009, Weiss 2009, Cromwell 2006, Chi 2003], induce undesired immunogenicity in patients, [Wang 2012, De Groot 2007], and is also commonly associated with various human diseases. For instance, proteins having undergone non-native aggregation (hereafter referred to as aggregation) are often observed to have increased levels of β-sheet, secondary structure [Weiss 2009], such as the cross-beta amyloid structure, commonly associated in many protein deposition diseases in humans [Rousseau 2006, Murphy, 2002, Fink 1998].

For example, the fibrillation of the beta-amyloid peptide (Aβ) within neurons is considered a critical aspect of Alzheimer's disease (AD) [Sanchez de Groot 2006]. Another example is the misfolding and misassembly of the extracellular, homotetrameric protein known as transthyretin (TTR), believed to result in familial amyloid polyneuropathy (FAP) [Du & Murphy 2010]. Additionally, the formation of insoluble protein aggregates, often caused by protein instability, is often related to Huntington's, Parkinson's and the spongiform encephalopathy diseases such as "mad cow" and Creutzfeld-Jacob disease [Murphy 2002, Valentine 2003, Nowak 2010].

Arguably one of the most devastating neurodegenerative diseases, however, is amyotrophic lateral sclerosis (ALS or Lou Gehrig's disease). ALS is characterized by the rapid degeneration of both the corticospinal (upper) and spinal (lower) motor neurons of those affected [Majoor-Krakauer 2003, Valentine 2003]. Although mental intellect is unaffected as the degeneration of motor neurons persists, denervation and muscle

atrophy result in weakness, eventual paralysis, and often death within five years of diagnosis [Strange 2003, Potter 2003]. ALS has a mid-to-late-life onset with an overall prevalence approximated at 4-6 per 100,000 in population that increases with age and is more frequent in men than women [Valentine 2003, Majoor-Krakauer 2003]. Most ALS cases occur sporadically with no known cause (~ 90%), while familial forms (fALS) are inherited from point mutations found in the gene encoding the cytosolic expression of human copper-zinc superoxide dismutase-1 (hSOD1) [Münch 2010, Strange 2003, Durazo 2009]. An enzyme, hSOD1 protects the body from oxidative damage caused by superoxide radicals by catalyzing the reaction of superoxide into hydrogen peroxide and oxygen [Valentine 2003]. Previous work has shown hSOD1 exists naturally as a 32-kDa homodimer with each subunit containing an eight-stranded anti-parallel β-barrel, individual zinc and copper binding sites, and a lone disulfide bond as depicted in Figure 4.1 [Hough 2004, Strange 2003].

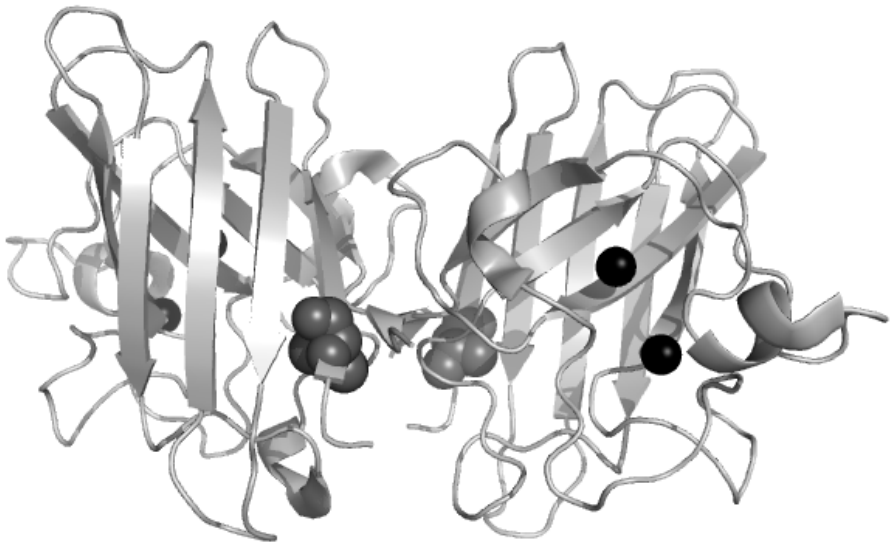


Figure 4.1: Crystal structure of human copper-zinc superoxide dismutase-1 (hSOD1) (pdb: 1N19) showing the homodimeric structure comprised predominantly of β-sheets. The copper and zinc metal binding sites are indicated on both subunits by black spheres, and the destabilizing A4V variant is shown on both subunits by gray spheres.

The mechanism(s) by which hSOD1 variants instigate fALS is still unclear, but studies have demonstrated toxicity is exerted through a gained toxic function rather than a loss of activity. Pioneering transgenic mice studies observed an absence of ALS-like symptoms in mice lacking hSOD1, providing evidence supporting this hypothesis [Reaume 1996, Shaw 2007]. Previous studies have suggested abnormal biochemical activity resulting in oxidative stress, increased levels of free radical species [Vucic 2009, Yim 1996, Bogdanov 1998, Liu 1998], or possibly improper zinc and copper metal binding [Vucic 2009, Lyons 1996, Crow 1997, Estevez 1999]. Yet another hypothesis explaining the toxicity of hSOD1 variants is the formation of neurotoxic protein aggregates resulting from misfolding or decreased conformational stability [Vucic 2009, Watanabe 2001, Rakhit 2006]. This aggregation hypothesis, in particular, is further supported by data showing increased aggregate formation for hSOD1 variants relative to wild type in transgenic mice models [Johnston 2000, Bruijn 1998, Bruijn 1997], cultured cells [Koide 1998], and *in vitro* studies [Stathopulos 2003]. Furthermore, transgenic mice expressing these hSOD1 variants have been observed to develop fALS-like symptoms [Gurney 1994, Wang 2002].

Proper copper and zinc binding and the oxidation of intramolecular disulfide bonds have both shown to increase the thermal stability of hSOD1 and favor dimer formation [Galaleldeen 2009, Doucette 2004, Lindberg 2004]. Thus, improper metal binding or reduction of intramolecular disulfide bonds may instigate dissociation of the hSOD1 homodimer into monomeric states prone to aggregate [Rakhit 2004, Niwa 2007, Banci 2007]. It has been well documented that many fALS-associated hSOD1 variants lead to destabilization of the homodimer and subsequently instigate aggregation [Hough 2004]. For instance, there are currently over 100 known ALS-associated variants that can occur in hSOD1, the majority being point mutations, however, sequence insertions,

deletions, and truncations can also take place [Chattopadhyay 2009]. However, one of the most prevalent hSOD1 variants, particularly in North America, is the alanine to valine substitution at the fourth residue (A4V-hSOD1), associated with nearly 50% of all fALS cases [Hough 2004]. Furthermore, A4V-hSOD1 is an extremely severe mutation as it induces a rapid progression of the disease resulting in death within 1-2 years from the onset of symptoms [Galaleldeen 2009, Hough 2004]. Past work involving A4V-hSOD1 variant has observed similar metal-binding and catalytic activity [Valentine 2003, Hayward 2002], but decreased thermal stability and intramolecular disulfide bond reduction relative to wild type hSOD1 [Rodriguez 2002, Tiwari 2003]. Additionally, local unfolding has been observed at the dimer interface of A4V-hSOD1 [Shaw 2006], and increases in protein unfolding and aggregation propensity of A4V-hSOD1 relative to wild type have also been shown [Stathopulos 2003].

As such, investigating the aggregation mechanism of prominent fALS-associated hSOD1 variants, such as A4V-hSOD1, would be valuable. Figure 4.1 shows the variant site is located in the first beta-strand that forms part of the dimer interface, and the amino acid side chain is shown to pack into the domain core of each subunit [Hough 2004]. Thus, a larger valine residue substituted at this position could conceivably cause an intra or inter-domain steric clash(s) with neighboring amino acids that may destabilize the protein, and instigate aggregation. In order to test this theory, neighboring residues of A4V that could engage in problematic interactions with the variant site would need to first be identified. This could potentially be accomplished by examining the crystal structure(s) of A4V-hSOD1. Then, by using computational design and protein engineering techniques, potential amino acid substitutions at these locations could be suggested that may prevent problematic interactions, and potentially improve the conformational stability and/or aggregation of A4V-hSOD1. These variants could then be

characterized experimentally to test the effectiveness of this strategy and the success rates of the computational design tools.

However, utilizing protein engineering to uncover and replace problematic molecular interactions caused by disease-related, protein variants can also present several challenges. For instance, in order to identify the neighboring residues that could engage in problematic interactions, the crystal structure of the protein must first be known. Further, if multiple crystal structures are available, they may differ structurally and suggest conflicting problematic interactions. Another challenge is determining whether the residues involved in the clash also contribute to aggregation, or if they expose other residues that become aggregation contacts. Finally, a third challenge becomes identifying promising amino acid substitutions that will successfully prevent problematic interactions and suppress conformational destabilization and/or aggregation. Nonetheless, implementing computational design in tandem with protein engineering to identify candidate variants can potentially address this latter challenge.

Thus, computational design tools were utilized in this work to test whether the aggregation mechanism of A4V-hSOD1 is potentially caused by a steric clash involving the inserted valine at the fourth residue, and if so, suggest additional variants that may prevent these steric clashes. In the past, investigators trying to improve conformational stability and reduce protein aggregation have used several techniques, besides computational design, to help choose candidate mutations. These have included rational or structure-based methods [Eijsink 2004, Lehmann 2001, Wetzel 1994, Chrunyk 1993], directed evolution techniques [Lehmann 2001, Eijsink 2005], as well as consensus design using homologous proteins [Lehmann 2001, Forrer 2004]. However, directed evolution involves a substantial amount of time and experimental resources, while consensus design requires identification of many homologous proteins as well as criteria

for selecting a specific variant when a clearly favorable option at a specific location is lacking [Bannen 2008].

On the other hand, there are many examples in literature where rational or structural-based mutational designs resulted in enhanced protein stability [Eijsink 2004, Fernandez-Lafuente 2009, Goihberg 2008, Berezovsky 2005, Gerk 2000, Strickler 2006, Schwehm 2003, Williams 1999, Makhatadze 2003, Melnik 2012]. Further, it has also been shown that some of these design strategies can stabilize the protein against irreversible processes such as aggregation [Logan 2010, Ray 2004, Sekijima 2006]. Nevertheless, at least two significant challenges to rational design remain: (1) the tertiary structure of the protein must be known, and (2) the lack of a reliable process for identifying the best variant(s) out of vast possibilities to best improve the desired property of the protein [Eijsink 2004, Bannen 2008].

However, as mentioned, implementing computational design methods [Das 2008, Cellmer 2007, Bratko 2007, Saven 2010] in conjunction with rational design methods may address the latter of these challenges. Furthermore, there are numerous investigations and reviews reported in literature where successes in computational protein re-engineering or *de novo* design have been established to improve conformational stability [Das 2008, Bratko 2007, Saven 2010, Dantas 2003, Shah 2007, Hu 2008, Schueler-Furman 2005, Tian 2010, Lu 2009]. However, to our knowledge the application of such computational algorithms to identify and replace amino acids involved in the aggregation mechanism of proteins has not been widely studied.

Thus in this work, we implemented computational design tools to test if they could successfully identify problematic molecular interactions in A4V-hSOD1 that may result in conformational destabilization and aggregation of the protein. These problematic interactions could 1) involve steric clashes between A4V and other residues located on the same domain, 2) involve steric clashes between A4V and other residues

located on the adjacent domain, or finally 3) involve structural changes that result in the exposure of cysteine residues and cause intermolecular disulfide bonding. In order to test these possibilities, the computational design program, RosettaDesign, was used to suggest second-site variants near the A4V variant site that were identified to prevent problematic interactions and improve the conformational stability of the protein. The location and type of candidate mutations that were identified were then analyzed to determine whether any of the aforementioned hypotheses could contribute to the aggregation mechanism of A4V-hSOD1. Further, primary-sequence-based aggregation calculators were also implemented to predict whether these second-site variants might also alter the aggregation propensity of the protein. Some of these second-site variants were expressed and characterized experimentally, and the experimental results were compared to the computational predictions to evaluate predictive success rates of the computational design tools.

4.2. Materials and Methods

Selection of second-site variants in A4V-hSOD1 using RosettaDesign

The fixed backbone design protocol in RosettaDesign 3.0 was utilized for identifying second-site variants of A4V-hSOD1. This protocol permitted the side chains of every amino acid, except those involved in metal binding, to move and repack, including inserted variants. Design runs were conducted for four different crystal structures (pdb files: 1N19, 1UXM, 1PU0, and 2C9V). The crystal structures in the pdb files 1N19 [Cardoso 2002] and 1UXM [Hough 2004] contained a main backbone based on the A4V-hSOD1 structure. However, the crystal structures in the pdb files 1PU0 [DiDonato 2003] and 2C9V [Strange 2006] contained a main backbone based on the wild type hSOD1 structure. Each crystal structure was modified to include the A4V variant on both subunits, generating an A4V-homodimer, to be used as the initial starting

structure. First RosettaDesign was used to "repack" the starting A4V homodimeric structures by allowing all side chains, except those chelated to metal ions, to adopt favorable conformations, and the corresponding energy scores were recorded and used as reference scores. The residues involved in metal chelation, and thus not allowed to "repack" or mutate were H46, H48, H63, H71, H80, D83, and H120.

RosettaDesign was then used to conduct global redesign runs. During these redesigns, those amino acids located less than 5 Å from the A4V variant site on each subunit, and not involved in metal ion chelation, were allowed to mutate to any other amino acid. Three replicate global design runs were performed for each starting structure, and each produced a variable protein design generated from the Monte Carlo optimization used by RosettaDesign. Variants that were identified from these global redesign runs using each starting crystal structure, were then pooled. Subsequently, each variant within the pool was analyzed one at a time by individually inserting the variant into the A4V-hSOD1 sequence, and the corresponding energy score ($\Delta\Delta G_f$ using Eq. 1.1) estimated by RosettaDesign was recorded. Hereafter, these variants are referred to as *second-site variant*s because when inserted into the A4V-hSOD1 sequence, an hSOD1 double variant was generated compared to wild type hSOD1. All energy scores were then compared to the energy score of the respective A4V-homodimeric starting structure. Those second-site variants that RosettaDesign estimated to have improved $\Delta\Delta G_f$ values were identified as candidate second-site variants that may prevent a steric clash with the A4V variant site and conformationally stabilize A4V-hSOD1.

Furthermore, because RosettaDesign utilizes Monte Carlo optimization some residues were redesigned to different amino acids within the three global design runs. For example, the native F20 residue was redesigned to F20G in one global design run, but to F20L in another run. Therefore, in some cases complete mutation scans were also

conducted for certain residues to estimate the energy scores of all possible variants at a given location.

Candidate variants were required to pass several filtering metrics before selection. For instance, only second-site variants identified by RosettaDesign to lower the change in total energy score, $\Delta\Delta G_f$, relative to the A4V-hSOD1 homodimeric starting structure, were considered (similar to Eq. 2.1). In addition, a breakdown of the hydrogen-bonding network (ΔE_{HB}) or Lennard-Jones (ΔE_{LJ}) contributions within the overall energy score function resulted in a given second-site variant being disqualified from consideration.

The command lines and resfiles used to perform these RosettaDesign calculations are shown in Appendix B. Finally, the aforementioned aggregation calculators and the 3D profiling method were also used to predict whether any of these second-site variants would alter the intrinsic aggregation propensity (IAP) of the A4V-hSOD1 homodimer as well. The 3D profiling method was used to estimate quantitative IAP changes since the other three aggregation calculators can only provide qualitative predictions. These computational predictions were then also compared to the experimental results.

Expression and purification of hSOD1 variants and wild type

hSOD1 variants and wild type were expressed in *Saccharomyces cerevisiae* SOD-EG118 strain lacking the endogeneous yeast *sod1* gene EG118 was graciously provided by the Valentine lab at UCLA, and the construction of expression vectors were then made by Simpson Gregoire using the QuikChange II site-directed mutagenesis kit (Stratagene). Expression and purification was based on the procedure of Hough et. *al*. [Hough 2004]. Cells were initially grown on leucine dropout media containing 2% w/v agar at 30 °C for 48 hours. Single yeast cell colonies were selected and grown as

primary cultures in 50-100 mL leucine dropout media in 250 mL baffled flasks at 30 °C for 48 hours at approximately 220 rpm. Fresh 2 L batches of yeast-extract-peptone-dextrose (YPD) growth media were then inoculated with the primary culture to produce secondary cultures with an optical density of 0.05 at 600 nm (OD_{600} = 0.05). Secondary cultures were grown further at 30 °C for 4-5 days at 220 rpm.

Yeast cells were then harvested by centrifuging at 3500 rpm for 35 minutes. The supernatant was decanted and the cell pellet was frozen at -80 °C for storage and to aid cell lysis. The frozen cell pellet was resuspended in lysis buffer (250 mM Tris-HCl, 150 mM NaCl, 0.1 mM EDTA, pH 8.0). Following resuspension, cells were lysed using a bead beater (BioSpec Products) with 0.5 mm diameter glass beads in a 15 mL aluminum chamber immersed in ice. Cell lysate was centrifuged at 11,000 x g for 45 minutes at 4 °C and the resulting supernatant was collected and gradually brought to 60% ammonium sulfate saturation while on ice to induce precipitation of impurities. The solution was then incubated on ice for an additional 30 minutes and subsequently centrifuged to discard the precipitate. The supernatant was decanted and diluted to 2.0 M ammonium sulfate before loading onto a pre-packed HiPrep Phenyl FF (high sub) 16/10 column (GE Healthcare Biosciences) pre-equilibrated with 50 mM sodium phosphate, 150 mM NaCl, and 0.3 mM ethylenediaminetetraacetic acid (EDTA), pH 7.0 containing 2.0 M ammonium sulfate buffer, denoted buffer A. The column was washed with 1 CV buffer A before a step gradient was initiated decreasing the ammonium sulfate concentration from 2.0-0.1 M. Fractions containing hSOD1, as determined by SDS-PAGE, were collected and pooled followed by dialysis against ddH_2O at 4 °C. The pooled samples were then loaded onto a HiTrap DEAE FF column (GE Healthcare Biosciences) and proteins were eluted with a linear gradient of potassium phosphate (2.25 mM-1M, pH 7.0). The protein solution was then exchanged to a 20 mM potassium phosphate, 80 mM NaCl buffer, pH 7.4, using a 10 kDa cutoff centrifuge filter (Millipore), and then subjected

to size exclusion chromatography on a commercial Superdex 75 10/30 GL column (GE Healthcare Biosciences). Fractions containing purified hSOD1 were combined and the final concentration of protein was determined via UV spectroscopy at 280 nm. Finally, aliquots of 10 mg/ml protein were stored at -80 °C for future experiments.

Equilibrium chemical unfolding using circular dichroism (CD)

Shaojie Zhang conducted all hSOD1 chemical unfolding experiments in the Department of Chemical Engineering at the University of Virginia.

Chemical denaturation experiments of hSOD1 variants and wild type were performed using high purity guanidine hydrochloride (GdnHCl) (MP Biomedical) as the denaturant. Frozen aliquots of hSOD1 were first thawed and dialyzed to phosphate-buffered saline (PBS) (10 mM sodium phosphate, 2 mM potassium phosphate, 137 mM NaCl, pH 7.4). Next, hSOD1 samples were prepared to a dimer concentration of 15 μM into solutions ranging from 0 to 6.8 M GdnHCl. Samples were incubated for approximately 16 hours at room temperature at 100 rpm to ensure proper mixing. Circular dichroism (CD) was used to monitor the extent of unfolding as a function of increasing denaturant. CD ellipticities at 218 nm were recorded for each protein sample on a JASCO J-710 spectrapolarimeter, under nitrogen purge, in kinetics mode using a 0.1 cm circular, quartz cuvette (Helma) at room temperature. Three measurements were recorded for 30 seconds and then averaged for each denaturant concentration tested. Final ellipticity values were estimated after deducting the contribution from the buffer solution. These ellipticity values were then converted to the fraction of unfolded protein at each denaturant concentration tested, denoted f_u, using Eq. 4.1

$$f_u = \frac{\theta_n - \theta}{\theta_n - \theta_d} \qquad \text{(Eq. 4.1)}$$

where θ_n, θ_d, and θ represent the ellipticity value observed for fully folded protein, fully denatured protein, and the ellipticity value observed for the protein at a given denaturant concentration being tested.

Isothermal aggregation and analysis by size-exclusion chromatography (SEC)

Shaojie Zhang conducted all hSOD1 isothermal aggregation experiments in the Department of Chemical Engineering at the University of Virginia.

The aggregation of hSOD1 variants and wild type was investigated via isothermal aggregation experiments. Frozen aliquots of hSOD1 were first thawed and then buffer exchanged to a tris-buffered saline (TBS) (20 mM Tris, 150 mM NaCl, pH 7.4). Protein samples were prepared to an initial hSOD1 dimer concentration of 120 μM, followed by incubation in upright, sealed glass vials at 37 °C in a circulating water bath with negligible temperature variability. After 10 minutes of incubation, 3mM EDTA was added to induce metal dissociation and aggregate formation as was done in previous work [Ray 2005]. Samples from the incubated protein solutions were periodically taken (~ 25 μL), and analyzed using size exclusion chromatography (SEC) using TBS as the mobile phase, operated at a flow rate of 1 mL/min. Prior to injection, protein samples were diluted with 10 μL of TBS and 25 μL was then injected into a Tosoh SEC-2000 column (Tosoh Bioscience) for wild type hSOD1, and a Superdex G75 column (GE Healthcare) for hSOD1 variants. The columns were connected to a Waters Alliance e2695 separation module (Waters Corporation) and a SpectraSystem UV1000 (ThermoSeparation Products) for separation and detection via UV at 280 nm, respectively.

4.3. Results

Estimating the energy scores for candidate, second-site variants

RosettaDesign was used to estimate energy scores for candidate second-site variants of A4V-hSOD1 that potentially could prevent a steric clash with the A4V variant site and restabilize the conformation of the protein. Within each crystal structure, 18 amino acids were found within a 5-Å diameter around the A4V variant site on both subunits. During global redesign runs, the substitution of these 18 amino acids to any other amino acid was permitted, allowing for 342 possible variants. Results from each crystal structure were combined and showed RosettaDesign redesigned 13 of these 18 amino acids. The 5 amino acids that were not redesigned were I18, V29, L106, I149, and G150, and evaluation of the A4V-hSOD1 tertiary structure shows these are all hydrophobic residues packed into the core of each subunit (not shown). This may indicate these 5 amino acids are critical to the conformational stability of hSOD1, and avoid problematic interactions with the A4V variant site, rendering RosettaDesign reluctant to replace them. In fact, previous work has reported the important structural role of L106 as it has been conserved in the sequence of hSOD1 from different species [Getzoff 1989].

On the other hand, RosettaDesign identified 39 variants from the 13 residues that were redesigned during the global design runs of each starting crystal structure. Several variants occurred at the same sequence location (e.g. F20G or F20L) during different design runs depending on the starting structure used because of the Monte Carlo optimization implemented by RosettaDesign. Next, each of the 39 variants was individually inserted into the four starting crystal structures, and the corresponding energy score estimated by RosettaDesign for each was recorded. The aforementioned scoring metrics were then applied to these energy scores to identify candidate second-site variants. Table 4.1 lists the change in the total energy score, $\Delta\Delta G_f$, relative to the

four A4V-hSOD1 homodimer starting structures for each candidate, second-site variant. Bolded energy scores indicated the corresponding second-site variant passed all scoring metrics for that particular starting crystal structure.

Table 4.1: Summary of the $\Delta\Delta G_f$ values, relative to each A4V-hSOD1 homodimeric starting crystal structure, estimated by RosettaDesign for second-site variants identified during the global redesign runs. $\Delta\Delta G_f$ values were estimated when the variants were individually inserted into each starting crystal structure. Bolded energy scores passed all scoring metrics for that given starting crystal structure.

Second-site variant	The change in total energy score: $\Delta\Delta G_f$ (kcal/mol)			
	1PU0	2C9V	1N19	1UXM
A1G	0.3	**-7.4**	0.8	0.4
T2G	2.3	**-97.6**	4.3	2.5
T2H	0.0	**-25.2**	0.6	0.1
T2L	**-0.7**	200.2	3.0	1.2
T2N	0.0	**-15.3**	0.5	1.1
T2R	**-0.9**	11.4	1.4	0.6
T2V	0.3	95.2	1.4	-0.1
K3E	0.7	0.9	0.8	0.3
K3R	**-1.0**	**-1.6**	**-1.2**	**-0.8**
K3T	-0.5	-0.6	0.1	**-0.8**
K3V	-0.8	0.3	-0.5	0.8
V5S	-0.5	-0.2	-0.1	-0.3
C6A	0.5	0.6	-0.1	0.1
C6V	0.6	0.6	1.4	-0.2
N19F	**-3.5**	**-2.9**	-3.4	-2.7
N19W	**-3.5**	-2.6	**-2.9**	-2.7
N19Y	**-3.5**	-2.9	-3.3	**-3.0**
F20G	**-37.9**	**-42.8**[a]	8.0	2.9
F20L	-14.9	**-13.9**	4.2	0.6
E21K	-0.2	0.3	-0.2	0.2
E21M	0.1	0.1	0.3	-0.1
E21R	0.0	0.4	0.1	0.2
E21V	-0.2	2.0	1.8	1.3
E21Y	1.7	11.7	-1.2	0.4
I112V	1.3	1.5	1.3	1.4
I113Q	1.2	1.0	0.9	1.1
I113R	**-0.6**	**-0.1**	**-0.3**	**-0.7**
I113V	0.2	0.2	-0.2	-0.1
I151D	0.9	0.8	0.2	1.3
I151V	0.2	0.2	-0.2	0.1
I151Y	-0.2	**-0.1**	2.5	0.4
A152C	-1.0	**-0.8**	-1.1	**-0.9**
A152H	75.6	**-1.6**	0.2	**-0.9**
A152S	**-1.0**	**-0.7**	**-0.1**	-0.8
Q153A	0.1	0.2	-0.1	0.4
Q153C	0.1	**-0.1**	-0.1	0.1
Q153D	-0.3	0.1	0.0	-0.1
Q153S	-0.1	0.1	-0.2	0.3

[a]The F20G variant only marginally failed the filtering metrics for 2C9V as the E_{HB} = 0.1 kcal/mol.

As seen in Table 4.1, more candidate second-site variants were shown to pass the scoring metrics when inserted into the crystal structures of 1PU0 and 2C9V (based on the wild type hSOD1 structure), compared to the crystal structures of 1N19 and 1UXM (based on the A4V-hSOD1 structure). Furthermore, the absolute magnitude of many $\Delta\Delta G_f$ values was larger for 1PU0 and 2C9V compared to 1N19 and 1UXM. These large values are caused by large fluctuations in certain terms within the energy function of RosettaDesign as a result of more steric clashes appearing or disappearing in wild type-based crystal structures compared to A4V-based crystal structures.

Table 4.1 also showed several second-site variants that passed each filtering metric for one starting structure, but not for the other three. This was the case for the variants A1G, several T2 variants, K3T, I151Y, and Q153C. Previous work has reported residues I151 and A152 are involved in inter-subunit, hydrogen bonding, suggesting structural importance [Hough 2004]. Additionally, A1 and Q153 are located at the terminals of the protein sequence, and do not directly interact with the A4V variant site, therefore they may not be essential to the protein stability or promote steric clashes with A4V. Further, the insertion of a cysteine residue with Q153C may promote incorrect intra- or intermolecular disulfide shuffling. Thus, as variants that would improve conformational stability and/or prevent a steric clash with A4V these variants were disqualified as candidates.

Additionally, significantly larger improvements in $\Delta\Delta G_f$ values were also observed for certain residues, indicating a potential higher degree of conformational stabilization, compared to others. For instance, this was observed for variants located at the T2 (e.g. T2G, T2H, and T2N) and at the F20 (e.g. F20G and F20L) positions. However, previous work has reported the native T2 residue is important to the hydrogen-bonding network [Hough 2004], and only showed favorable energy scores in one crystal structure (2C9V).

Therefore, variants located at the T2 position were disregarded as candidate variants to test in this initial study.

On the other hand, the variants observed at the F20 position showed promise, both rationally and computationally, specifically when inserted into the wild type-based starting structures (e.g. 1PU0 and 2C9V). As a result, a complete mutation scan was also conducted via RosettaDesign, and the energy scores for the other possible F20 variants were estimated to see if similar results were observed. Upon visual examination of F20G and F20L within these crystal structures, the removal of the phenylalanine (Phe) prevented a steric clash with the valine (Val) side chain at the A4V variant site (Figure 4.2). The steric clash occurred between a hydrogen atom on the beta carbon of the native Phe residue and a hydrogen atom on a branched carbon of the Val residue as seen in Figure 4.2. The prevention of the steric clash at this location significantly improved the attractive and repulsive components of the Lennard Jones potential (E_{LJ}) within the Rosetta energy function (Eq. 1.1). Further, the change in the E_{LJ} term was the dominant contributor to the change in the total energy score as shown in Table 4.2. Values for $\Delta\Delta G_f$ and ΔE_{LJ} for all possible F20 variants, estimated by RosettaDesign relative to A4V-hSOD1 homodimer, are shown in Table 4.2.

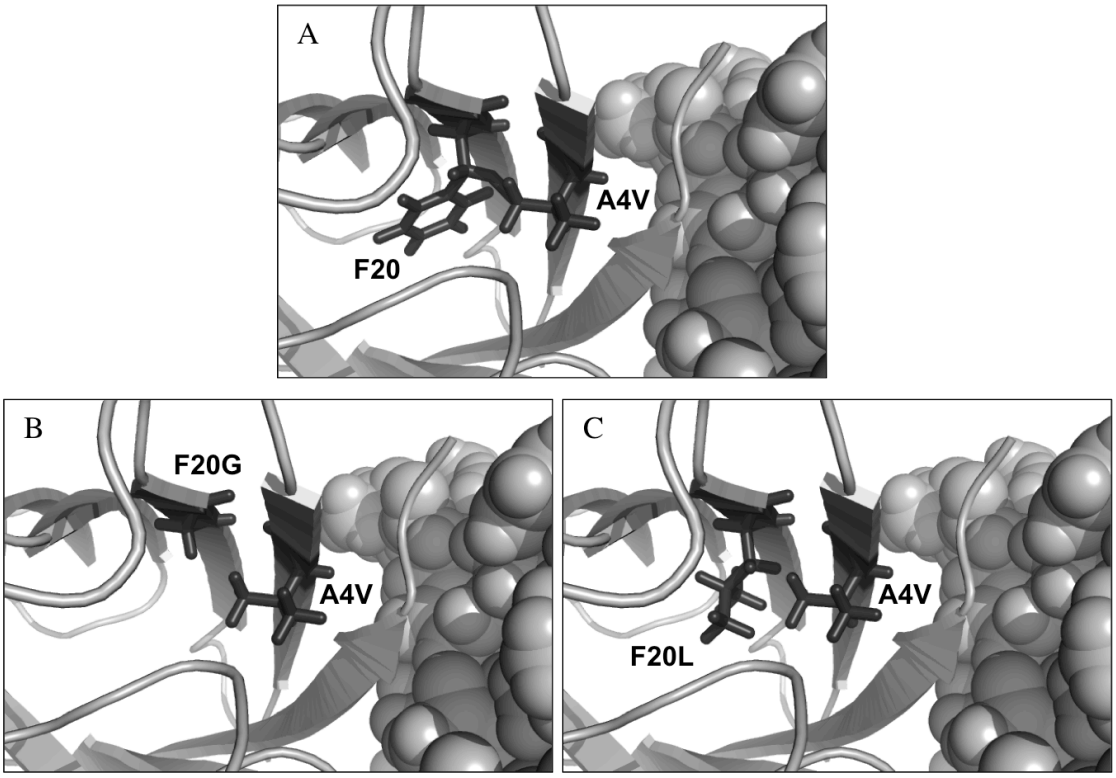


Figure 4.2: Molecular images from *PyMOL* showing A) the steric clash between residues F20 and the variant A4V, B) the lack of a steric clash between the variants F20G and A4V, and C) the lack of a steric clash between variants F20L and A4V. The side chains at residues 4 and 20 are shown as sticks in dark grey, and include hydrogens. The filled spheres denote the adjacent hSOD1 subunit. Variants were inserted into the structure and all side chains were repacked using RosettaDesign. The corresponding energy scores for each molecular configuration are shown in Table 4.1.

Table 4.2: $\Delta\Delta G_f$ and ΔE_{LJ} values estimated by RosettaDesign for each F20 variant relative to the A4V-hSOD1 homodimeric starting structure. Values for E_{LJ} were estimated by adding the attractive and repulsive contributions of the Lennard Jones potential found within the RosettaDesign energy function. Here negative values represent favorable changes in the energy score.

F20	1PU0		2C9V		1N19		1UXM	
variant	$\Delta\Delta G_f$	ΔE_{LJ}	$\Delta\Delta G_f$	ΔE_{LJ}	$\Delta\Delta G_f$	ΔE_{LJ}	$\Delta\Delta G_f$	ΔE_{LJ}
F20A	5.0	5.5	3.4	3.9	5.3	6.0	3.2	4.8
F20C	9.9	10.1	11.6	11.6	4.9	5.9	2.7	4.2
F20D	3.5	-3.5	2.6	-3.5	7.5	3.5	5.0	0.3
F20E	-2.3	-7.1	-2.5	-7.2	5.9	2.2	3.5	0.0
F20G	-37.9	-39.9	-42.8	-44.8	8.0	6.9	2.9	1.7
F20H	1.8	-1.5	1.2	-1.9	3.8	0.9	2.7	0.1
F20I	11.6	11.6	9.4	9.8	6.9	7.5	9.3	11.4
F20K	-13.2	-19.0	-11.8	-17.2	5.4	2.4	3.8	0.6
F20L	-14.9	-15.2	-13.9	-13.6	4.2	4.8	0.6	2.1
F20M	-8.7	-10.2	-8.4	-10.2	3.0	3.6	0.4	0.1
F20N	-3.4	-9.6	1.4	-3.2	6.6	3.5	4.3	1.7
F20P	94.3	78.0	39.4	23.6	23.2	7.8	21.4	6.3
F20Q	-3.8	-8.7	-1.1	-6.2	5.7	1.7	2.9	0.2
F20R	-9.5	-16.2	-9.6	-16.3	7.5	1.9	3.3	-1.1
F20S	1.6	0.7	6.0	6.3	5.6	5.5	2.5	3.3
F20T	10.6	10.4	0.1	-0.8	9.2	9.3	10.3	11.2
F20V	10.5	12.4	6.9	8.9	7.1	9.1	8.4	12.0
F20W	12.0	7.8	5.6	1.3	19.5	13.4	12.0	8.4
F20Y	-2.4	-3.4	-2.3	-3.5	3.1	2.0	0.7	-0.1

As seen in Table 4.2, RosettaDesign showed multiple F20 variants exhibited favorable changes in the total energy score, $\Delta\Delta G_f$, relative to the A4V-hSOD1 homodimeric structure. However, favorable changes were only observed for wild type-based crystal structures (1PU0 and 2C9V). F20G was found to be the most favorable variant in both 1PU0 and 2C9V, but replacing the phenylalanine with other hydrophobic amino acids, such as leucine (F20L) and methionine (F20M), also prevented the steric clash and produced favorable changes in the total energy score as well. As such, these variants were also considered for experimental testing.

Somewhat surprisingly, inserting other hydrophobic amino acids intermediate in size between Phe and Gly at residue 20, such as alanine, valine, isoleucine, and

tryptophan did not prevent the steric clash between the hydrogen on the beta carbon of residue 20 and the hydrogen on the branched carbon of residue 4 from occurring. According to RosettaDesign, these variants had the same orientation of hydrogen atoms at the beta carbon compared to the native Phe residue, resulting in the steric clash to persist with the valine at residue 4. This resulted in some amino acids with smaller side chains, such as alanine, failing to prevent the steric clash, but other amino acids with larger side chains, such as methionine, to successfully prevent the clash. Notably however, F20A was rationally selected for experimentation, despite unfavorable scores from RosettaDesign, because it possessed the second smallest side chain of all the amino acids and may provide more freedom to accommodate the valine at residue 4.

Further, substituting uncharged polar amino acids for the native Phe at residue 20, such as asparagine (F20N), glutamine (F20Q), or tyrosine (F20Y) as well as charged amino acids such as glutamic acid (F20E), lysine (F20K), and arginine (F20R) also produced favorable total energy scores. However, the solvation term in the energy function of RosettaDesign was not improved for these variants, and no significant change was observed for the hydrogen bonding term as well. This is most likely because the insertion of polar or charged amino acids into the hydrophobic core of a protein is often undesired, therefore these F20 variants were not considered as candidates to test experimentally.

Notably, other non-F20 variants were observed to have improved $\Delta\Delta G_f$ values ($\Delta\Delta G_f < 0.0$) when inserted into all four starting structures. These included the variants K3R, V5S, three N19 variants, I113R, A152C, and A152S (Table 4.1). Notably, upon examination of the tertiary structure, the variants at residue 19 were observed to be extensively solvent exposed and not interactive with the A4V side chain, thus they were assumed to be less essential to the conformational stability of A4V-hSOD1. Further, the variant A152C was also not considered because an additional cysteine residue inserted

into the sequence could increase the propensity for incorrect intra- or intermolecular disulfide formation. As a result, these variants, besides those at residue 19 and A152C, were also considered as candidate second-site variants, despite not passing all defined filtering metrics.

Overall, RosettaDesign identified several variants that were located near the A4V variant site and were suggested to increase the conformational stability of A4V-hSOD1, potentially via prevention of a steric clash with the A4V variant. A final group of candidate, second-site variants was selected, and included K3R, V5S, F20A, F20G, F20L, F20M, I113R, and A152S. The location of the native residues involving these variants, with respect to the A4V variant site are shown in Figure 4.3. As seen in Tables 4.1 and 4.2, significantly more favorable changes in the total energy score were observed for the F20 variants compared to the others. Consequently, in this work, initial experimentation was only conducted on F20 variants.

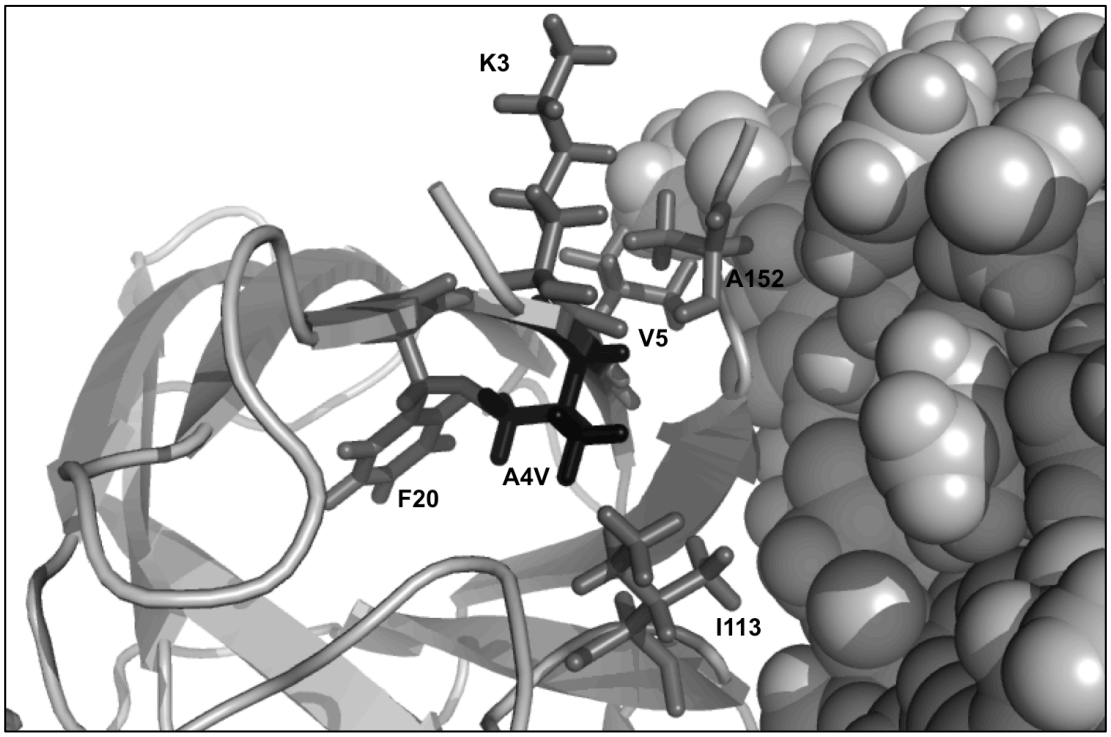


Figure 4.3: A molecular image from *PyMOL* illustrating the location of each candidate, second-site variant identified by RosettaDesign with respect to the A4V variant site. The second-site variants are labeled and denoted by dark gray sticks including hydrogen atoms. The A4V variant site is labeled and shown by black sticks, also including hydrogen atoms. The filled spheres denote the adjacent hSOD1 subunit.

Molecular analysis of non-F20 variants using RosettaDesign and PyMOL

For this preliminary investigation, only the F20 variants were tested experimentally. Nonetheless, a more in-depth examination of the non-F20 variants was also conducted to determine the intermolecular interactions that may result in their predicted conformational stabilization. These intermolecular interactions were based on changes to the individual energy terms, located within the RosettaDesign energy function (Eq. 1.1), after the insertion of each variant and relative to the A4V-hSOD1 homodimeric starting structure. For the most part, the changes in these individual energy terms were similar on a qualitative basis for each crystal structure tested. Notably, however, some differences were still observed.

For instance, K3R as a second-site variant showed favorable changes in both the Lennard-Jones potential term, E_{LJ}, and the hydrogen bonding potential term, E_{HB}, of approximately -4.0 kcal/mol total. However, a smaller, unfavorable change was observed in the solvation term, E_{sol}, of approximately 2.0 kcal/mol, thus yielding an overall, favorable $\Delta\Delta G_f$ value of approximately -2.0 kcal/mol (based on the crystal structure 2C9V) Notably, the insertion of a charged, arginine side-chain favors increased hydrogen-bonding, as indicated by RosettaDesign and visualized on a molecular level in *PyMOL*. Additionally, the multiple chi bonds of arginine, plus the proximity of the residue near the dimer interface, may allow or promote hydrogen bonding across the interface, and potentially reduce conformational destabilization.

On the other hand, V5S as a second-site variant showed unfavorable changes for both E_{LJ} and E_{HB} energy terms, totaling approximately 2.0 kcal/mol. Thus, V5S did not pass all of the filtering metrics, however, because the variant was identified overall to be conformationally stabilizing for all the starting crystal structures tested, it was considered a candidate. Favorable changes within the RosettaDesign energy function were observed for the torsion term, E_{tors}, and E_{sol}, of approximately -2.5 kcal/mol, yielding an overall, favorable $\Delta\Delta G_f$ value of approximately -0.5 kcal/mol (based on the crystal structure 1PU0). Upon examination of the tertiary structure, V5S is partially solvent exposed; therefore substituting a hydrophobic residue, like valine, to a polar residue, like serine, may be favorable and avoid undesired, hydrophobic intermolecular interactions on the protein surface.

As a second-site variant I113R showed favorable changes for both E_{LJ} and E_{HB} energy terms, of approximately -4.5 kcal/mol total. On the other hand, an unfavorable change of approximately 3.5 kcal/mol was observed for the E_{sol} energy term, thus yielding an overall favorable $\Delta\Delta G_f$ value of approximately -1.0 kcal/mol (based on crystal structure 1UXM). Upon examination of the I113R variant within the tertiary structure of

hSOD1, the variant was shown to be relatively buried and near the dimer interface. Thus, the insertion of a charged amino acid like arginine into a buried location would most likely create unfavorable solvation interactions; however increased hydrogen bonding may be realistic.

Finally, as a second site variant A152S also showed favorable changes for both E_{LJ} and E_{HB} energy terms, totaling approximately -2.0 kcal/mol. On the other hand, an unfavorable change was also observed for the E_{sol} term of approximately 1.0 kcal/mol, yielding an overall, favorable $\Delta\Delta G_f$ value of approximately -1.0 kcal/mol (based on crystal structure 2C9V). Upon examination of the tertiary structure, A152 was also shown to be partially solvent exposed. Thus, the replacement of a hydrophobic residue, such as alanine, with a polar residue, such as serine, would most likely yield favorable solvation interactions. Additionally, the hydrogen-bonding network could also be improved with the addition of a serine in place of an alanine, as indicated by RosettaDesign.

Identification of aggregation "hot spots" for wild type hSOD1 and variants

The three aggregation calculators PASTA, AGGRESCAN, and TANGO were also used to identify aggregation-prone, "hot spots", located within the wild type hSOD1 primary sequence. Additionally, calculations were also conducted to estimate if these "hot spots" were altered when the A4V variant or the candidate, second-site variants were inserted into the primary sequence.

Figure 4.4 shows the predictions from the aggregation calculators for wild type hSOD1 and after the A4V variant was inserted into the primary sequence. Figure 4.4.A shows five regions were agreed upon by 2 of the 3 calculators, and thus identified as potential "hot spots" within the wild type hSOD1 sequence. These five regions included residues A4-L8, Q15-F20, E100-L106, I113-H120, and A145-A152. After inserting the

A4V variant into the primary sequence, all three calculators agreed upon a “hot spot” between residues A4-V8, suggesting an increase in the intrinsic aggregation propensity (IAP) of the molecule. None of the other identified “hot spots” were altered after inserting the A4V variant since these calculators are primary-sequence-based.

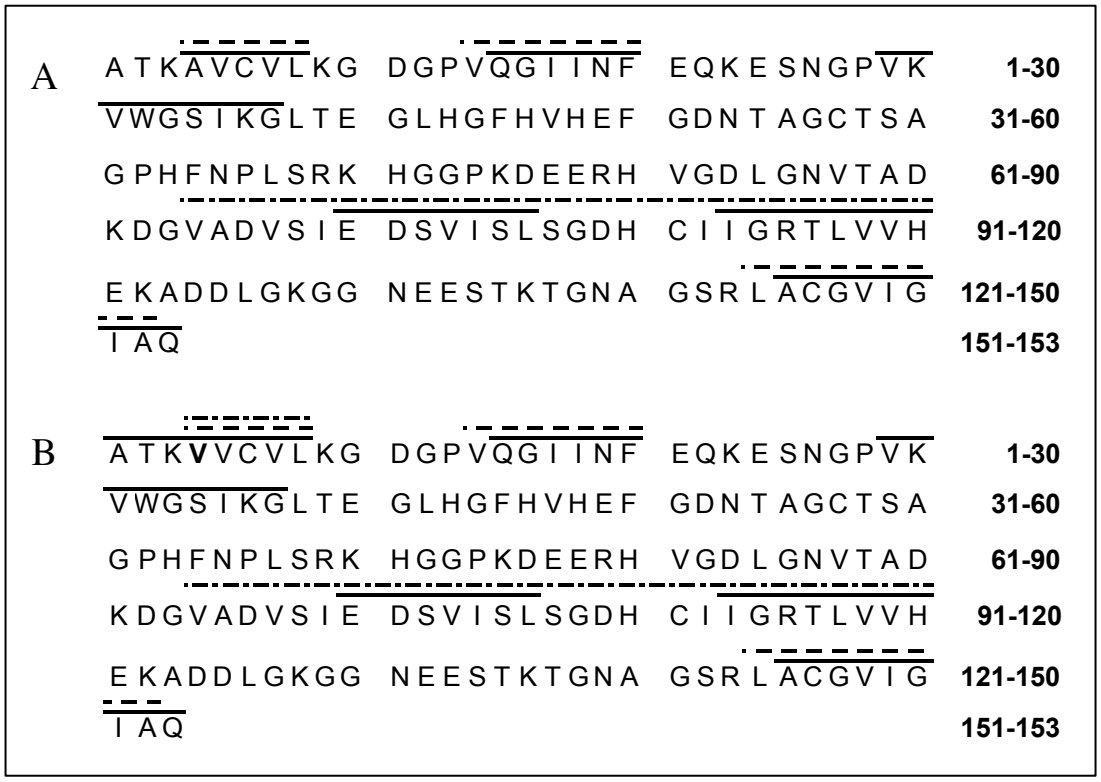


Figure 4.4: Potential aggregation-prone, “hot spots”, in the primary sequence of A) wild type hSOD1 and B) A4V-hSOD1 predicted by three aggregation calculators. Predicted “hot “spots” are denoted by lines above the sequence for AGGRESCAN (solid line), PASTA (dash-dotted line), and TANGO (dashed line), and the A4V variant site is bolded.

The aggregation calculators were also used to estimate whether any of the identified “hot spot” regions were altered after the addition of a candidate, second-site variant. Interestingly, four of the five “hot spot” regions identified in the wild type hSOD1 and A4V-hSOD1 sequence (Figure 4.4) contained at least one candidate, second-site variant. In fact, the aggregation calculators predicted favorable IAP changes, relative to the A4V-hSOD1 sequence, when the second-site variants V5S, F20A/F20G, and I113R were inserted into the primary sequence, as seen in Figure 4.5.

For instance, inserting V5S as a second-site variant into A4V-hSOD1 resulted in only two of calculators identifying a “hot spot” between residues V4-L8 (Figure 4.5A), relative to all three calculators as observed when the A4V variant was inserted alone (Figure 4.4B). Furthermore, inserting F20A and F20G as second-site variants was predicted to eliminate the consensus “hot spot” between residues Q15-F20 (Figure 4.5B). Likewise, inserting I113R as a second-site variant eliminated the consensus “hot spot” between residues I113-H120 (Figure 4.5C). On the other hand, inserting F20M as a second-site variant only decreased the corresponding, consensus “hot spot” at Q15-F20, on the A4V-hSOD1 primary sequence, to include Q15-I18; however the variant failed to eliminate the “hot spot” (data not shown). Finally, the aggregation calculators did not predict any alteration to “hot spots” after K3R, F20L, and A152S were inserted as second-site variants into the A4V-hSOD1 sequence (data not shown).

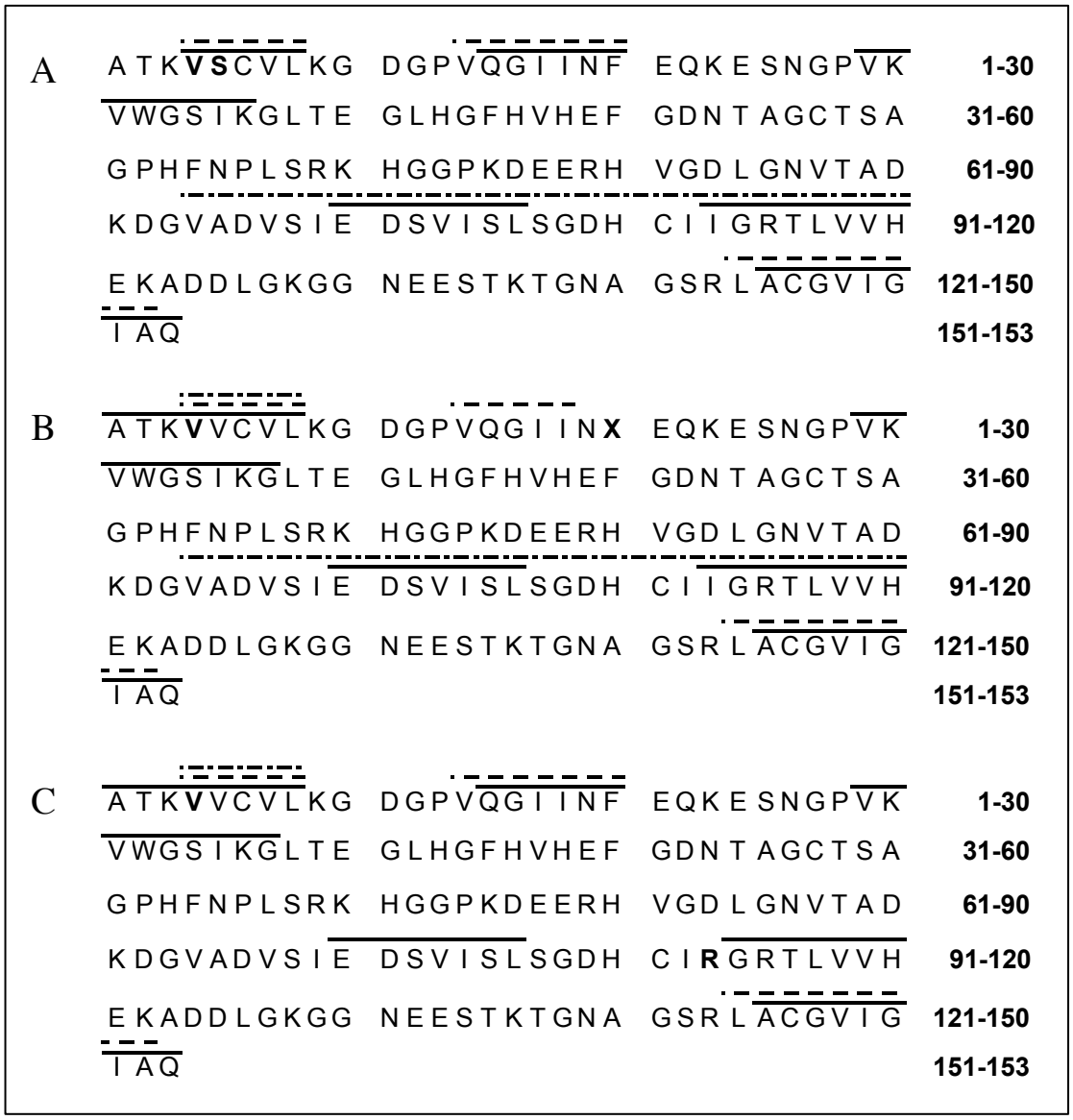


Figure 4.5: Potential aggregation-prone "hot spots", in the primary sequence of A) A4V-V5S, B) A4V-F20A/F20G where X denotes A or G, and C) A4V-I113R variants in hSOD1 predicted by three aggregation calculators. Lines above the sequences denote predicted "hot spots" for AGGRESCAN (solid line), PASTA (dash-dotted line), and TANGO (dashed line), and all variant sites are bolded.

Quantitative estimations to changes in the IAP of hSOD1

In addition to the qualitative IAP predictions obtained using the aggregation calculators, predictions from the 3D profiling method [Thompson 2006] were also obtained to quantitatively estimate changes in the IAP for candidate, second-site variants relative to A4V-hSOD1 using Eq. 4.2.

$$\Delta\Delta G_{assoc} = \Delta G_{assoc}^{\mathrm{var}} - \Delta G_{assoc}^{ref} \qquad \text{(Eq. 4.2)}$$

Here, $\Delta\Delta G_{assoc}$ represents the free energy for a variant sequence relative to a reference sequence of pairing two identical hexapeptides of a given sequence (containing the mutation site of each variant), using the known crystal structure of an amyloid peptide as the template. Using Eq. 4.2 for these calculations, the second-site variant sequences were used as the variant sequence, and the A4V-hSOD1 sequence was used as the reference sequence. As such, positive values of $\Delta\Delta G_{assoc}$ indicate a less conformationally stable amyloid structure is formed with the insertion of the variant sequence, and thus a favorable change in IAP, relative to the reference sequence. Since hSOD1 has previously been suggested to form amyloid fibrils [DiDonato, 2003, Chattopadhyay 2008], and the 3D profile method uses the same energy score function as RosettaDesign, this was considered the most direct way to quantitatively compare folding vs. IAP changes on a similar scale. For this work, $\Delta\Delta G_{assoc}$ was computed for each candidate, second-site variant in keeping with the original implementation of the 3D profiling method. Estimated values for $\Delta\Delta G_{assoc}$ are listed in Table 4.3, and more specific details regarding the 3D profiling method are provided in the following references [Thompson 2006, Nelson 2005].

Results showed the 3D profiling method estimated favorable IAP changes for all second-site variants, except for F20L. In general, the favorable IAP changes estimated for each second-site variant correlated well with the qualitative predictions from the other

three aggregation calculators. Notably, an unfavorable change in IAP was observed using this method when the A4V variant was inserted alone, without any second-site variant, into the hSOD1 sequence (data not shown).

Table 4.3: Summary of $\Delta\Delta G_{assoc}$ values, estimated by Eq. 4.2 using the 3D profile method, for several second-site variants identified by RosettaDesign. Here, positive values indicate a favorable change, or a reduction in the intrinsic aggregation propensity (IAP) of the molecule, after the insertion of A4V and the second-site variant.

Second-site variants	$\Delta\Delta G_{assoc}$ (kcal/mol)
A4V + K3R	4.8
A4V + V5S	0.9
A4V + F20A	0.1
A4V + F20G	0.6
A4V + F20L	-0.1
A4V + F20M	0.1
A4V + I113R	6.4
A4V + A152S	0.1

Equilibrium denaturation using circular dichroism (CD)

The conformational stability for wild type hSOD1 and three variants was measured via equilibrium, chemical unfolding experiments monitored by circular dichroism (CD). Shaojie Zhang conducted these experiments in the Department of Chemical Engineering at the University of Virginia. Figure 4.6 show the unfolding curves generated from these experiments for wild type hSOD1, A4V-hSOD1, and the second-site hSOD1 variants, A4V-F20G and A4V-F20A. Notably, *in vivo* experiments were previously conducted for the second-site variants A4V-F20L and A4V-F20M, however; the results showed aggregation was accelerated relative to A4V-hSOD1 (Appendix C), thus *in vitro* chemical unfolding and aggregation experiments were not performed for F20L and F20M.

A two-state unfolding model was applied to the unfolding curves shown in Figure 4.6 to try and quantify thermodynamic parameters, however, the fit was poor and produced large statistical uncertainties. This may suggest intermediate species, such as hSOD1 monomers, are present along the unfolding pathway. In fact, upon a visual examination of these data in Figure 4.6, an intermediate transition may occur between 2 to 4 M GdnHCl for the hSOD1 variants. Thus, future studies should try to apply a three-state unfolding model to these data to see if a better fit is obtained. Further, refolding experiments have thus far not been conducted for these variants and wild type hSOD1, and therefore any thermodynamic parameters that would be estimated by fitting these data to a two or three-state unfolding model should be considered apparent until true equilibrium conditions can be determined. Nonetheless, qualitative trends can still be visually observed. For instance, the unfolding transition for A4V-hSOD1 took place at lower GdnHCl concentrations, relative to wild type hSOD1, indicating decreased conformational stability. Further, the unfolding transitions for the hSOD1 second-site variants, A4V-F20A and A4V-F20G, take place at higher GdnHCl concentrations, relative to A4V-hSOD1, suggesting they may partially suppress the conformational destabilization observed for A4V-hSOD1; however they were still observed to be less stable than wild type hSOD1. It should be noted these experimental results are merely initial findings in an ongoing study, and only used here to qualitatively correlate the computational predictions to experimental results.

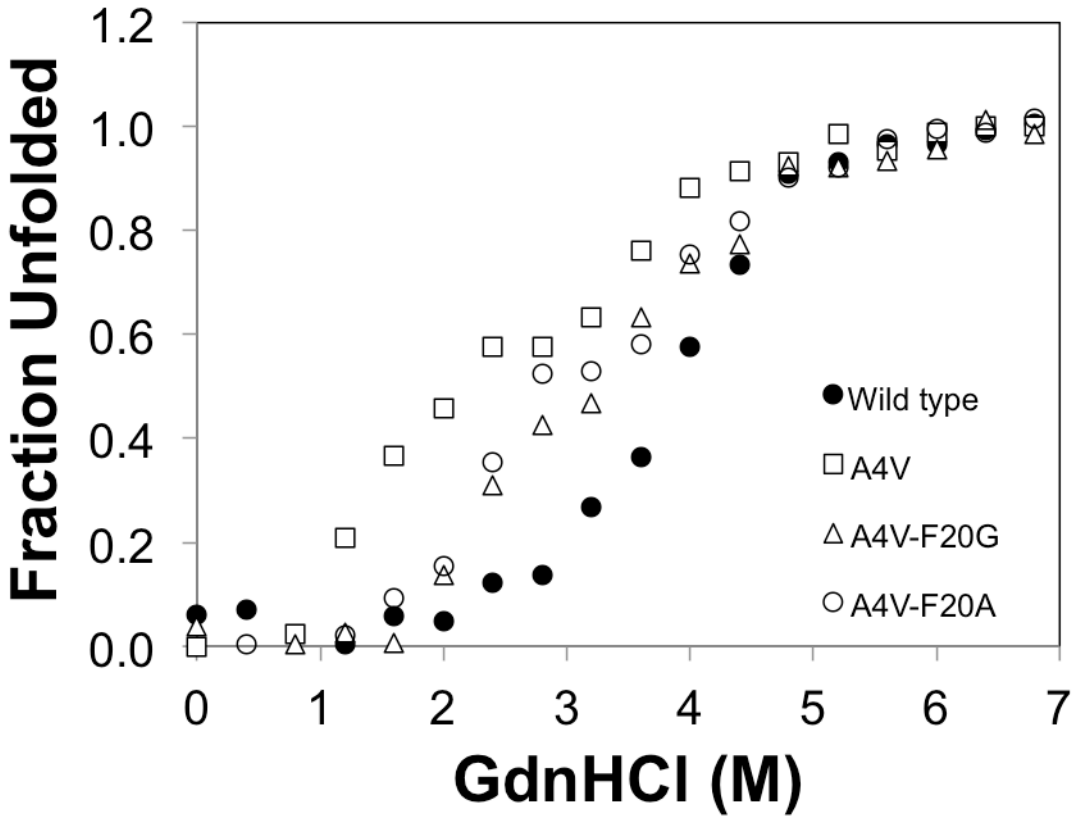


Figure 4.6: Chemical denaturation curves for wild type hSOD1, A4V-hSOD1, and the double hSOD1 variants, A4V-F20G and A4V-F20A. Points represent experimental data converted from CD ellipticity values at 218 nm to fractions of unfolded protein as a function of GdnHCl concentration.

Isothermal aggregation and analysis via size-exclusion chromatography

The aggregation of wild type hSOD1 and three variants was investigated by conducting isothermal aggregation experiments at physiological conditions (PBS buffer, T = 37 °C), and monitoring the extent of aggregation via HPLC-SEC. Shaojie Zhang conducted these aggregation experiments in the Department of Chemical Engineering at the University of Virginia. The aggregation behavior of each hSOD1 variant was compared qualitatively to wild type by visually comparing the elution peaks of hSOD1 dimer as a function of incubation time. These results are shown in the corresponding SEC chromatograms in Figure 4.7A-D.

Upon visual examination of Figure 4.7A, the chromatograms for wild type hSOD1 predominantly overlapped each other indicating a relatively constant dimer concentration remained, even after several hours of incubation. On the other hand, Figure 4.7B shows the native homodimer peak for A4V-hSOD1 decreased at a faster rate relative to wild type hSOD1. Further, the larger oligomer species were present at intermediate incubation times, which did not appear in wild type hSOD1, indicating enhanced aggregation behavior for A4V-hSOD1. However, these oligomer peaks were not observed upon incubation of the hSOD1 second-site variants A4V-F20A (Figure 4.7C) and A4V-F20G (Figure 4.7D), and the native homodimer peak area of these second-site variants was much slower to decay than A4V-hSOD1, but decayed faster compared to wild type hSOD1, even after several hours of incubation. These results indicate these second-site variants resisted aggregation more than A4V-hSOD1, although remain somewhat more aggregation-prone compared to wild type hSOD1. Notably, this ordering of aggregation correlates with the conformational stability trends observed in Figure 4.6, and the predictions from the aggregation calculators. As with the equilibrium denaturation experiments, these experimental results regarding aggregation kinetics are merely initial findings in an ongoing study. They are used here to qualitatively correlate the computational predictions to experimental results.

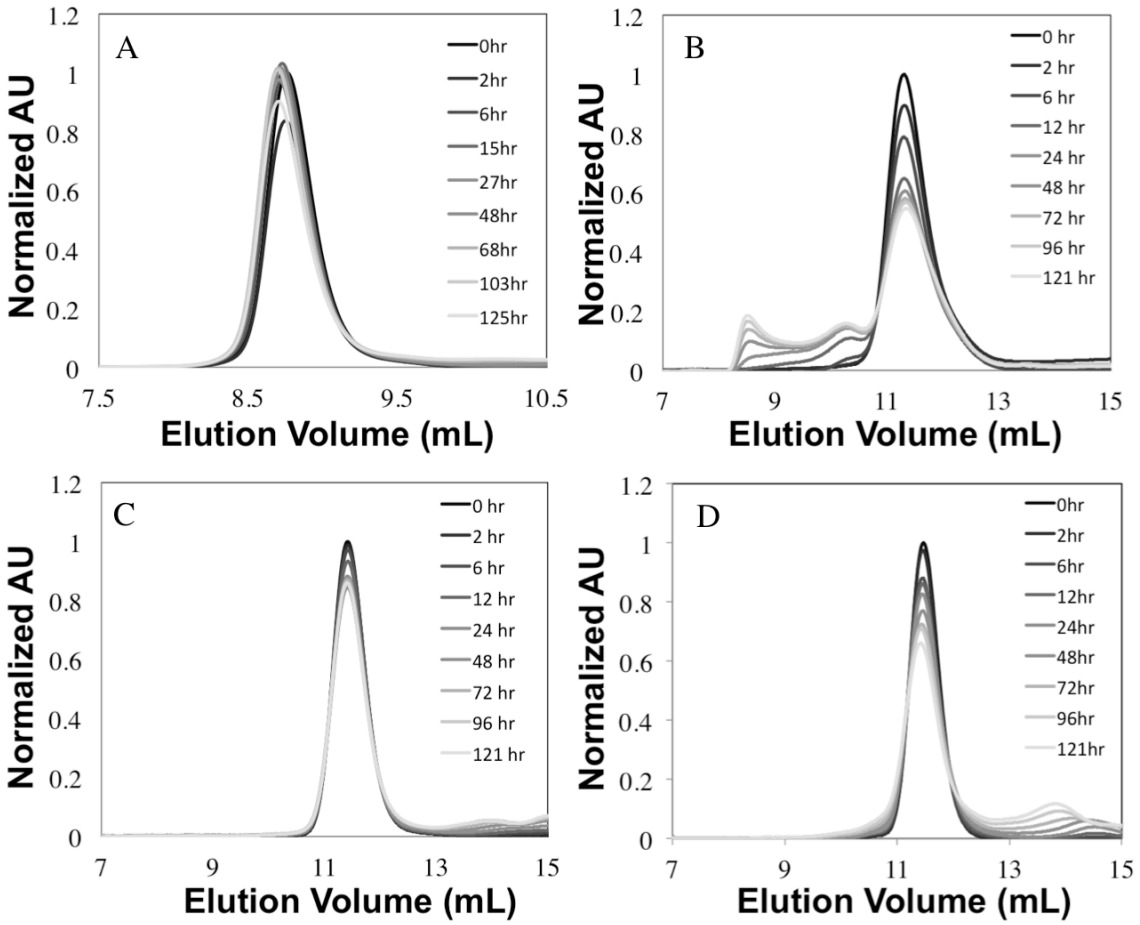


Figure 4.7: Aggregation time course at 37 °C monitored by HPLC-SEC for A) wild type hSOD1, B) A4V-hSOD1, and for the hSOD1 double variants, C) A4V-F20A and D) A4V-F20G.

4.4.Discussion

An intra-domain steric clash between F20 and A4V may instigate aggregation

The A4V variant is one of the most prevalent hSOD1 variants, particularly in North America, and is associated with nearly 50% of all fALS cases [Hough 2004]. Furthermore, A4V-hSOD1 is an extremely devastating variant as it induces rapid disease progression resulting in death within 1-2 years from the onset of symptoms [Galaleldeen 2009, Hough 2004]. Consequently in this study, the aggregation mechanism of A4V-hSOD1 was investigated using protein engineering and computational design

Upon examination of the A4V-hSOD1 tertiary structure, an intra-domain, steric clash was observed between F20 and the A4V variant site. This steric clash was hypothesized to destabilize the protein, and potentially initiate aggregation. By combining rational and computational design, two variants at the F20 location, F20A and F20G, were identified as candidate second-site variants in A4V-hSOD1 that would conformationally stabilize the protein and reduce the aggregation propensity by preventing the intra-domain, steric clash (Table 4.1 and 4.2).

Notably, even after proper energy minimization of the rotamer side chains, RosettaDesign did not suggest an altered orientation for hydrogens on the beta carbon of residue A20 compared to the native F20 residue, and thus the steric clash persisted with the A4V variant. As a result, the repulsive Lennard-Jones contribution within the energy function was negatively affected, and resulted in an overall, unfavorable energy score for the second-site variant F20A. It is conceivable that increasing the size of the rotamer library or using a post-design “relaxation” step might result in more favorable side-chain packing between F20A and A4V (and therefore more accurate energy predictions); however, this also increases the computation time considerably. Thus, there are other more expensive computational design approaches within RosettaDesign,

such as the flexible-backbone design or the relax mode application, that could also be implemented to try and achieve more accurate “hits” in design.

Nonetheless, RosettaDesign did manage to identify residue 20 in hSOD1 as being the most important residue to relieving the deleterious effects of the A4V mutation. Like the small side chain associated with alanine, glycine could also prevent the clash from taking place. Previous work has indicated that mutating non-glycine residues to glycine in an attempt to relieve a steric strain may increase conformational stability [Anil 2004]. Other studies have also shown the success of glycine to alanine substitutions intending to decrease the entropy of unfolding of a protein, thus increasing the stability of the folded state [Matthews 1987]. Thus, these variants were experimentally characterized and both were observed to restabilize the A4V-hSOD1 homodimer closer to the conformational stability of wild type hSOD1, as well as reduce aggregation relative to A4V-hSOD1 (Figures 4.6 and 4.7, respectively).

Further, RosettaDesign identified two other variants at the F20 location (e.g. F20L and F20M), to also improve the conformational stability of the A4V-hSOD1. These variants were also observed to prevent the intra-domain steric clash, despite being larger hydrophobic amino acids. However, the aggregation calculators did not predict a significant decrease in the IAP for F20L or F20M, and *in vivo* experiments conducted prior to this work indicated increased aggregation behavior for these second-site variants compared to A4V-hSOD1 (Appendix C). Thus they were not considered for further *in vitro* studies in this work.

The unfavorable change in conformational stability identified by RosettaDesign for F20A but not for larger, aliphatic residues such as F20L and F20M may indicate a unique molecular conformation is required to avoid the steric clash that is independent of side chain size. On the other hand, previous work discussed in Chapter 2 has shown that RosettaDesign successfully identifies stabilizing variants in approximately one-third

of design runs conducted, thus F20A may also be a scenario where RosettaDesign simply fails to identify the correct, repacked molecular conformation. As such, further experimentation of other F20 variants that incorporate substitutions of larger, hydrophobic amino acids into the protein sequence would be valuable. Nonetheless, these results suggested the aggregation mechanism of A4V-hSOD1 is intra-domain-related, and potentially caused by a steric clash between F20 and A4V.

Other potential aggregation mechanisms of A4V-hSOD1

Substituting the F20 residue in A4V-hSOD1 may not be the only way to prevent the suggested A4V-hSOD1 intra-domain steric clash. It is plausible that substituting other residues neighboring the A4V variant site may also provide the additional space to accommodate the valine residue. Thus, experimentally testing the other candidate second-site variants identified within this study as well as additional variants at these same residues may be valuable.

Of these variants, one particular residue to test is the I113 residue because it is located near the A4V variant site as well as the dimer interface, and thus, may be important for hSOD1 dimer binding (Figure 4.3). In fact, *in vivo* experiments conducted by Simpson Gregoire, prior to this work, showed I113M and I113V as second-site variants in A4V-hSOD1 that exhibited slight increases in aggregation resistance, relative to A4V-hSOD1 (Appendix C). Furthermore, RosettaDesign also identified I113V as having a favorable $\Delta\Delta G_f$ value, but only for the crystal structures 1N19 and 1UXM (Table 4.1); however the aggregation calculators showed no change in the IAP relative to A4V-hSOD1 after inserting I113V into the primary sequence (data not shown). Further, the computational design tools also did not identify I113M as a favorable second-site variant.

The *in vivo* experiments also identified increased aggregation resistance for C111S as a second-site variant, relative to A4V-hSOD1 (Appendix C). Further,

RosettaDesign identified C111S as a favorable variant as well. These results may indicate replacing free thiols within hSOD1 may also be important to reducing aggregation, particularly aggregation instigated by intermolecular disulfide bonding. Previous work has also reported mutating free thiols in hSOD1 may be critical to reducing aggregation [Cozzolino 2008, Niwa 2007, Tiwari 2003]. Further, intra-domain clashes, such as the one observed between F20 and A4V shown in this work, might result in protein unfolding that could expose free thiols and promote intermolecular disulfide bonding.

Finally, it is also noteworthy that using the approaches outlined in this work may also enable the elucidation of aggregation mechanisms for other ALS-associated variants in hSOD1, or other amyloidogenic proteins [Ross 2004]. For instance, a particularly relevant group of ALS-associated variants to study would be those located near the dimer interface that may affect interfacial binding. Previous work has shown the variants I113T, V148G, and I149T in hSOD1 are all located near the interface and associated with ALS [Valentine 2005].

Evaluating the success rates of computational design tools

Our previous work summarized in Chapter 2 implementing RosettaDesign and the aforementioned aggregation calculators concluded moderate success rates (one-third success rate for identifying conformationally stabilizing variants and aggregation-resistant variants, and a two-thirds success rate for identifying conformationally stabilizing and destabilizing variants) for the computational tools when applied to preventing protein aggregation via protein engineering techniques. Here, F20G was correctly identified by RosettaDesign and the aggregation calculators to improve the conformational stability and decrease the aggregation propensity of the protein, respectively, compared to the experimental results. However, the experimental results

for F20A indicated the variant was incorrectly identified by RosettaDesign to be conformationally destabilizing, but was correctly predicted by the aggregation calculators to decrease the aggregation propensity. Therefore, despite only experimentally characterizing two variants *in vitro*, the overall success rates of the computational design tools seem to be moderate as well (approximately 50 percent). Further, *in vivo* studies of additional second-site variants, such as I113V and C111S, indicated decreased aggregation (Appendix C), however; RosettaDesign only indicated a favorable change for C111S (approximately -0.5 kcal/mol for each crystal structure tested), and the aggregation calculators predicted no favorable changes in the IAP for these two variants. Therefore, a predictive success rate of around 50 percent was again observed using the computational tools for these variants as well.

It is noteworthy to mention that determining whether or not these success rates are generally acceptable is difficult because, to our knowledge, no standard benchmark for meaningful predictive yields or acceptable predictive success rates of these computational design tools has been set. Although a 50 percent success rate may be considered by some to be inadequate, to others it may be significantly better than selecting variants randomly without the use of these computational tools [Eijsink 2005, Richardson 2002, Palackal 2004]. Overall, the correlations that were observed demonstrate the benefits these computational design tools can provide; yet the moderate success rates highlight the need for further improved and optimized algorithms as well.

One possible confounding factor in using the aggregation calculators is that they are heavily influenced by the presence of β-sheet structure since that is important during amyloid formation [Caflisch 2006, Tartaglia 2005, Fernandez-Escamilla 2004, Conchillo-Sole 2007, Trovato 2007, Thompson 2006]. For hSOD1, previous work has reported conflicting results in terms of amyloid-like aggregate formation [Hwang 2010,

Chattopadhyay 2008]. Thus, sufficient β-sheet structure and previously observed amyloid-like aggregate formation ideally should be known for the protein under study to produce more reliable computational results. In addition, the ability to alter the environmental solution conditions (e.g. pH, ionic strength, etc.) for both RosettaDesign and the aggregation calculators would also most likely produce more confident computational predictions.

Interestingly however, RosettaDesign did identify favorable second-site variants located within four of the five “hot spots” identified by the aggregation calculators. Previous work also showed RosettaDesign identified favorable variants within each “hot spot” identified by the aggregation calculators for the human eye lens protein, γD-crys (C41T in the N-td and S130 variants in the C-td shown in Chapter 2). However, experimental studies did not show decreased aggregation for C41T and many of the S130 variants as discussed in Chapter 2. These results again highlight the need to balance changes in the intrinsic aggregation propensity and the conformational stability simultaneously when implementing protein engineering and computational design to deter protein aggregation.

General factors influencing the use of RosettaDesign to elucidate aggregation mechanisms and reduce protein aggregation

Examining the overall trends observed after using RosettaDesign within this study indicate general factors that may be important to successfully applying this tool to deterring protein aggregation. For instance, RosettaDesign was observed to mutate certain residues more than others during the global redesign runs. For example, the residues I18, V29, L106, I149, and G150 within the hSOD1 structure were not redesigned during any of the global design runs. This observation suggests RosettaDesign may have the ability to identify certain residues that are essential to

maintaining the overall conformational stability of the molecule, and thus must be conserved in protein design.

Additionally, RosettaDesign was observed to only identify favorable variants at the F20 position for starting structures based on wild type hSOD1 (e.g. 1PU0 and 2C9V). Comparing the main-backbone of the four crystal structures used in this analysis showed a bulge in the β-sheet where the A4V variant was located, that was not present in the wild type structure. This bulge is most likely caused by the substitution of the larger valine residue inserted into the sequence. Thus, if a second-site variant prevents the steric clash with the valine from A4V, and restabilizes the molecule to near wild type-like behavior, then a starting crystal structure containing a fixed, wild type-like, main backbone should theoretically produce more realistic, computational predictions. This may explain why favorable $\Delta\Delta G_f$ values were observed for F20G when using starting structures based on wild type hSOD1 (e.g. 1PU0 and 2C9V), but not for starting structures based on A4V-hSOD1 (e.g. 1N19 and 1UXM) (Table 4.2). Further, this suggests the starting crystal structure implemented during fixed-backbone computational design of proteins may be an important factor to consider, as well as analyzing multiple starting structures for the molecule when they are available.

Other factors are also worth mentioning that can significantly affect the predictions of RosettaDesign. For instance, using a flexible backbone design protocol versus a fixed backbone design protocol would most likely identify different variants, as well as more realistic alterations to the backbone structure as a result of the amino acid substitution. However, utilizing a flexible backbone design requires significantly more computational power compared to fixed backbone design. Another important factor to consider is the version of RosettaDesign being used, as different terms within the energy function and altered scoring weights have been observed to change the total energy scores of variants, as was seen with S130P and M69W as discussed in Chapter 2.

Furthermore, RosettaDesign estimates the lowest energy score via Monte Carlo optimization, therefore; conducting a sufficient number of design runs and outputted structures to ensure the lowest energy score has been determined is also important. Maintaining consistency of these factors throughout all design runs within a given study should produce valuable computational predictions.

Overall, however, the successes observed when implementing these computational design approaches during this study and previous studies indicate conformational stabilization can be a valid way to predict residues that should be targeted for mutation. Notably though, kinetics also are suggested to be important as some correlation was observed between the aggregation calculators and the 3D profiling method to the rates of aggregation observed experimentally.

4.5. Conclusions

In this work, various aggregation calculators and RosettaDesign were utilized to investigate the aggregation mechanism of the extremely severe A4V variant in human copper-zinc superoxide dismutase-1 (A4V-hSOD1), associated with amyotrophic lateral sclerosis (ALS). An intra-domain clash was identified between the F20 residue and the A4V variant site, potentially instigating conformational destabilization and subsequent aggregation. Via computational design, F20G and F20A were identified as candidate, second-site variants to restabilize the molecule and/or reduce the intrinsic aggregation propensity, relative to A4V-hSOD1. Preliminary experimental characterization showed both variants restabilized A4V-hSOD1 closer to the conformational stability of wild type hSOD1, and reduced aggregation relative to A4V-hSOD1. Nonetheless this study is ongoing and further experimentation is needed to more accurately report thermodynamic and aggregation parameters of these variants and wild type hSOD1.

Nevertheless, the prevention of this intra-domain steric clash seems critical to limiting the aggregation of A4V-hSOD1. Substituting exposed free thiols within hSOD1 as well as other neighboring residues of the A4V variant site may also aid in limiting aggregation. Furthermore, the computational design tools exhibited some correlation with the experimental results demonstrating their promise as protein engineering tools, and important factors were discussed regarding RosettaDesign to successfully identify aggregation resistant variants. Overall, the computational tools yielded moderate success rates as was seen in previous work, and further highlighted the need for improved algorithms for use in future studies involving protein engineering and computational design.

Chapter 5: Project Summary and Avenues for Future Work

5.1. Project Summary

In this dissertation, mutational strategies intending to reduce non-native protein aggregation were investigated for a multi-domain protein system guided by computational design. These strategies aimed to increase the conformational stability or reduce the aggregation propensity of the molecule. The human eye lens protein, gamma D crystallin (γD-crys), was used as the model system to investigate the aforementioned mutational strategies, and several computational design tools were implemented in tandem with rational design to select candidate variants that would populate each mutational strategy, some of which were experimentally characterized.

For instance, the computational design algorithm RosettaDesign and several sequence-based aggregation correlations (e.g. PASTA, AGGRESCAN, TANGO, 3D profiling method) were used to identify point variants that would alter the conformational stability or intrinsic aggregation propensity (IAP), respectively, of γD-crys. This was a broader, theoretical assessment of previous work as nine variants were tested to evaluate three mutational strategies intended to reduce the aggregation: (1) stabilizing the less stable domain, (2) increasing the binding interaction between the domains, and (3) decreasing the IAP of the polypeptide chain.

One variant from each strategy was observed to reduce aggregation rate coefficients relative to wild type γD-crys; however, incorrect predictions were also notable among each mutational strategy. Nonetheless, these results demonstrated each mutational strategy is capable of reducing aggregation in multi-domain proteins, which is applicable to many biopharmaceutical products and some human diseases. Further, experimentally measured unfolding free energies correlated well with aggregation rates coefficients; however, notable outliers were observed, in particular for those molecules

designed to alter the IAP while minimally affecting conformational stability. These results highlighted a need to consider a balance between altered conformational stability and IAP when using computational design and protein engineering to mitigate protein aggregation.

This work also demonstrated utilizing these computational tools in tandem could identify proteins variants with improved aggregation relative to wild type. This is particularly noteworthy as, to our knowledge; no prior studies have used these computational design tools together with these mutational strategies to successfully deter aggregation using single, point variants for multi-domain proteins. Thus, these computational tools are presented as approaches to be used to identify protein variants, *a priori* to experimental studies, which will deter non-native aggregation. Nonetheless, the systematic analysis of these computational design tools produced moderate success rates that were somewhat subjective as benchmarks for meaningful predictive values for protein designs or acceptable predictive yields for a given computational design tool have not yet been established. This indicates improved design tools incorporating both IAP and conformational stability are still desired to yield higher predictive success rates on a quantitative as well as qualitative basis.

Experimental and computational approaches were also presented to investigate the aggregation mechanism of γD-crys as well as another protein, A4V-human superoxide dismutase-1 (A4V-hSOD1), a variant of wild type hSOD1 associated to amyotrophic lateral sclerosis (ALS). Investigating the aggregation mechanism of each protein system was valuable as potential aggregation contacts and problematic molecular interactions were identified that if mutated, significantly deterred protein aggregation.

First, the aggregation mechanisms of three γD-crys variants (e.g. H22T, S130P, and S130T), all displaying diverse conformational stability and aggregation behavior, were examined experimentally relative to wild type using hydrogen-deuterium exchange coupled with mass spectrometry (HX-MS). Qualitative and quantitative assessments of these data were conducted to compare and contrast the aggregate and monomer conformations. Results showed the monomeric structures of all three γD-crys variants and wild type γD-crys were similar, however; the aggregate conformations showed significant differences.

For instance, HX-MS measurements suggested wild type γD-crys and S130T might form well-structured amyloid-like aggregates, while H22T and S130P form more flexible, less-structured aggregates. These observations correlated with the aggregation behavior previously measured for these γD-crys species, and highlight the promise of the mutational strategies to which they were affiliated as discussed in Chapter 2. Several aggregation contacts were identified experimentally for wild type and S130T, but few were identified for H22T and S130P. Nonetheless, a potential aggregation contact at N125-L133 was identified experimentally among all γD-crys species tested, and correlated with residues N125-L133 predicted by the aggregation calculators to be an aggregation-prone, “hot spot”. These observations suggested this region might be important to the aggregation mechanism of multiple γD-crys species. The correlation observed between the experimental results and computational predictions is promising, but notably, incorrect predictions were also observed. Nevertheless, these results further highlighted the use of HX-MS and these computational design tools as approaches to identify aggregation contacts of multi-domain proteins, although the need for improved computational design tools and optimized experimental protocols is still desired.

The aggregation mechanism of the severely destabilizing A4V variant in human copper-zinc superoxide dismutase-1 (A4V-hSOD1) was also investigated. The

aggregation of this protein is associated to ALS, and thus investigating the aggregation mechanism has physiological relevance. Here, RosettaDesign was utilized to investigate the aggregation mechanism of A4V-hSOD1 by identifying residues that could engage in problematic clashes with the A4V variant, and then suggest substitutions of these residues that would prevent clashes and restabilize the protein.

An intra-domain clash was identified between the phenylalanine at residue 20 with the valine at residue 4 in A4V-hSOD1, and was hypothesized to destabilize the protein and instigate aggregation. RosettaDesign identified F20G as a second-site variant that prevented the steric clash and improved conformational stability of A4V-hSOD1. In addition, several aggregation calculators identified both F20G and F20A as second-site variants that potentially could reduce the IAP of A4V-hSOD1. Experimental characterization of these variants showed both restabilized A4V-hSOD1 closer to the conformational stability of wild type hSOD1, and also reduced aggregation relative to A4V-hSOD1. Thus, the prevention of this intra-domain steric clash was observed to be critical in limiting the aggregation of A4V-hSOD1. Furthermore, the computational design tools exhibited some correlation to the experimental results demonstrating this methodology as a promising approach to investigate the aggregation mechanisms of proteins.

5.2. Potential Avenues for Future Work

This section discusses recommendations for additional experimentation regarding work conducted in this dissertation. Additionally, the development and optimization of various protocols that were used in this work is also suggested and discussed.

Integration of new and optimized computational design tools

One such avenue would be to integrate new versions of the computational design algorithms implemented within this dissertation, or utilize new computational design tools that were not used in this work to refine the design process.

Regarding, RosettaDesign, utilizing a flexible-backbone design protocol, rather than the fixed-backbone design protocol used in this work, may better identify alterations to the protein structure that can occur upon protein design. Considering perturbations along the protein backbone during protein design may estimate more realistic energy scores for candidate variants, and improve predictive yields. In this work, strong, quantitative correlations were not observed between energy scores from RosettaDesign and the experimental results; however, implementing flexible-backbone design may produce energy scores that better correlate, quantitatively, with experimentally determined free energies of unfolding. Further, using the most recent version of the RosettaDesign software is recommended, as optimized energy terms and scoring weights corresponding to each energy term within the energy function will be included. These energy terms and scoring weights can affect the overall energy score of protein designs.

Regarding the primary-sequence-based aggregation calculators used in this work, the development and implementation of algorithms capable of measuring changes in IAP are desired, along with algorithms that yield more quantitative IAP predictions to correlate with experimentally obtained, aggregation kinetic data. Further, using aggregation calculators that account for the tertiary structure of proteins rather than just the primary sequence would also be valuable [Chennamsetty 2010]. One such algorithm, known as the spatial aggregation propensity (SAP), has been shown in previous work to identify aggregation-prone regions of a protein based on the dynamic exposure and spatial proximity of hydrophobic residues [Chennamsetty 2010]. SAP

could be combined with RosettaDesign for this work to identify and redesign any exposed, hydrophobic patches that are close spatially, but not sequentially, for these two protein systems. Therefore, the primary-sequence-based aggregation calculators used here may not predict these regions as “hot spots”, although they may contribute to the protein aggregation mechanism.

Finally, the version of RosettaDesign used here and many of the aggregation calculators used in this work can only estimate changes in the conformational stability or aggregation propensity of a protein under physiological, environmental conditions. Therefore, if experiments are conducted at different conditions (e.g. acidic pH, low ionic strength), they cannot be inputted into the design protocol. This could alter protein design results, particularly for charged-related variants, and thus, the ability to redesign proteins and identify aggregation contacts at various solution conditions would be beneficial, and may also yield more reliable predictive results.

Implementing additional experimental techniques to characterize aggregates

Other experimental techniques could also be implemented in future studies to further characterize the aggregation behavior of both protein systems. Here, SEC was used to monitor the extent of aggregate formation via the disappearance of monomer for both proteins, however; using this analytical method alone only determines the composition of oligomers present in the protein solution, and requires the use of molecular standards to estimate the molecular weight of species. On the other hand, utilizing static or dynamic light scattering techniques inline with SEC can estimate additional properties of the aggregate species, such as their molecular weight without the use of molecular standards, radius of gyration, polydispersity, and second virial coefficients. Therefore, monomer concentration and aggregate growth can be simultaneously monitored. Previous work has successfully characterized γD-crys [Sahin

2011] and hSOD1 [Banci 2008] using SEC-MALS, and therefore, applying this technique to the variants studied in this work would be feasible. These data can then be modeled to estimate individual aggregation rate constants for nucleation and growth regimes, and may improve the correlation between aggregation rate coefficients and unfolding energies shown in Chapter 2 for γD-crys.

Implementing other experimental techniques such as several dye binding assays may also be helpful in further characterizing aggregate species. For instance, thioflavin-T and Congo red binding assays have been used in previous work to detect amyloid-like aggregates [Zhang 2010], even for hSOD1 [Banci 2008] and γD-crys [Sahin 2011, Papanikolopoulou 2008]. Conducting these experiments on wild type γD-crys and the S130T variant would be useful to determine if amyloid-like aggregates form for these species, as was suggested in Chapter 3 of this work.

Recommendations on further evaluating mutational strategies

Another potential area for future work is to simultaneously evaluate multiple mutational strategies intending to deter aggregation. This could be feasible by experimentally characterizing double protein variants. For example, characterizing a protein that includes a conformationally stabilizing variant, such as H22T in γD-crys, as well as a variant that reduces the aggregation propensity of the protein, such as S130P, would be interesting. One could then test whether the stability and aggregation behavior of one variant dominated over the other, or if a combination of behavior from both variants was observed.

Future studies could also focus on variants that were considered to populate multiple mutational strategies to distinguish whether they adhered to one particular strategy. For instance in γD-crys, M69Q and C41T were each placed in two of the three defined mutational strategies. As such, performing experiments to identify which

mutational strategy the variant best belonged to would also be beneficial to this work. As an example, the M69Q variant was originally considered a potential domain stabilizer and interface stabilizer. Thus, to elucidate the region of the molecule that was most stabilized by the variant, equilibrium unfolding experiments could be conducted for the isolated N-terminal domain, that includes the M69Q variant, and then subsequently compared to unfolding experiments on the whole molecule. This may then highlight whether the M69Q variant solely stabilizes the N-td or the domain interface of the entire protein.

Finally, experimentally characterizing a larger subset of variants that populate each mutational strategy would also help to more clearly distinguish the most promising strategy to deter aggregation. However, a very large group of variants would be needed to confidently and statistically identify one strategy as more promising than another, and thus this would require an abundant amount of time and experimental resources.

Optimization of HX-MS experimental protocol

Further optimization and development of the HX-MS experimental protocol implemented in this work would also be suggested. For instance, the statistical uncertainties observed for deuterium labeling of reporter peptides was higher in this work than in previous work by Houde et. *al*. [Houde 2011]. As such, testing several more sample replicates may reduce the statistical uncertainties, and identify more statistically significant differences in labeling for peptides.

Additionally, more reporter peptides were tested by Houde et *al*. for interferon-beta-1a (IFN), a protein similar in molecular weight to γD-crys, compared to the amount of reporter peptides analyzed in this work [Houde 2011]. Thus, to increase the number of reporter peptides, improvements and optimization of the protein digestion step, peptide chromatography, and MS/MS analysis is suggested. As an example, inefficient digestion

of γD-crys aggregates, in particular, may result from difficulties associated with γD-crys aggregate dissociation. Thus, it is suggested to conduct experiments that will identify chemical agents (e.g. chaotropic salts, detergents, etc.) that more effectively dissociate γD-crys aggregates and subsequently aid protein digestion.

Furthermore, the current HX-MS protocols require a substantial amount of time. Thus, the development and optimization of higher throughput experimental methods, such as those used by Houde et *al*., that still are capable of obtaining short peptide-level or residue-level resolution is desired.

Characterize additional second-site variants of A4V-hSOD1

Finally, studying additional second-site variants for A4V-hSOD1 is suggested. The F20G and F20A second-site variants were shown experimentally to improve the conformational stability and reduce aggregation of A4V-hSOD1, by potentially preventing an intra-domain, steric clash with the A4V variant. However, wild type hSOD1 still exhibited higher conformational stability and less aggregation compared to the second-site variants. Although F20G and F20A were suggested to prevent this intra-domain clash, the small side chains associated with glycine and alanine may have decreased important hydrophobic interactions within the subunit core of A4V-hSOD1, and resulted in more conformational destabilization compared to wild type. Therefore, substituting amino acids with larger hydrophobic side chains at this location, that also prevent the intra-domain steric clash, may re-establish more wild type-like conformational stability. However, amino acids with large hydrophobic side chains are more inclined to produce additional, destabilizing steric clashes as well.

RosettaDesign did identify second-site variants at this location that substituted amino acids with large hydrophobic side chains, such as F20L and F20M. These second-site variants were predicted by RosettaDesign to prevent the intra-domain steric

clash, but initial *in vivo* experiments conducted for these variants did not show favorable aggregation behavior relative to A4V-hSOD1 (Appendix C), and as such, subsequent *in vitro* experiments were not performed. Therefore, performing *in vitro* studies on these variants, as well as other F20 variants is suggested to determine if amino acids with larger aliphatic side chains inserted at F20 can prevent the steric clash, and also improve structural packing relative to the second-site variants F20A and F20G.

Likewise, it is recommended to test whether substituting other neighboring residues of the A4V variant site would also prevent destabilizing, steric clashes. For instance, the side chain of the I113 residue is close to the A4V variant, but located on the opposite side compared to the F20 residue. Previous work has shown interactions may occur between the A4V variant and the native I113 residue [Cardoso 2002]. Therefore, experimentally characterizing the stability and aggregation behavior of several I113 second-site variants may determine if other steric clash(s) involving A4V occur, and can also be prevented. Notably, *in vivo* studies (Appendix C) and computational results identified certain I113 variants (e.g. I113V) to reduce aggregation of A4V-hSOD1, strengthening this suggestion for future work.

Nonetheless, these studies would also require a substantial amount of time and experimental resources in order to express and purify adequate amounts of each protein needed for *in vitro* studies. Currently, final yields for expressing wild type hSOD1 and several variants are only on the order of a few milligrams, yet take weeks to complete. Thus, the development of new plasmid constructs and the optimization or increased-scale of current purification methods is also recommended to strive for increased protein expression and purification yields.

Investigating the aggregation mechanism of other disease-related proteins

Lastly, the work in Chapter 4 highlighted the promise of using computational design to investigate the aggregation mechanism of an ALS-associated protein variant. As such, a similar approach could also be promising to apply to other ALS-associated variants in hSOD1, as well as other proteins whose aggregation is associated with human diseases. For instance, past work has suggested the dissociation of hSOD1 dimer as the initial step in the aggregation mechanism [Hough 2004]. Therefore, utilizing RosettaDesign to investigate the aggregation mechanisms of other ALS-associated variants, particularly those located near the dimer interface, such as I113T, V148G, and I149T, may also be worthwhile. Alternatively, previous work has postulated the L166P mutation in the DJ-1 protein is associated with Parkinson's disease, and is caused by the destabilization of the dimer interface [Wilson 2003]. As a result, using RosettaDesign to determine if problematic interactions arise from the insertion of this mutation in DJ-1, as well as other disease-associated variants in other protein structures, could also be feasible.

References

Acosta-Sampson, L., King, J., Partially folded aggregation intermediates of human gammaD-, gammaC-, and gammaS-crystallin are recognized and bound by human alphaB-crystallin chaperone. *J. Mol. Biol.* 2010, *401*, 134–152.

Ahmad, S., Gromiha, M., Fawareh, H., Sarai, A., ASAview: Database and tool for solvent accessibility representation in proteins. *BMC Bioinf.* 2004, *5*, 1-5.

Aldington, S., Bonnerjea, J., Scale-up of monoclonal antibody purification processes. *J. Chromatogr. B.* 2007, *848*, 64–78.

Andrews J.M., Roberts, C.J., A Lumry-Eyring Nucleated Polymerization Model of Protein Aggregation Kinetics: 1. Aggregation with Pre-Equilibrated Unfolding. *J. Phys. Chem.* 2007, *111*, 7897-7913.

Anil, B., Song, B., Tang, Y., Raleigh, D.P., Exploiting the right side of the Ramachandran plot: substitution of glycines by D-alanine can significantly increase protein stability. *J. Am. Chem. Soc.* 2004, *126*, 13194–13195.

Bai, Y., Milne, J.S., Mayne, L., Englander, S.W. Primary structure effects on peptide group hydrogen exchange. *Proteins Struct. Funct. Genet.* 1993, *17,* 75-86.

Banci, L., Bertini, I., Boca, M., Girotto, S., et *al*., SOD1 and amyotrophic lateral sclerosis: mutations and oligomerization. PloS One. 2008, *3*, e1677.

Banci, L., Bertini, I., Durazo, A., Girotto, S., et. *al*., Metal-free superoxide dismutase forms soluble oligomers under physiological conditions: a possible general mechanism for familial ALS. *Proc. Natl. Acad. Sci. USA*. 2007, *104*, 11263–11267.

Banerjee, P.R., Puttamadappa, S.S., Pande, A., Shekhtman, A., et. *al*., Increased hydrophobicity and decreased backbone flexibility explain the lower solubility of a cataract-linked mutant of γD-crystallin. *J. Mol. Biol.* 2011, *412*, 647–659.

Bannen, R.M., Suresh, V., Phillips, G.N., Wright, S.J., et. *al*., Optimal design of thermally stable proteins. *Bioinformatics.* 2008, *24*, 2339-2343.

Berezovsky, I.N., Chen, W.W., Choi, P.J., and Shakhnovich, E.I., Entropic stabilization of proteins and its proteomic consequences. *PLOS Comput. Biol.* 2005, *1*, 0322-0332.

Bogdanov, M.B., Ramos, L.E., Xu, Z., and Beal, M.F. Elevated “hydroxyl radical” generation in vivo in an animal model of amyotrophic lateral sclerosis., *J. Neurochem.* 1998, *71*, 1321-1324.

Bratko, D., Cellmer, T., Prausnitz, J.M., and Blanch, H.W., Molecular Simulation of Protein Aggregation. *Biotechnol. Bioeng.* 2007, *96*, 1-8.

Bruijn, L.I. Aggregation and Motor Neuron Toxicity of an ALS-Linked SOD1 Mutant Independent from Wild-Type SOD1, *Science.* 1998, *281*, 1851-1854.

Bruijn, L.I., Becher, M.W., Lee, M.K., Anderson, K.L., et. *al*. ALS-linked SOD1 mutant G85R mediates damage to astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions., *Neuron.* 1997, *18*, 327-338.

Caflisch, A., Computational models for the prediction of polypeptide aggregation propensity. *Curr. Opin. Chem. Biol.* 2006, *10*, 437-444.

Cardoso, R.M.F., Thayer, M.M., DiDonato, M., Lo, T.P., Insights into Lou Gehrig's Disease from the Structure and Instability of the A4V Mutant of Human Cu,Zn Superoxide Dismutase. *J. Mol. Biol.* 2002, *324*, 247–256.

Carpenter, J.F., Ludwig, D.B., Webb, J.N., Ferna, C., et. *al*., Quaternary Conformational Stability : The Effect of Reversible Self-Association on the Fibrillation of Two Insulin Analogs. *Biotech. Bioeng.* 2011, *108*, 2359–2370.

Cellmer, T., Bratko, D., Prausnitz, J.M., and Blanch, H.W., Protein aggregation in silico. *Trends Biotechnol.* 2007, *25*, 254-261.

Chattopadhyay, M., Durazo, A., Sohn, S.H., Strong, C.D.,et. *al*., Initiation and elongation in fibrillation of ALS-linked superoxide dismutase. *Proc. Natl. Acad. Sci. USA.* 2008, *105*, 18663–18668.

Chattopadhyay, M., Valentine, J.S., Aggregation of copper-zinc superoxide dismutase in familial and sporadic ALS. Antioxidants & redox signaling 2009, *11*, 1603–1614.

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.

Chi, E.Y., Krishnan, S., Randolph, T.W., and Carpenter, J. F., Physical Stability of Proteins in Aqueous Solution: Mechanism and Driving Forces in Nonnative Protein Aggregation. *Pharm. Res.* 2003, *20*, 1325-1336.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., et *al*., Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*. 2003, *424*, 805-808.

Chrunyk, B.A., Wetzel, R., Breakdown in the relationship between thermal and thermodynamic stability in an interleukin-1-beta point mutant modified in a surface loop. *Protein Engr.* 1993, *6*, 733-738.

Conchillo-Solé, O., de Groot, N.S., Avilés, F.X., Vendrell, J., et. *al*., AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinf.* 2007, *8*, 65.

Cozzolino, M., Amori, I., Pesaresi, M.G., Ferri, A., et. *al*., Cysteine 111 affects aggregation and cytotoxicity of mutant Cu,Zn-superoxide dismutase associated with familial amyotrophic lateral sclerosis. *J. Biol. Chem.* 2008, *283*, 866–874.

Crabbe, M.J., Goode, D., Protein Folds and Functional Similarity; The Greek Key/Immunoglobin Fold. *Computers Chem.* 1995, *19*, 343-349.

Cromwell, M.E.M., Hilario, E., Jacobsen, F., Protein Aggregation and Bioprocessing. *AAPS J.* 2006, *8*, E572-E579.

Crow, J.P., Sampson, J.B., Zhuang, Y.X., Thompson, J.A., et *al*. Decreased zinc affinity of amyotrophic lateral sclerosis-associated superoxide dismutase mutants leads to enhanced catalysis of tyrosine nitration by peroxynitrate., *J. Neurochem.*1997, *69*, 1936-1944.

Dai, S.Y., Fitzgerald, M.C., A mass spectrometry-based probe of equilibrium intermediates in protein-folding reactions. *Biochemistry*. 2006, *45*, 12890–12897.

Dantas, G. Kuhlman, B., Callender, D., Wong, et. *al*., A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* 2003, *332*, 449-460.

Das, R., and Baker, D., Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* 2008, *77*, 363-382.

De Groot, A.S., Scott, D.W., Immunogenicity of protein therapeutics. *Trends Immunol.* 2007, *28*, 482–490.

de Groot, N.S., Aviles, F.X., Vendrell, J., Ventura, S., Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *The FEBS J.* 2006, *273*, 658–668.

Del Mar, C., Greenbaum, E.A., Mayne, L., Englander, S.W., et. *al*., Structure and properties of alpha-synuclein and other amyloids determined at the amino acid level. *PNAS*. 2005, *102*, 15477-15482.

DiDonato, M., Craig, L., Huff, M.E., Thayer, M.M., ALS Mutants of Human Superoxide Dismutase Form Fibrous Aggregates Via Framework Destabilization. *J. Mol. Biol.* 2003, *332*, 601-615.

Doucette, P.A., Whitson, L.J., Cao, X., Schirf, V., et. *al*., Dissociation of human copper-zinc superoxide dismutase dimers using chaotrope and reductant. Insights into the molecular basis for dimer stability. *J. Biol. Chem.* 2004, *279*, 54558–54566.

Du, J., Murphy, R.M., Characterization of the interaction of beta-amyloid with transthyretin monomers and tetramers. Biochemistry. 2010, *49,* 8276–8289.

Dunbrack Jr., R.L. Cohen, F.E., Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* 1997, *6*, 1661-1681.

Durazo, A., Shaw, B.F., Chattopadhyay, M., Faull, K.F., et. *al*., Metal-free superoxide dismutase-1 and three different amyotrophic lateral sclerosis variants share a similar partially unfolded beta-barrel at physiological temperature. *J. Biol. Chem.* 2009, *284*, 34382–34389.

Dzwolak, W., Loksztejn, A., Smirnovas, V., New Insights into the Self-Assembly of Insulin Amyloid Fibrils: An H-D Exchange FT-IR Study. *Biochemistry*. 2006, *45*, 8143-8151.

Ebrahim-Habibi, A., Morshedi, D., Rezaei-Ghaleh, N., Sabbaghian, M., et. *al*., Protein-Protein Interactions Leading to Aggregation: Perspectives on Mechanism, Significance and Control. *J. Iran. Chem. Soc.* 2010, *7*, 521-544.

Eijsink, V.G.H., Bjørk, A., Sigrid, G., Sirevåg, R., et. *al*., Rational engineering of enzyme stability. *J. Biotechnol.* 2004, *113*, 105-120.

Eijsink, V.G.H., Gåseidnes, Borchert, T.V., van den Burg, B., Directed evolution of enzyme stability. *Biomol. Eng.* 2005, *22*, 21-30.

Englander S.W., Kallenbach, N.R., Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q. Rev. Biophys.* 1983, *16*, 521-655.

Estevez, A.G., Crow, J.P., Induction of nitric oxide-dependent apoptosis in motor neurons by zinc-deficient superoxide dismutase. *Science*. 1999, *286*, 2498-2500.

Fernandez-Escamilla, A.M. Rousseau, F., Schymkowitz, J., Serrano, L., Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 2004, *22*, 1302-1306.

Fernandez-Lafuente, R., Stabilization of multimeric enzymes: Strategies to prevent subunit dissociation. *Enzyme Microb. Tech.* 2009, *45*, 405-418.

Fink, A. L., Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold. Des.* 1998, *3*, R9-R23.

Flaugh, S.L., Kosinski-Collins, M.S., King, J., Contributions of hydrophobic domain interface interactions to the folding and stability of human γD-crystallin. *Protein Sci.* 2005a, *14*, 569-581.

Flaugh, S.L., Kosinski-Collins, M.S., King, J., Interdomain side-chain interactions in human γD crystallin influencing folding and stability. *Protein Sci.* 2005b, *14*, 2030-2043.

Flaugh, S.L., Mills, I.A., King, J., Glutamine deamidation destabilizes human γD-crystallin and lowers the kinetic barrier to unfolding. *J. Biol. Chem.* 2006, *281*, 30782-30793.

Foguel, D., Robinson, C. R., de Sousa, P. C. Jr., Robinson, A.S., et. *al*. Hydrostatic pressure rescues native protein from aggregates. *Biotechnol. Bioeng.* 1999, *63*, 552-558.

Forrer, P., Binz, H.K., Stumpp, M.T., and Plückthun, A. Consensus design of repeat proteins. *ChemBioChem.* 2004, *5*, 183-189.

Frand, A.R., Cuozzo, J.W., Kaiser, C.A., Pathways for protein disulphide bond formation. *Trends Cell Biol.* 2000, *10*, 203–210.

Gagnon, P., Technology trends in antibody purification. *J. Chromatogr. A.* 2012,*1221*, 57–70.

Galaleldeen, A., Strange, R.W., Whitson, L.J., Antonyuk, S.V., et. *al*., Structural and biophysical properties of metal-free pathogenic SOD1 mutants A4V and G93A. *Arch. Biochem. Biophys.* 2009, *492*, 40–47.

Gamble, C.N., The role of soluble aggregates in the primary immune response of mice to human gamma globulin. *Int. Arch. Allergy Appl. Immunol.*1966, *30*, 446–455.

Gerk, L.P., Leven, O., and Müller-Hill, B., Strengthening the dimerisation interface of Lac repressor increases its thermostability by 40 deg. C. *J. Mol. Biol.* 2000, *299*, 805-812.

Getzoff, E.D., Tainer, J.A., Stempien, M.M., Bell, G.I., et. *al*. Evolution of CuZn Superoxide Dismutase and the Greek Key beta-Barrel Structural Motif. *Proteins*, *5,* 322-336.

Goihberg, E., Dym, O., Tel-Or, S., Shimon, L., Frolow, F., et. *al*., Thermal stabilization of the protozoan Entamoeba histolytica alcohol dehydrogenase by a single proline substitution. *Proteins* 2008, *72*, 711-719.

Gurney, M.E., Pu, H., Chiu, A.Y., Dal Canto, M.C., Polchow, C.Y., et. *al*. Motor neuron degeneration in mice that express a human Cu,Zn superoxide dismutase mutation., *Science*. 1994, *264*, 1772-1775.

Hayward, L.J., Rodriguez, J.A., Kim, J.W., Tiwari, A., et. *al*., Decreased metallation and activity in subsets of mutant superoxide dismutases associated with familial amyotrophic lateral sclerosis. *J. Biol. Chem.* 2002, *277*, 15923–15931.

Hopper, E.D., Pittman, A.M.C., Tucker, C.L., Campa, M.J., et *al*., Protein-Ligand Binding Detection. *Anal. Chem.* 2009, *81*, 6860–6867.

Hoshino, M., Katou, H., Hagihara, Y., Hasegawa, K., et. *al*., Mapping the core of the beta(2)-microglobulin amyloid fibril by H/D exchange. *Nature Struct. Biol.* 2002, *9*, 332–336.

Houde, D., Berkowitz, S.A., Engen, J.R., The Utility of Hydrogen / Deuterium Exchange Mass Spectrometry in Biopharmaceutical Comparability Studies. *J. Pharm. Sci.* 2011, *100*, 2071–2086.

Hough, M.A., Grossmann, J.G., Antonyuk, S.V., Strange, R.W., et. *al*. Dimer destabilization in superoxide dismutase may result in disease-causing properties: structures of motor neuron disease mutants. *Proc. Natl. Acad. Sci. USA.* 2004, *101*, 5976–5981.

Hu, X., Wang, H., Ke, H., Kuhlman, B., Computer-Based Redesign of a β Sandwich Protein Suggests that Extensive Negative Design Is Not Required for De Novo β Sheet Design. *Structure*. 2008, *16*, 1799-1805.

Hvidt A., Nielsen S.O., Hydrogen exchange in proteins, Adv. Prot. Chem. 1966, *21*, 287-386.

Hwang, Y., Stathopulos, P.B., Dimmick, K., Yang, H., et. *al*., Nonamyloid aggregates arising from mature copper/zinc superoxide dismutases resemble those observed in amyotrophic lateral sclerosis. *J. Biol. Chem.* 2010, *285*, 41701–41711.

Ivanova, M.I., Thompson, M.J., Eisenberg, D., A systematic screen of beta2-microglobulin and insulin for amyloid-like segments. *Proc. Natl. Acad. Sci. U.S.A.* 2006, *103*, 4079-4082.

Jiang, L., Althoff, E.A., Clemente, F.R., Baker, D. et *al*., De Novo Computational Design of Retro-Aldol Enyzmes. *Science*. 2008, *319*, 1387-1391.

Johnston, J.A., Dalton, M.J., Gurney, M.E., and Kopito, R.R. Formation of high molecular weight complexes of mutant Cu, Zn-superoxide dismutase in a mouse model for familial amyotrophic lateral sclerosis., *Proc. Natl. Acad. Sci. USA.* 2000, *97*, 12571-12576.

Jung, J. Byeon, I.L., Wang, Y., King, J., et. *al*., The Structure of the Cataract-Causing P23T Mutant of Human γD-Crystallin Exhibits Distinctive Local Conformational and Dynamic Changes. *Biochemistry*. 2009, *48*, 2597-2609.

Kamerzell, T.J., Esfandiary, R., Joshi, S.B., Middaugh, C.R., et. *al.,* Protein-excipient interactions: mechanisms and biophysical characterization applied to protein formulation development. *Adv. Drug Delivery Rev.* 2011, *63*, 1118–1159.

Kheterpal, I., Cook, K.D., Wetzel, R., Hydrogen/deuterium exchange mass spectrometry analysis of protein aggregates. *Methods Enzymol.* 2006, *413*, 140–166.

Koide, T., Igarashi, S., Kikugawa, K., Nakano, R., et. *al*., Formation of granular cytoplasmic aggregates in COS7 cells expressing mutant Cu/Zn superoxide dismutase associated with familial amyotrophic lateral sclerosis. *Neuroscience letters.* 1998, *257*, 29–32.

Kortemme, T., Morozov, A.V., Baker, D., An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes. *J. Mol. Biol.* 2003, *326*, 1239-1259.

Kosinski-Collins, M.S. Flaugh, S.L., King, J., Probing folding and fluorescence quenching in human γD crystallin Greek key domains using triple tryptophan mutant proteins. *Protein Sci.* 2004, *13*, 2223-2235.

Kosinski-Collins, M.S., King, J., In vitro unfolding, refolding, and polymerization of human γD crystallin, a protein involved in cataract formation. *Protein Sci.* 2003, *12*, 480-490.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., et. *al*., Design of a novel globular protein fold with atomic-level accuracy. *Science*. 2003, *302*, 1364–1368.

Lazaridis, T., Karplus, M., Effective Energy Function for Proteins in Solution. *Proteins*. 1999, *35*, 133-152.

Lee, S., Mahler, B., Toward, J., Jones, B., et. *al*., A single destabilizing mutation (F9S) promotes concerted unfolding of an entire globular domain in gammaS-crystallin. *J. Mol. Biol.* 2010, *399*, 320–330.

Lefebvre, B.G., Gage, M. J., and Robinson, A.S., Maximizing recovery of native protein from aggregates by optimizing pressure treatment. *Biotechnol. Progr.* 2004, *20*, 623-629.

Lehmann, M., and Wyss, M., Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotech.* 2001, *12*, 371-375.

Lindberg, M.J., Normark, J., Holmgren, A., Oliveberg, M., Folding of human superoxide dismutase: disulfide reduction prevents dimerization and produces marginally stable monomers. *Proc. Natl. Acad. Sci. USA.* 2004, *101*, 15893–15898.

Liu, R., Althaus, J.S., Ellerbrock, B.R., Becker, D.A., et. *al*. Enhanced oxygen radical production in a transgenic mouse model of familial amyotrophic lateral sclerosis., *Ann. Neurol.* 1998, *44*, 763-770.

Liu, Y., Kuhlman, B., RosettaDesign server for protein design. *Nucleic Acids Res.* 2006, *34*, W235-W238.

Logan, T., Clark, L., and Ray, S.S., Engineered disulfide bonds restore chaperone-like function of DJ-1 mutants linked to familial Parkinson's disease. *Biochemistry* 2010, *49*, 5624-5633.

Lu, Y., Yeung, N., Sieracki, N., Marshall, N.M., Design of functional metalloproteins. *Nature*. 2009, *460*, 855-862.

Lumry, R., Eyring, H., Conformation changes of proteins'. *J. Phys. Chem.* 1954, *58*, 110–120.

Lyons, T.J., Liu, H., Goto, J.J., Nersissian, A., et. *al*. Mutations in copper-zinc superoxide dismutase that cause amyotrophic lateral sclerosis alter the zinc binding site and the redox behavior of the protein., *Proc. Natl. Acad. Sci. USA.* 1996, *93*, 12240-12244.

Mahler, B., Doddapaneni, K., Kleckner, I., Yuan, C., et. *al*., Characterization of a transient unfolding intermediate in a core mutant of γS-crystallin. *J. Mol. Biol.* 2010, *405*, 840-850.

Mahler, H.C., Friess, W., Grauschopf, U., Kiese, S., Protein Aggregation: Pathways, Induction Factors and Analysis *J. Pharm. Sci.* 2009, *98*, 2909-2934.

Majoor-Krakauer, D., Willems, P.J., Hofman, A., Genetic epidemiology of amyotrophic lateral sclerosis. *Clin. Genet.* 2003, *63*, 83–101.

Makhatadze, G. I., Loladze, V.V., Ermolenko, D. N., Chen, X., et. *al*., Contribution of Surface Salt Bridges to Protein Stability: Guidelines for Protein Engineering. *J. Mol. Biol.* 2003, *327*, 1135-1148.

Matthews, B.W., Nicholson, H., Becktel, W.J., Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. USA.* 1987, *84*, 6663–6667.

Meersman, F., Dobson, C. M., and Heremans, K., Protein unfolding, amyloid fibril formation and configurational energy landscapes under high pressure conditions. *Chem. Soc. Rev.* 2006, *35*, 908-917.

Melnik, B.S., Povarnitsyna, T.V., Glukhov, A.S., Melnik, et. *al*., SS-Stabilizing Proteins Rationally: Intrinsic Disorder-Based Design of Stabilizing Disulphide Bridges in GFP. *J. Biomol. Struct. Dyn.* 2012, *29*, 817-824.

Miklos, A.E., Kluwe, C., Der, B.S., Pai S., et *al*. Structure-Based Design of Supercharged, Highly Thermoresistant Antibodies. *Chemistry & Biology.* 2012, *19*, 449-455.

Mills, I.A., Flaugh, S.L., Kosinski-Collins, M.S., King, J.A., Folding and stability of the isolated Greek key domains of the long-lived human lens proteins γD-crystallin and γS-crystallin. *Protein Sci.* 2007, *16*, 2427-2444.

Moore, W.V., Leppert, P., Role of Aggregated Human Growth-Hormone (HGH) in Development of Antibodies to HGH. *J. Clin. Endocrinol. Metab.* 1980, *5*, 691–697.

Münch, C., Bertolotti, A., Exposure of hydrophobic surfaces initiates aggregation of diverse ALS-causing superoxide dismutase-1 mutants. *J. Mol. Biol*. 2010, *399*, 512–525.

Murphy, R.M., Peptide aggregation in neurodegenerative disease. *Annu. Rev. Biomed. Eng.* 2002, *4*, 155-74.

Nelson, R. Sawaya, M.R., Balbirnie, M., Eisenberg, D. et. *al.,* Structure of the cross-β spine of amyloid-like fibrils. *Nature*. 2005, *435*, 773-778.

Niwa, J., Yamada, S., Ishigaki, S., Sone, J., et. *al*., Disulfide bond mediates aggregation, toxicity, and ubiquitylation of familial amyotrophic lateral sclerosis-linked mutant SOD1. *J. Biol. Chem.* 2007, *282*, 28087–28095.

Nowak, R.J., Cuny, G.D., Choi, S., Lansbury, P.T., Improving binding specificity of pharmacological chaperones that target mutant superoxide dismutase-1 linked to familial amyotrophic lateral sclerosis using computational methods. *J. Med. Chem.* 2011, *53*, 2709–2718.

Pace, C.N., Shirely, B.A., Thomson, J.A., Measuring the conformational stability of a protein, in: Protein Structure, a Practical Approach (Creighton, T.E., Ed.), IRL Press at Oxford University Press, 1989, 331-330.

Palackal, N., Brennan, Y., Callen, W.N., Dupree, P., et. *al*., An evolutionary route to xylanase process fitness. *Protein Sci.* 2004, *13*, 494-503.

Pande, A., Ghosh, K.S., Banerjee, P.R., Pande, J., Increase in Surface Hydrophobicity of the Cataract-Associated, P23T Mutant of Human GammaD-Crystallin is Responsible for its Dramatically Lower, Retrograde Solubility. Biochemistry. 2011, *49*, 6122–6129.

Pande, A., Pande, J., Asherie, N., Lomakin, A., et. *al*., Molecular basis of a progressive juvenile-onset hereditary cataract. *PNAS*. 2000, *97*, 1993-1998.

Papnikolopoulou, K., Mills-Henry, I., Thol., S.L., Wang, Y. et. *al.,* Formation of amyloid fibrils in vitro by human γD-crystallin and its isolated domains. *Mol. Vis.* 2008, *14*, 81-89.

Paterson, Y., Englander, S.W., Roder, H. An Antibody-Binding Site on Cytochrome-c Defined by Hydrogen-Exchange and 2-Dimensional NMR. *Science*. 1990, *4970*, 755-759.

Potter, S.Z., Valentine, J.S., The perplexing role of copper-zinc superoxide dismutase in amyotrophic lateral sclerosis (Lou Gehrig's disease). *J. Biol. Inorg. Chem.* 2003, *8*, 373–380.

Qi, W., Zhang, A., Patel, D., Lee, S., et. *al*., Simultaneous Monitoring of Peptide Aggregate Distributions, Structure and Kinetics Using Amide Hydrogen Exchange: Application to Abeta(1-40) Fibrillogenesis. *Biotechnol. Bioeng.* 2008, *100*, 1214–1227.

Rakhit, R., Chakrabartty, A., Structure, folding, and misfolding of Cu,Zn superoxide dismutase in amyotrophic lateral sclerosis. *Biochim. Biophys. Acta.* 2006, *1762*, 1025–1037.

Rakhit, R., Crow, J.P., Lepock, J.R., Kondejewski, L.H., et. *al*., Monomeric Cu,Zn-superoxide dismutase is a common misfolding intermediate in the oxidation models of sporadic and familial amyotrophic lateral sclerosis. *J. Biol. Chem.* 2004, *279*, 15499–15504.

Rakhit, R., Cunningham, P., Furtos-Matei, A., Dahan, S., et. *al*. Oxidation-induced misfolding and aggregation of superoxide dismutase and its implications for amyotrophic lateral sclerosis., *J. Biol. Chem.* 2002, *277*, 47551-47556.

Ray, S.S., Nowak, R.J., Brown, R.H., Lansbury, P.T., Small-molecule-mediated stabilization of familial amyotrophic lateral sclerosis-linked superoxide dismutase mutants against unfolding and aggregation. *Proc. Natl. Acad. Sci. USA* 2005, *102*, 3639–3644.

Ray, S.S., Nowak, R.J., Strokovich, K., Brown, R.H., et. *al*., An intersubunit disulfide bond prevents in vitro aggregation of a superoxide dismutase-1 mutant linked to familial amytrophic lateral sclerosis. *Biochemistry* 2004, *43*, 4899-4905.

Reaume, A.G., Elliott, J.L., Hoffman, E.K, Kowall, N.W., et. *al*. Motor neurons in Cu/Zn superoxde dismutase-deficient mice develop normally but exhibit enhanced cell death after axonal injury. *Nat. Genet.* 1996, *13*, 43-47.

Remmele Jr., R.L., Bhat, S.D., Phan, D.H., Gombotz, W.R., Minimization of Recombinant Human Flt3 Ligand Aggregation at the T_m Plateau: A Matter of Thermal Reversibility. *Biochemistry*. 1999, *38*, 5241-5247.

Richardson, T.H., Tan, X.Q., Frey, G., Callen, W., et. *al.*, A novel, high performance enzyme for starch liquefaction – Discovery and optimization of a low pH, thermostable alpha-amylase. *J. Biol. Chem.* 2002, *277*, 26501-26507.

Rodriguez, J.A., Valentine, J.S., Eggers, D.K., Roe, J.A., et. *al*. Familial amyotrophic lateral sclerosis-associated mutations decrease the thermal stability of distinctly metallated species of human copper/zinc superoxide dismutase. *J. Biol. Chem.* 2002, *277*, 15932–15937.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D., Protein structure prediction using Rosetta. Methods in enzymology 2004, *383*, 66–93.

Rosenberg, A.S., Immunogenicity of biological therapeutics: a hierarchy of concerns. *Dev. Biol.* 2003, *112*, 15-21.

Ross, C.A., Poirier, M.A.,Protein aggregation and neurodegenerative disease. *Nat. Med.* 2004, *10*, S10–S17.

Rousseau, F., Schymkowitz, J., and Serrano, L., Protein aggregation and amyloidosis: confusion of the kinds? *Curr. Opin. Struc. Biol.* 2006, *16*, 118-126.

Routledge, K.E., Tartaglia, G.G., Platt, G.W., Vendruscolo, M., et. *al*. Competition between Intramolecular and Intermolecular Interactions in an Amyloid-Forming Protein. *J. Mol. Biol.* 2009, *389*, 776-786.

Sahin, E., Jordan, J.L., Spatara, M.L., Naranjo, A.N., Costanzo, J.A., et *al*., Computational Design and Biophysical Characterization of Aggregation-Resistant Point Mutations for γD Crystallin Illustrate a Balance of Conformational Stability and Intrinsic Aggregation Propensity. *Biochemistry*. 2011, 628-639.

Sammond, D.W., Eletr, Z.M., Purbeck, C., Kimple, et. *al*., Structure-based Protocol for Identifying Mutations that Enhance Protein-Protein Binding Affinities. *J. Mol. Biol.* 2007, *371*, 1392-1404.

Sauerborn, M., Brinks, V., Jiskoot, W., Schellekens, H., Immunological mechanism underlying the immune response to recombinant human protein therapeutics. *Trends Pharmacol. Sci.* 2010, *31*, 53–59.

Saven, J.G., Computational protein design: Advances in the design and redesign of biomolecular nanostructures. *Curr. Opin. Colloid Interface Sci.* 2010, *15*, 13-17.

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., et. *al*., Progress in Modeling of Protein Structures and Interactions. *Science*. 2005, *310*, 638-642.

Schuster, M.C., Ricklin, D., Papp, K., Molnar, K.S., et. *al*., Dynamic Structural Changes During Complement C3 Activation Analyzed by Hydrogen/Deuterium Exchange Mass Spectrometry. *Mol. Immunol.* 2008, *45*, 3142–3151.

Schwehm, J. M., Fitch, C.A., Dang, B. N., Bertrand García-Moreno E., et. *al*., Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects. *Biochemistry* 2003, *42*, 1118-1128.

Sekijima, Y., Dendle, M.T., Wiseman, R.L., White, J.T., et. *al*., R104H may suppress transthyretin amyloidogenesis by thermodynamic stabilization, but not by the kinetic mechanism characterizing T119 interallelic trans-suppression. *Amyloid*. 2006, *13*, 57-66.

Shah, P.S., Hom, G.K., Ross, S.A., Lassila, J.K., et. *al*., Full-sequence Computational Design and Solution Structure of a Thermostable Protein Variant. *J. Mol. Biol.* 2007, *372*, 1-6.

Shaw, B.F., Durazo, A., Nersissian, A.M., Whitelegge, J.P., et. *al*. Local unfolding in a destabilized, pathogenic variant of superoxide dismutase-1 observed with H/D exchange and mass spectrometry. *J. Biol. Chem.* 2006, *281*, 18167–18176.

Shaw, B.F., Valentine, J.S., How do ALS-associated mutations in superoxide dismutase-1 promote aggregation of the protein? *Trends Biochem. Sci.* 2007, *32*, 78–85.

Stathopulos, P.B., Rumfeldt, J.A.O., Scholz,G.A., Irani, R.A., et. *al*., Cu/Zn superoxide dismutase mutants associated with amyotrophic lateral sclerosis show enhanced formation of aggregates in vitro. Proc. Natl. Acad. Sci. U.S.A. 2003,*100*, 7021–7026.

Strange, R.W., Antonyuk, S., Hough, M.A., Doucette, P.A., et. *al*., The Structure of Holo and Metal-deficient Wild-type Human Cu, Zn Superoxide Dismutase and its Relevance to Familial Amyotrophic Lateral Sclerosis. *J. Mol. Biol.* 2003, *328*, 877–891.

Strange, R.W., Antonyuk, S., Hough, M.A., Doucette, P.A., et. *al*., Variable Metallation of Human Superoxide Dismutase: Atomic Resolution Crystal Structure of Cu-Zn, Zn-Zn, and As-isolated Wild-type enzymes. *J. Mol. Biol.* 2006, *356*, 1152-1162.

Strickler, S.S., Gribenko, A.V., Gribenko, A.V., Keiffer, et. *al*., Protein Stability and Surface Electrostatics: A Charged Relationship. *Biochemistry*. 2006, *45*, 2761-2766.

Takata, T., Smith, J.P., Arbogast, B., David, L.L., et. *al.*, Solvent accessibility of betaB2-crystallin and local structural changes due to deamidation at the dimer interface. *Exp. Eye Res.* 2010, *91*, 336–346.

Tartaglia, G.G., Cavalli, A., Pellarin, R., Caflisch, A., Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.* 2005, *14*, 2723-2734.

Tartaglia, G.G., Pawar, A.P., Campioni, S., Dobson, C.M., et. *al*. Prediction of Aggregation-Prone Regions in Structured Proteins. *J. Mol. Biol.* 2008, *380*, 425-436.

Thompson, M.J., Sievers, S.A., Karanicolas, J., Ivanova, M.I., et. *al*.,The 3D profile method for identifying fibril-forming segments of proteins. *PNAS*. 2006, *103*, 4074-4078.

Tian, P., Computational protein design, from single domain soluble proteins to membrane proteins. *Chem. Soc. Rev.* 2010, *39*, 2071-2082.

Tiwari, A., Hayward, L.J., Familial amyotrophic lateral sclerosis mutants of copper/zinc superoxide dismutase are susceptible to disulfide reduction. *J. Biol. Chem.* 2003, *278*, 5984–5992.

Tobler, S.A., Fernandez, E.J., Structural features of interferon-gamma aggregation revealed by hydrogen exchange. *Protein Sci.* 2002, *11*, 1340–1352.

Trautmann, M.E., Effect of the Insulin Analog [Lys(B28),Pro(B29)] on Blood-Glucose Control. *Horm. Metab. Res.*1994, *26*, 588–590.

Trovato, A., Seno, F., Silvio, F., Tosatto, C.E., The PASTA server for protein aggregation prediction. *Protein Eng. Des. Sel.* 2007, *20*, 521-523.

Tsemekhman, K., Goldschmidt, L., Eisenberg, D., Baker, D., Cooperative hydrogen bonding in amyloid formation. *Protein Sci.* 2007, *16*, 761–764.

Valentine, J.S., Doucette, P. a, Zittin Potter, S., Copper-zinc superoxide dismutase and amyotrophic lateral sclerosis. *Annu. Rev. Biochem.* 2005, *74*, 563–593.

Valentine, J.S., Hart, P.J., Misfolded CuZnSOD and amyotrophic lateral sclerosis. *P. Natl. Acad. Sci. USA*. 2003,*100*, 3617–3622.

van Reis, R., Zydney, A., Membrane separations in biotechnology. *Curr. Opin. Biotechnol.* 2001,*12*, 208–211.

Vázquez-Rey, M., and Lang, D.A., Aggregates in monoclonal antibody manufacturing processes. *Biotech. Bioeng*. 2011, *108*, 1494-1508.

Voynov, V., Chennamsetty, N., Kayser, V., Helk, B., Trout, B.L., Predictive tools for stabilization of therapeutic proteins. *Design.* 2009, *1*, 580–582.

Vucic, S., Kiernan, M.C. Pathophysiology of neurodegeneration in familial amyotrophic lateral sclerosis. *Curr. Mol. Med.* 2009, *9*, 255-272.

Wan, Y., Vasan, S., Ghosh, R., Hale, G., Cui, Z., Separation of Monoclonal Antibody Alemtuzumab Monomer and Dimers Using Ultrafiltration. *Biotech. Bioeng.* 2005, *90*, 422–432.

Wang, J. Fibrillar Inclusions and Motor Neuron Degeneration in Transgenic Mice Expressing Superoxide Dismutase-1 with a Disrupted Copper-Binding Site, *Neurobiol. Dis.* 2002, *10*, 128-138.

Wang, S.S.S., Tobler, S.A., Good, T.A., Fernandez, E.J., Hydrogen exchange-mass spectrometry analysis of beta-amyloid peptide structure. *Biochemistry.* 2003, *42*, 9507–9514.

Wang, Q., Johnson, J.L., Agar, N.Y.R., Agar, J.N., Protein Aggregation and Protein Instability Govern Familial Amyotrophic Lateral Sclerosis Patient Survival. *PLoS Biol*. 2008, *6*, 1508-1526.

Wang, W., Protein aggregation and its inhibition in biopharmaceutics. *Int. J. Pharm.* 2005, *289*, 1-30.

Wang, W., Nema, S., Teagarden, D., Protein aggregation--pathways and influencing factors. *Inter. J. Pharm.* 2010, *390*, 89–99.

Wang, W., Singh, S.K., Li, N., Toler, M.R., et. *al*., Immunogenicity of protein aggregates-concerns and realities. *Int. J. Pharm.* 2012, *31*, 1–11.

Watanabe, M., Dykes-Hoberg, M., Culotta, V.C., Price, D.L., et. *al*. Histological evidence of protein aggregation in mutant SOD1 transgenic mice and in amyotrophic lateral sclerosis neural tissues., *Neurobiol. Dis.* 2001, *8*, 933-941.

Webb, J.N., Webb, S.D., Cleland, J.L., Carpenter, et *al*., Partial molar volume, surface area, and hydration changes for equilibrium unfolding and formation of aggregation transition state: High-pressure and cosolute studies on recombinant human IFN-γ. *PNAS*. 2001, *98*, 7259-7264.

Weiss IV, F. W., Young, T. M., and Roberts, C. J., Principles, Approaches, and Challenges for Predicting Protein Aggregation Rates and Shelf Life. *J. Pharm. Sci.* 2009, *98*, 1246-1277.

Weksler, M.E., Bull, G., Schwawrz, G.H., Stenzel, K.H., et. *al*. Immunologic Responses of Graft Recipients to Antilymphocyte Globulin: Effect of Prior Treatment with Aggregate-Free Gamma Globulin. *J. Clin. Invest.*1970, *49*, 1589–1595.

Wetzel, R., Mutations and off-pathway aggregation of proteins. *Trends Biotechnol*. 1994, *12*, 193-198.

Williams, A.D., Portelius, E., Kheterpal, I., Guo, J., et. *al*., Mapping Aβ Amyloid Fibril Secondary Structure Using Scanning Proline Mutagenesis. *J. Mol. Biol.* 2004, *335*, 833–842.

Williams, J.C., Zeelen, J.P., Gitte, N., Vriend, G., et. *al*. Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. *Protein Engr.* 1999, *12*, 243-250.

Wilson, M.A., Collins, J.L., Hod, Y., Ringe, D., et. *al.*, The 1.1-Å resolution crystal structure of DJ-1, the protein mutated in autosomal recessive early onset Parkinson's disease. *Proc. Natl. Acad. Sci. USA*. 2003, *100*, 9256-9261.

Wörn, A., Plückthun, A., Different Equilibrium Stability Behavior of ScFv Fragments: Identification, Classification, and Improvement by Protein Engineering. *Biochemistry*. 1999, *38*, 8739-8750.

Wörn, A., Plückthun, A., Mutual Stabilization of VL and VH in Single-Chain Antibody Fragments, Investigated with Mutants Engineered for Stability. *Biochemistry*. 1998, *37*, 13120-13127.

Yim, M., Kang, J., Yim, H., Kwak, H., et. *al*. Cu , Zn-superoxide dismutase mutant : An enhancement of free radical formation due to a decrease in Km for hydrogen peroxide, *Proc. Natl. Acad. Sci. USA.* 1996, *93*, 5709-5714.

Zhang, A., Jordan, J.L., Ivanova, M.I., Weiss, W.F., Roberts, C.J., Fernandez, E.J., Molecular level insights into thermally induced α-chymotrypsinogen A amyloid aggregation mechanism and semiflexible protofibril morphology. Biochemistry 2010, *49*, 10553–10564.

Zhang, A., Singh, S.K., Shirts, M.R., Kumar, S., Fernandez, E.J., Distinct aggregation mechanisms of monoclonal antibody under thermal and freeze-thaw stresses revealed by hydrogen exchange. Pharmaceutical research 2012, *29*, 236–250.

Zhang, W., Czupryn, M.J., Free sulfhydryl in recombinant monoclonal antibodies. *Biotechnol. Prog*. 2002,*18*, 509–513.

Zheng, J., Ma, B., Tsai, C.-J., Nussinov, R., Structural stability and dynamics of an amyloid-forming peptide GNNQQNY from the yeast prion sup-35. *Biophys. J.* 2006, *91*, 824–833.

Appendix A

RosettaDesign command lines used for γD-crys studies

1) Repacking:

```
./fixbb.macosgccrelease -s 1HK0.pdb -resfile resfile.NATAA.txt -
ignore_unrecognized_res -no_optH –database
/Users/cheuser/Rosetta_3.0/minirosetta_database -score:weights
./reoptimized_standardwts.txt -mute core.io core.conformation core.pack
core.scoring -ndruns 10 -nstruct 10
```

2) Global redesign[a,b]:

```
./fixbb.macosgccrelease -s 1HK0_packed.pdb –resfile resfile.X.txt -
ignore_unrecognized_res -no_optH -database
/Users/cheuser/Rosetta_3.0/minirosetta_database -score:weights
./reoptimized_standardwts.txt -mute core.io core.conformation core.pack
core.scoring -ndruns 3 -nstruct 3
```

3) Point mutation[b,c]:

```
./fixbb.macosgccrelease -s 1HK0_packed.pdb –resfile
resfile.pointmut.txt ignore_unrecognized_res -no_optH -database
/Users/cheuser/Rosetta_3.0/minirosetta_database -score:weights
./reoptimized_standardwts.txt -mute core.io core.conformation core.pack
core.scoring -ndruns 10 -nstruct 10
```

[a]resfile.X.txt: file name for resfiles associated with N-td or interface design runs
[b]1HK0_packed.pdb: the renamed outputted pdb file with the lowest energy score produced from the ten repacking runs
[c]resfile.pointmut.txt: file name for resfiles associated with a given point mutation

RosettaDesign resfiles used for γD-crys studies

1) Repacking resfile name: resfiles.NATAA.txt

```
NATAA
EX 1 EX 2
USE_INPUT_SC

start
```

2) N-td or interface redesign resfile name: resfile.X.txt

```
NATAA
EX 1 EX 2
USE_INPUT_SC

start
x A ALLAAwc
x A EX 1 EX 2
x A USE_INPUT_SC
```

where *x* represents the residue number (e.g. x = 1 to 85 for N-td redesign)

3) Example of H22T point mutation resfile:

```
NATAA
EX 1 EX 2
USE_INPUT_SC

start
22 A EX 1 EX 2
22 A PIKAA T
```

Appendix B

RosettaDesign command lines used for hSOD1 studies

1) Repacking:

```
./fixbb.macosgccrelease -s X.pdb -resfile resfile.NATAA.txt -
ignore_unrecognized_res -no_optH –database
/Users/cheuser/Rosetta_3.0/minirosetta_database -score:weights
./reoptimized_standardwts.txt -mute core.io core.conformation core.pack
core.scoring -ndruns 10 -nstruct 10
```

2) Global redesign[b,c]:

```
./fixbb.macosgccrelease -s X_packed.pdb –resfile resfile.Y.txt -
ignore_unrecognized_res -no_optH -database
/Users/cheuser/Rosetta_3.0/minirosetta_database -score:weights
./reoptimized_standardwts.txt -mute core.io core.conformation core.pack
core.scoring -ndruns 3 -nstruct 3
```

3) Point mutation[b,d]:

```
./fixbb.macosgccrelease -s X_packed.pdb –resfile resfile.pointmut.txt
ignore_unrecognized_res -no_optH -database
/Users/cheuser/Rosetta_3.0/minirosetta_database -score:weights
./reoptimized_standardwts.txt -mute core.io core.conformation core.pack
core.scoring -ndruns 10 -nstruct 10
```

[a]X.pdb: file name for starting pdb file for each crystal structure analyzed (e.g. X = 1PU0, 2C9V, 1N19, or 1UXM)

[b]X_packed.pdb: the renamed outputted pdb file with the lowest energy score produced from the ten repacking runs for each crystal structure analyzed (e.g. X = 1PU0, 2C9V, 1N19, or 1UXM)

[c]resfile.Y.txt: file name for resfiles associated with global design runs involving a 5 Å diameter area around the A4V variant site

[d]resfile.pointmut.txt: file name for resfiles associated with a given point mutation

RosettaDesign resfiles used for hSOD1 studies

1) Repacking resfile name: resfiles.NATAA.txt

```
NATAA
EX 1 EX 2
USE_INPUT_SC

start
46 z NATRO
48 z NATRO
63 z NATRO
71 z NATRO
80 z NATRO
83 z NATRO
120 z NATRO
```

where *z* represents chain A or B. Residues H46, H48, H63, H71, H80, D83, and H120 are all involved in metal binding sites, and therefore were not allowed to be repacked or substituted.

2) Global redesign resfile name: resfile.Y.txt

```
NATAA
EX 1 EX 2
USE_INPUT_SC
```

```
start
x A ALLAAwc
x A EX 1 EX 2
x A USE_INPUT_SC
46 z NATRO
48 z NATRO
63 z NATRO
71 z NATRO
80 z NATRO
83 z NATRO
120 z NATRO
```

where *x* represents residue numbers located < 5 Å from the A4V variant site, and *z* represents chain A or B.

3) Example of F20G point mutation resfile:

```
NATAA
EX 1 EX 2
USE_INPUT_SC

start
20 A EX 1 EX 2
20 A PIKAA G
46 z NATRO
48 z NATRO
63 z NATRO
71 z NATRO
80 z NATRO
83 z NATRO
120 z NATRO
```

where *z* represents chain A or B.

Appendix C

C.1 Materials and Methods

Flow Cytometric Analysis of Cellular Fluorescence

Simpson Gregoire developed the following protocol and conducted these experiments in the Department of Chemical Engineering at the University of Virginia. Two days post-transfection, transfected HEK293T and NSC-34 cells were prepared for flow cytometry. The cells were first were trypsinized, washed twice with PBS, and resuspended in 500μL of 1X PBS. The fluorescence intensities of the cells expressing SOD1 variant-EGFP fusion protein were measured using a C6 flow cytometer (Accuri, Ann Arbor, MI). The excitation and emission wavelengths used were 488 nm and 585 nm, respectively. Only GFP positive cells were used to calculate the mean cellular fluorescence. Transfection efficiency was determined by dividing the number of fluorescence positive cells by the number of total cells analyzed. Three replicates of each sample were analyzed and an average cellular fluorescence was recorded. Statistical uncertainties were estimated using a two-sided, Student's paired t-test.

C.2 Results

Figure C.1 shows the cellular fluorescence from flow cytometric analysis of several second-site variants in A4V-hSOD1, relative to A4V-hSOD1 and wild type hSOD1. Here, an increase in cellular fluorescence suggests decreased protein aggregation. Not surprisingly, wild type hSOD1 was shown to produce the highest cellular fluorescence among all protein sequences tested. Nonetheless, decreased aggregation was suggested from these results for several F20 variants (e.g. F20A, F20G, and F20V), but for F20L or F20M (data not shown for F20M). Additionally, increased cellular fluorescence was also observed for I113V relative to A4V-hSOD1, but not for I113M, because it behaved more similarly to A4V-hSOD1. Furthermore, replacing the free thiol at C111 with serine showed a significant increase in cellular fluorescence, but several variants at I149 that were also tested all showed decreased cellular fluorescence relative to A4V-hSOD1.

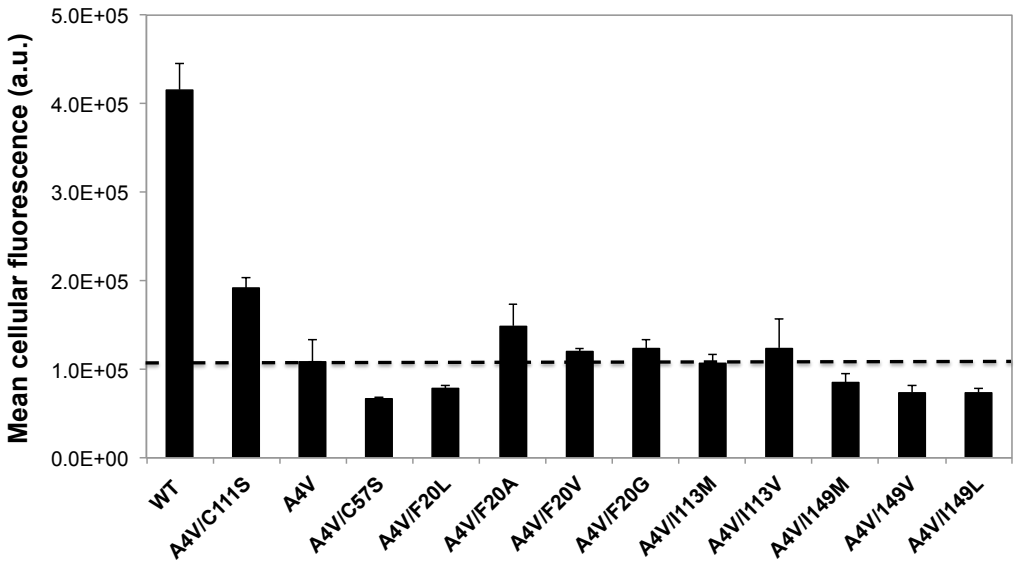


Figure C.1: *In vivo* experimental studies monitoring the cellular fluorescence upon flow cytometric analysis of several second-site variants in A4V-hSOD1 relative to A4V-hSOD1 and wild type hSOD1. Here, an increase in cellular fluorescence suggests decreased protein aggregation. The black dashed line helps compare the cellular fluorescence observed for A4V-hSOD1 compared to the other proteins tested.