# Monte Carlo Simulation Approaches to Protein Stability and Aggregation Prediction

# Monte Carlo simulation approaches to protein stability and aggregation prediction

a dissertation presented

by

Jaie Christina Woodard

to

The Committee on Higher Degrees in Biophysics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biophysics

Harvard University

Cambridge, Massachusetts

March 2017

Monte Carlo simulation approaches to protein stability and aggregation prediction

Abstract

A protein's sequence and set of covalent modifications determine its stability and aggregation propensity in a given environment. Given a change in sequence or covalent structure, we would like to be able to predict the change in stability and tendency to aggregate. Such knowledge would enable us to engineer more stable proteins and to better understand protein misfolding and aggregation diseases. In addition, knowledge of the protein folding pathway and aggregate structure could aid in structure-based design of therapeutics.

This thesis employs Monte Carlo simulations to predict protein stability, aggregation propensity, and aggregate structure. First, we describe the use of short unfolding simulations to predict stabilized mutants of the enzyme Dihydrofolate Reductase. Next, we describe a simple model of protein domain swapping that predicts the tendency of proteins to domain swap at intermediate temperature and predicts a concentration dependence where proteins domain swap at intermediate concentration but exhibit non-specific interactions between unfolded proteins at high concentration. Finally, we predict that cataract-associated mutations within γD-crystallin destabilize the protein and that these mutations, along with an experimentally observed disulfide bond, increase the protein's propensity to aggregate. Based on two-molecule simulations, we propose an aggregation model whereby the N-terminal hairpin of one molecule forms antiparallel beta sheet interactions with the C-terminal domain of the next molecule. We also suggest a

mechanism by which the wild-type protein could accelerate mutant aggregation, an experimentally observed phenomenon. We expect our methods to be applicable to stability and aggregation prediction in other proteins.

# Contents

*To my parents*

*and to my colleagues at Harvard*

# Acknowledgements

First, I'd like to thank those who helped me to get started within the Shakhnovich group, particularly Muyoung Heo for his help with the simulation code. Thank you also to Martha Bulyk and Leonid Mirny for serving as mentors in my first rotation, which influenced my approach to my thesis project.

I benefited from discussions, both on and off topic, with colleagues: Louis, Chris, Adrian, Nicholas, Amy, Murat, Will, Michael, Ariel, Sanchari, Jeong-Mo, and many others. Thank you to members of the Biophysics program and to Jim and Michele for their encouragement and for helping me navigate through the program.

This work stemmed from extensive collaboration with others, especially Eugene Serebryany, Jian Tian, and Sachith Dunatunga. Finally, I'd like to acknowledge my thesis committee: Collin Stultz, Shamil Sunyaev, and Jonathan King, and my advisor Eugene Shakhnovich, who encouraged me to pursue my ideas and who served as my mentor throughout graduate school.

# 1. Introduction

Most proteins must fold stably into their specified three-dimensional structure in order to carry out their biological function. While much has been learned about the factors contributing to protein stability, we still cannot predict with full accuracy how a given mutation or a covalent modification will affect the folding stability of a protein or the propensity of the protein to aggregate. This limits our understanding of how proteins can evolve and how we might engineer a protein to be more stable and less aggregation prone. This thesis utilizes molecular simulation approaches to study protein stability and protein-protein interactions such as domain swapping and aggregation. In the Introduction, we outline basic concepts from studies of protein folding and aggregation. We then describe a computational tool that can be used to simulate protein dynamics, which is used in the work described in this thesis but was introduced in previous publications by other authors.

## *1.1. Protein stability*

Protein folding is a cooperative process, often approximated by a two-state model, where the protein resides either in the native (folded) or unfolded state (Privalov, 1979; Shakhnovich and Finkelstein, 1989; Zeldovich et al., 2007). The equilibrium stability of a protein is then quantified by the difference in free energy

between these states, $\Delta G_{eq}$ (see Figure 1.1). Proteins unfold at high temperatures, due to the higher entropy of the unfolded state. The temperature at which the folded and unfolded states are equally populated ($\Delta G = 0$) is called the unfolding temperature, $T_m$. Mutations in protein sequence can either stabilize or destabilize the protein, and the change in stability is quantified by $\Delta\Delta G$ or $\Delta T_m$.

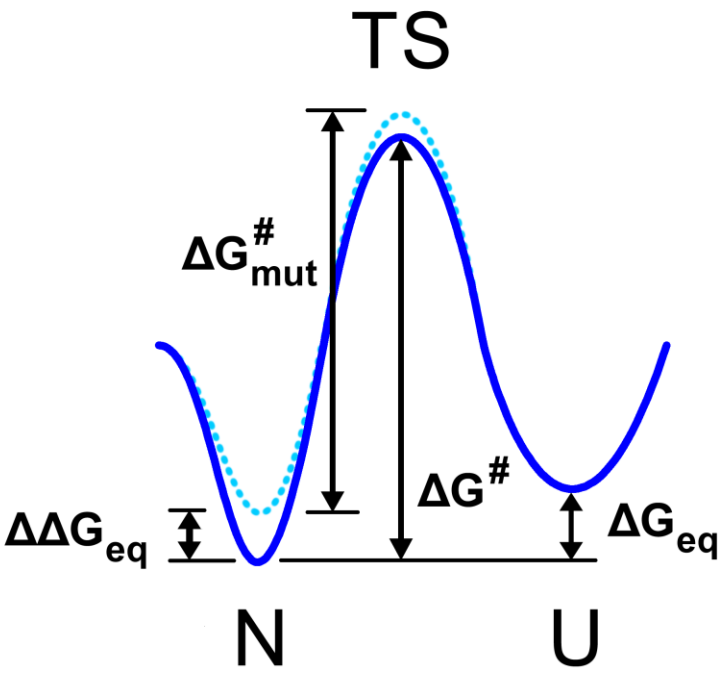



**Figure 1.1.** Two-state model of protein stability. N represents the folded native state, TS the transition state, and U the unfolded state.

The fraction of folded protein depends sigmoidally on the equilibrium stability, according to Boltzmann statistics, as shown in Figure 1.2. A destabilizing mutation ($\Delta\Delta G > 0$) can therefore reduce the amount of folded, functional protein at equilibrium. In addition, in the cellular environment, unfolded and partially unfolded proteins can aggregate or be degraded (Bershtein et al., 2013). A sufficient amount of folded protein must be present in order for a protein to carry out its

function (Goldstein, 2011; Wylie and Shakhnovich, 2011), and misfolded/aggregating proteins can themselves be deleterious to the cell (Drummond and Wilke, 2008). Protein stability therefore constrains protein evolution, avoiding mutational trajectories that would lead to unstable proteins.

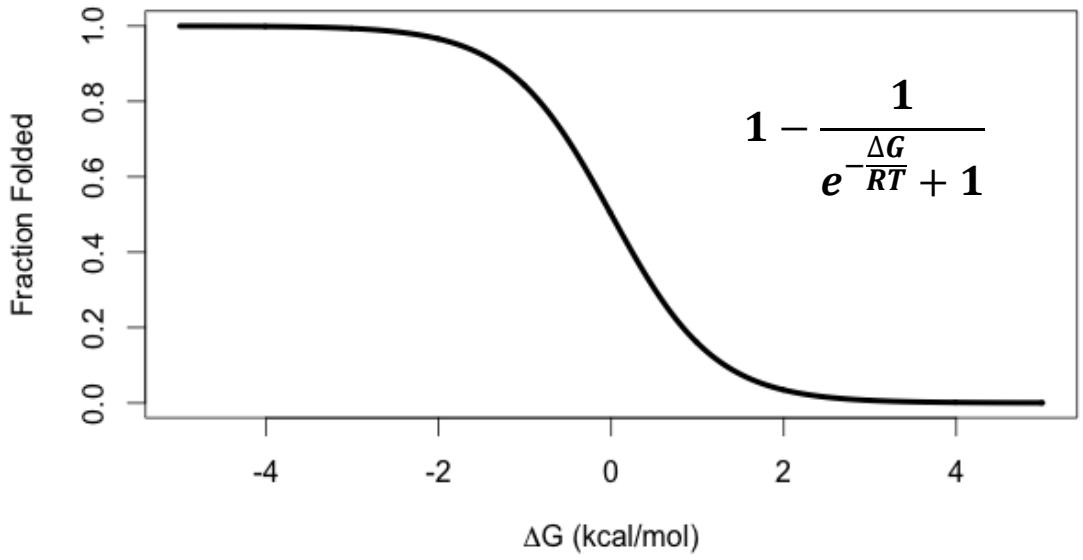



**Figure 1.2.** Fraction of folded protein as a function of equilibrium stability.

Most of the possible single mutations in a biological protein will be destabilizing, with mutations in the protein core leading to a greater stability loss on average, compared with mutations at the protein surface (Tokuriki et al., 2007). A small fraction of mutations would stabilize the protein. These generally include mutation of an active site residue to a hydrophobic amino acid, which in addition to stabilizing the protein would render it inactive. Costs of high stability can include loss of the ability to efficiently regulate the amount of protein and, possibly, loss of functionally-relevant dynamics for overly-stable proteins (Beadle and Shoichet,

2002; DePristo et al., 2005), although this thesis argues against the universality of this second type of stability-activity trade-off.

The folding transition state is defined as the state or ensemble of states that is equally likely to proceed to the native state as it is to proceed to the unfolded state (see Figure 1.1). The height of the energy barrier at the transition state determines the rate of interconversion between native and unfolded states. A destabilizing mutation will generally increase the free energy of the native state by some amount ΔΔG and will increase the free energy of the transition state by some fraction of this amount. This fraction is quantified by a residue's ϕ value, which describes the role of the residue in the transition state and thus the effect that a mutation will have on the folding and unfolding rates (Fersht and Sato, 2004; Matouschek et al., 1989). For ϕ = 1, the residue is folded in the transition state, so that $\Delta\Delta G^{\#}$, or the change in energy difference between the native and transition state for a mutation, is zero. In this case, the rate of folding decreases, while the rate of unfolding remains the same. For ϕ = 0, the residue is unfolded in the transition state, so that $\Delta\Delta G^{\#}$ is equal to $\Delta\Delta G_{eq}$, and the rate of unfolding increases, while the rate of folding remains the same. More generally, for residue *i*,

$$\Delta\Delta G_i^{\#} = (1 - \varphi_i)\Delta\Delta G_i^{eq} \qquad \text{(Equation 1.1)}$$

where $\Delta\Delta G^{\#}$ is defined as the change in the free energy barrier height relative to the native state, as depicted in Figure 1.1.

The validity of the two-state approximation to protein folding has been questioned in some cases, particularly for large proteins, with greater than about 110 amino acids (Braselmann et al., 2013). Proteins may contain one or more folding intermediates, which may be on pathway to the folded state, or off pathway, producing a "misfolded" state. Intermediate species may be prone to aggregation or to degradation within the cell. The two-state model is also challenged in the case of intrinsically disordered proteins, and in rare cases where the protein may adopt multiple folds (Bryan and Orban, 2010).

### *1.2. Protein folding in the cell*

Models of protein folding in the cellular environment must be extended to include protein production and degradation, chaperones, and the potential for aggregation (Figure 1.3). It has been hypothesized that many proteins do not fully unfold in the cell but are degraded, sequestered by chaperones, or aggregate from their partially unfolded forms (Braselmann et al., 2013; Dobson, 2003). Chaperones can prevent aggregation of partially unfolded proteins and can help to restore proteins to their native state. Chaperones therefore act to buffer deleterious mutations, allowing a larger number of destabilizing mutations to accumulate and speeding the rate of protein evolution (Tokuriki and Tawfik, 2009). At the same time, errors in protein production at the transcriptional or translational levels can increase the number of proteins that harbor destabilizing mutations.

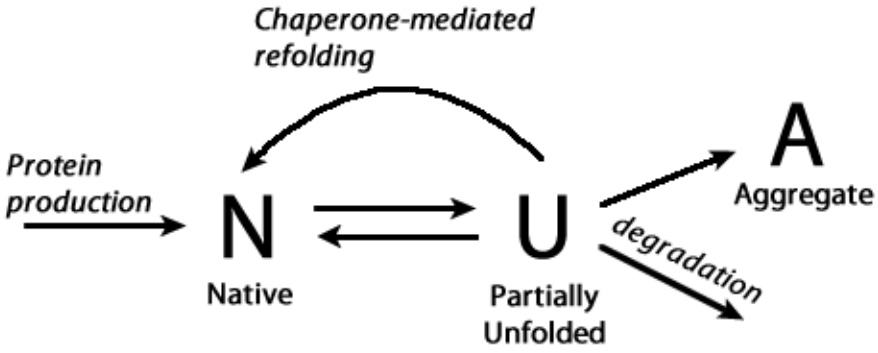



**Figure 1.3.** A simple model of protein folding in the cell

Molecular crowding, due to the high density of proteins and other biomolecules in the cell, can shift the population of folded vs. unfolded protein and can affect aggregation rate. For instance, excluded volume effects will stabilize the native state, which is more compact than the unfolded state. However, the nature of the interactions with surrounding molecules can also be important, and in some cases crowding can increase the population of the unfolded state (Danielsson et al., 2015; Elcock, 2010; Kuznetsova et al., 2015).

### *1.3. Domain swapping*

Domain swapping is a type of protein-protein interaction in which a structural element is exchanged between proteins, such that native-like contacts are formed with the complementary portion of the other protein (Liu and Eisenberg, 2002). An example of a protein that exhibits domain swapping, RNase A, is shown in Figure 1.4. The "swapped" region is often an alpha helix, beta strand, or beta hairpin, although it can be a larger structural element or an entire domain. The region of the

protein that changes conformation to allow partial unfolding of the monomer into an "open" state prone to domain swapping is known as the "hinge loop."

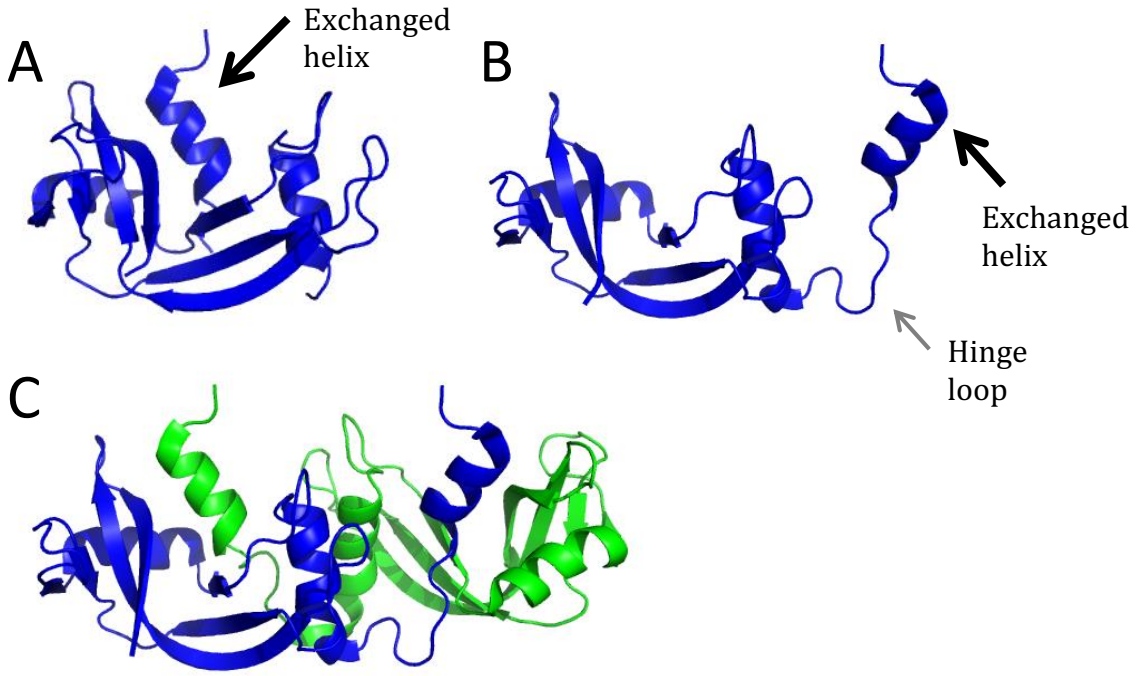



**Figure 1.4.** Protein domain swapping in RNase A (PDB structures 3BCM, 3BCO). The N-terminal alpha helix (black arrows) is exchanged between structures. A) Closed monomer. B) Open monomer from domain swapped structure. C) Domain-swapped dimer.

Many proteins have been crystallized as domain-swapped dimers, although the functional relevance of this interaction in most cases remains uncertain. One hypothesis is that "run-away" domain swapping may lead to aggregation (Rousseau et al., 2003). In this model, the first protein binds to the second protein, which binds to a third protein, etc., via domain-swap interactions. Domain swapping may also provide insight into protein folding pathways and intermediate folding states that may be involved in aggregation.

### *1.4. Protein aggregation*

Protein aggregation is usually categorized into two types: amyloid and amorphous. In amyloid aggregation, beta strands stack via hydrogen bonding interactions to form an ordered structure (see Figure 1.5), leading to the formation of long fibrils. By contrast, amorphous aggregates are less ordered, and it has been hypothesized that this type of aggregation involves the interaction of unfolded segments via non-specific contacts. However, there is evidence that amorphous aggregation may in fact involve specific interactions (Horwich, 2002; Speed et al., 1996). As noted in the previous section, domain swapping has been proposed as a possible mechanism leading to aggregate formation.

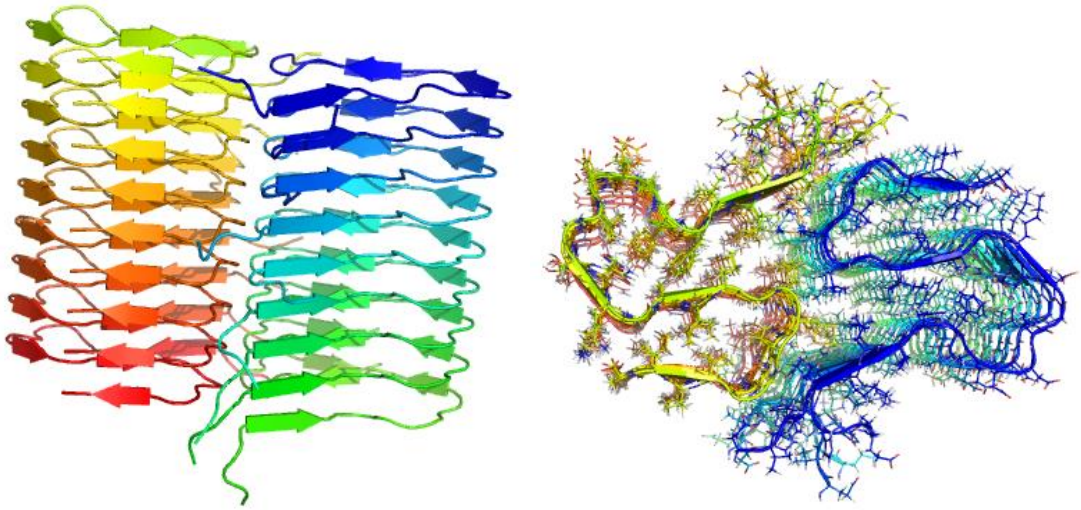


**Figure 1.5.** NMR structure of amyloid beta fibrils (PDB ID 5KK3). Left: side view. Right: Top-down view.

### *1.5. Monte Carlo protein simulation*

Molecular Dynamics (MD) and Monte Carlo (MC) simulations have been used extensively to study protein folding and dynamics. While both techniques make use

of simplified models to simulate molecular motion, they differ in their fundamental approach. Molecular dynamics simulations integrate classical equations of motion to predict the position of each atom as a function of time, given an initial set of positions and velocities. Monte Carlo simulations perform a set of moves, such as rotation about a torsional angle, which are accepted or rejected depending on the change in energy relative to the simulation temperature. The commonly-used Metropolis criterion for accepting a move is

$$p_{accept} = \begin{cases} 1, & \Delta E < 0 \\ e^{-\frac{\Delta E}{kT}}, & \Delta E > 0 \end{cases} \qquad \text{(Equation 1.2)}$$

The Shakhnovich group has developed an all non-hydrogen atom Monte Carlo simulation program. The program is described in detail in previous publications (Xu et al., 2011; Yang et al., 2007; Yang et al., 2008). Briefly, the move set consists of rotations about torsional angles, with bonds and angles held fixed. The energy function is a sum of terms:

$$E = E_{con} + w_{trp}E_{trp} + w_{hb}E_{hb} + w_{sct}E_{sct} + w_{trp}E_{trp} + w_{aro}E_{aro} \qquad \text{(Equation 1.3)}$$

where $E_{con}$ is the contact energy, $E_{trp}$ represents the torsional preferences of amino acid triplets, $E_{hb}$ is a directional hydrogen bonding term, $E_{sct}$ represents side chain torsional preferences, and $E_{aro}$ biases the relative orientations of aromatic residues towards the orthogonal direction. Two residues are said to be in contact if their

outer radii overlap. A "clash" occurs if inner radii overlap; in this case, the attempted move is rejected. The potential is knowledge based, with energies derived from statistics of real protein structures in the protein data bank. For instance, the contact energy between atom types A and B,

$$E_{AB} = \frac{-\mu N_{AB} + (1-\mu)\widetilde{N}_{AB}}{\mu N_{AB} + (1-\mu)\widetilde{N}_{AB}} \qquad \text{(Equation 1.4)}$$

where $N_{AB}$ is the number of contacts observed between atom types A and B, $\widetilde{N}_{AB}$ is the number of pairs not in contact, and $\mu$ is chosen so that the average value of $E_{AB}$ is zero.

The program outputs atomic coordinates, as well as simulation energy, number of contacts between atoms, and RMSD from the native structure, at frequencies specified by the user.

# 2. Predicting stabilized mutants of Dihydrofolate Reductase using Monte Carlo unfolding simulations

*Abstract*

Mutations in amino acid sequence can alter protein stability, an important factor in protein evolution and protein design. Here we introduce a method based on Monte Carlo unfolding simulations to predict stability effects of mutations. We predict relative stabilities for all possible point mutants of the enzyme Dihydrofolate Reductase. We find good agreement between simulation-based predictions and experimental measurements ($r$ = 0.68) for WT and 42 mutants. We identify 10 new stabilizing mutations, out of 23 experimentally tested mutations predicted to be stabilizing. The most stabilizing mutation, D27F, is located in the active site and renders the protein inactive. However, in general we see a positive correlation between stability and catalytic activity. By combining stabilizing mutations, we engineer a catalytically active DHFR mutant with experimental denaturation temperature 7.2 °C higher than WT.

## *2.1. Introduction*

Accurate prediction of protein stability is important in enzyme design, as well as in understanding aspects of protein evolution and human disease (Dobson, 2003; Liberles et al., 2012; Serohijos and Shakhnovich, 2014; Shakhnovich, 2006).

While most mutations will be destabilizing for a biological protein, some will generally be stabilizing (Tokuriki et al., 2007). By stabilizing a protein, we increase the fraction of the protein that resides in the folded, functional state, while decreasing the propensity of the protein to aggregate from the unfolded or partially unfolded state. However, some stabilizing mutations will negatively affect protein function, including those in the active site of an enzyme, where specific residues are required for catalysis. The question of whether there exists a more general stability-activity trade-off, due to a requirement that a protein must be sufficiently dynamic, is still debated (Adamczyk et al., 2011; Beadle and Shoichet, 2002; Bloom et al., 2006; DePristo et al., 2005; Studer et al., 2014).

Several computational methods to predict protein stability or stability change upon mutation have been developed and tested. However, the performance of these popular methods is still relatively low (Khan and Vihinen, 2010; Potapov et al., 2009; Thiltgen and Goldstein, 2012). While ideally, simulation-based prediction of protein stability would involve the simulation of multiple folding and unfolding events, this is not computationally feasible for moderate to large sized proteins. We propose a new method to predict stability change upon mutation, using Monte Carlo unfolding simulations. We use our method to predict relative melting temperatures of all possible point mutants of *E. coli* Dihydrofolate Reductase (DHFR), an essential enzyme that is an important target of antibiotics and chemotherapeutic drugs. For several mutants that are predicted to be stabilized relative to WT, we experimentally determine melting temperatures and catalytic activities. We find a small but significant positive correlation between stability and activity. Our

approach allows us to identify several stabilized mutants, and we obtain good agreement with experiments ($r$ = 0.68), competitive with existing methods.

## *2.2. Methods*

Note: Experimental methods are described in (Tian et al., 2015), as this work was completed by other authors. Computational methods are described below.

### *2.2.1. Monte Carlo simulations*

Simulations were carried out using an all-non-hydrogen-atom Monte Carlo simulation program with a knowledge based potential, described in the Introduction section 1.5 of this thesis and in previous publications (Xu et al., 2011; Yang et al., 2007; Yang et al., 2008). Mutations were generated using Modeller v9.2 (Eswar et al., 2006). Next, an energy minimization was carried out in NAMD (Phillips et al., 2005) for 5,000 steps, using the default minimization algorithm and par_all27_prot_lipid.inp parameter file, without solvent. An additional minimization step was carried out by running the Monte Carlo simulation program at low temperature ($T$ = 0.100 in simulation units) for 2,000,000 steps. A 2,000,000-step simulation was then run at each of 32 temperatures, averaging over the final 1,000,000 steps and over 50 separate simulations to obtain Energy, RMSD, and number of contacts. Data was fit to a sigmoidal function to obtain the computationally-predicted melting temperature for each of Energy, RMSD, and number of contacts. Longer simulations of 20,000,000 steps were carried out on select mutants with 30 replications, averaging over the last 2,000,000 steps.

1,000,000 steps were completed in about one hour of simulation time on a single CPU.

### *2.2.2. Bioinformatics Analysis*

DHFR protein sequences from 290 bacterial species were aligned using the program MUSCLE and online server (Edgar, 2004). The MATLAB Bioinformatics Toolbox was used to create sequence logo representations and to determine the consensus sequence.

### *2.2.3. Simulation analysis*

Sigmoidal fits were performed using the module "Sigmoidal, 4PL" within the software program Prism 6. The sigmoid function has the form:

Y = Bottom + (Top-Bottom)/(1+10^((LogIC50-X)0HillSlope))

## *2.3. Results*

### *2.3.1. Theory justifying the use of non-equilibrium unfolding simulations to obtain equilibrium stability effects of mutations*

We assume two-state folding kinetics, as described in the Introduction section 1.1 and Figure 1.1. The time spent in the native state waiting for sufficient thermal fluctuation to cross the unfolding free energy barrier is given by:

$$\tau_u^{fp} = \tau_0 e^{\frac{\Delta G^{\#}}{kT}} \qquad \text{(Equation 2.1)}$$

where $\tau_u^{fp}$ is the first-passage time from the folded to the unfolded state, $\Delta G^{\#}$ is the free energy difference between the folded state and the transition state, and $\tau_0$ is the elementary time constant. Unfolding is observed when the simulation time $\tau_{sim}$ approaches $\tau_u^{fp}$. Therefore, the temperature at which unfolding events are observed in simulations depends on the simulation time according to:

$$kT_m^{app} = \frac{\Delta G^{\#}}{\ln\left(\frac{\tau_{sim}}{\tau_0}\right)} \qquad \text{(Equation 2.2)}$$

Next, we introduce a relative melting temperature:

$$\Delta T_m^{rel}(i) = \left(T_m^{app}(i) - T_m^{app}(WT)\right) / T_m^{app}(WT) \qquad \text{(Equation 2.3)}$$

and use Equations 2.2, 2.3, and 1.1 to obtain

$$\Delta T_m^{rel}(i) = \frac{(1-\varphi_i)\Delta\Delta G_i^{eq}}{\Delta G^{\#}} \qquad \text{(Equation 2.4)}$$

A recent study demonstrated that most mutations have approximately the same $\varphi$ value (Naganathan and Muñoz, 2010), where $\varphi_i \approx 0.24$. Therefore,

$$\Delta T_m^{rel}(i) = 0.76\frac{\Delta\Delta G_i^{eq}}{\Delta G^{\#}} \qquad \text{(Equation 2.5)}$$

i.e., the relative unfolding temperature is independent of simulation time and proportional to the equilibrium free energy effect of mutation, provided that simulations have equilibrated in the native basin.

*2.3.2. Monte Carlo unfolding simulations*

Unfolding simulations were performed on Dihydrofolate Reductase (PDB ID: 4DFR) using Monte Carlo simulations. A sample unfolding trajectory for the WT

protein is shown in Figure 2.1. Unfolding tends to proceed by a single pathway, which begins at the C-terminal hairpin.

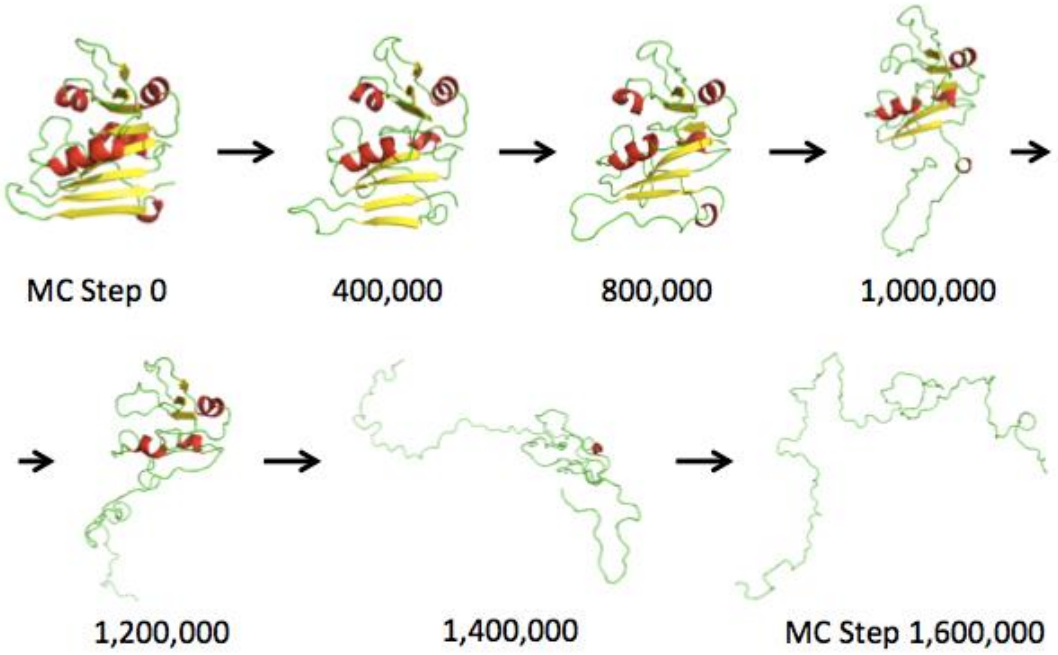



**Figure 2.1.** DHFR unfolding trajectory from MC simulations.

As a preliminary test of the effect of mutation on unfolding in Monte Carlo simulations, we carried out five simulations for WT and destabilized mutants I155A and W133V/I91L. Figures 2.2 – 2.4 show that destabilized mutants appear to unfold more rapidly than WT, with the double mutant unfolding the fastest. For mutants and for WT, unfolding begins at the C-terminal hairpin, which detaches from the rest of the protein prior to the major unfolding event encompassing the rest of the structure.

Figure 2.6 shows RMSD versus Monte Carlo step for the trajectories. While variation is seen among the trajectories, mutants appear to unfold before WT on

average. Table 2.1 shows estimated time of the first major unfolding event for each of the simulations. The mean time of the first unfolding event is found to be less for the double mutant than for WT ($p = 0.03$, T-test).

To obtain simulated melting curves, DHFR was first subjected to a brief MD and MC energy minimization, followed by unfolding simulation at each of 32 separate temperatures (see Methods section 2.2 for details). As expected, at higher temperatures the protein displayed higher RMSD from the initial structure, higher energy, and fewer contacts between atoms than at lower temperatures (Figure 2.7). Dependence on temperature was roughly sigmoidal, and $T_m$ was calculated by fitting to a sigmoidal function, for each of RMSD, energy, and number of contacts.

*2.3.3. Computational prediction of stabilizing mutations*

All 3,021 possible single point mutants of DHFR were simulated, and $T_m$ was calculated as described above, using each of three metrics: energy, RMSD from the folded structure, and number of contacts. The values calculated using energy, RMSD, and contacts were highly correlated, as shown in Figure 2.8. 523 mutations (17.3%) were predicted to be stabilizing by all three metrics, while 42.1% of mutations were predicted to be destabilizing by all three metrics. The distribution of predicted melting temperatures averaged over the three metrics for all 3021 point mutants is shown in Figure 2.9. The majority of mutations are predicted to be destabilizing, in agreement with published experimental data and FoldX predictions (Tokuriki et al., 2007; Zeldovich et al., 2007).

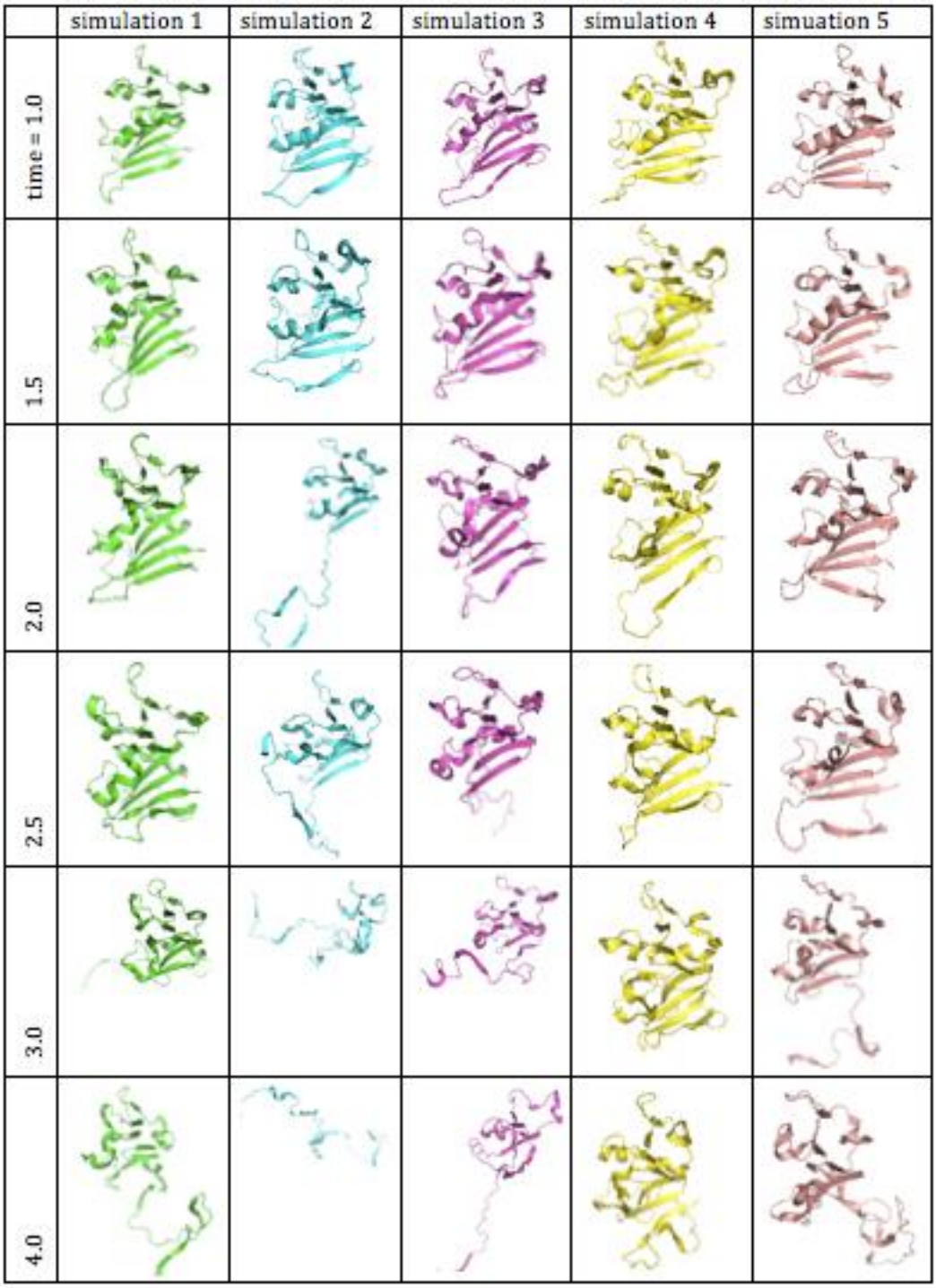



**Figure 2.3.** WT unfolding trajectories, $T$ = 1.3, steps in units of $10^6$.

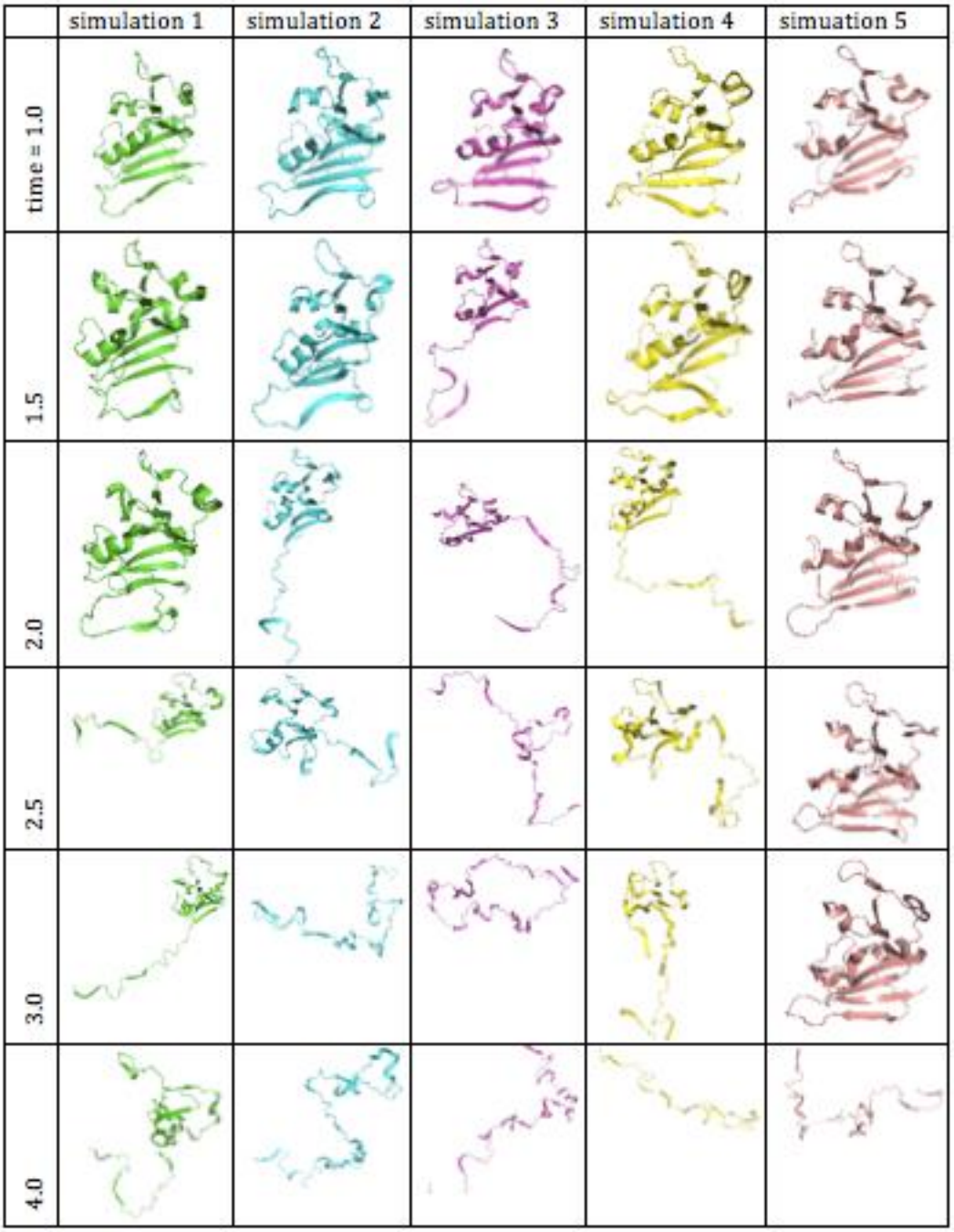



**Figure 2.4.** I155A unfolding trajectories, $T$ = 1.3, steps in units of $10^6$.

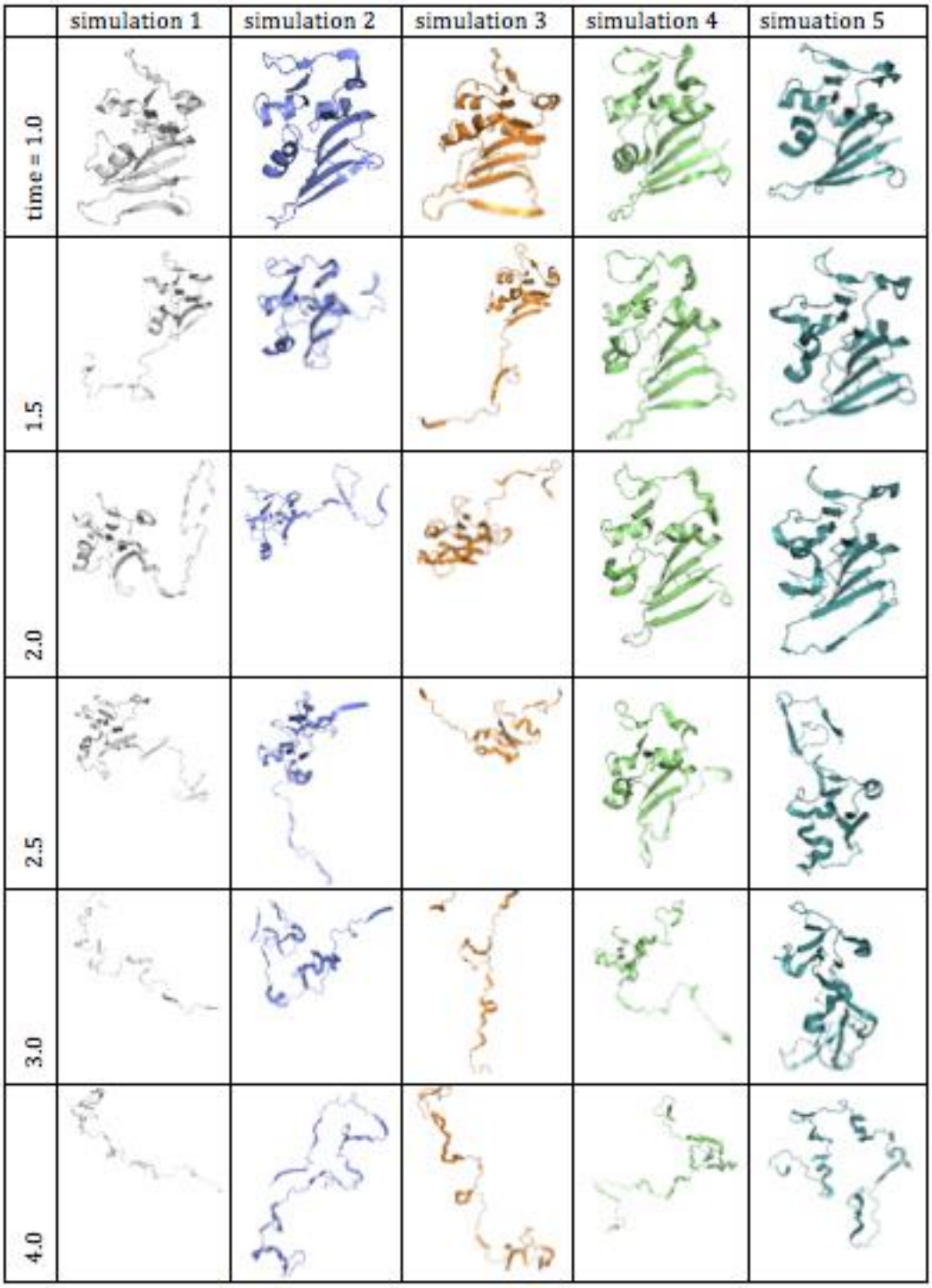



**Figure 2.5.** W133V/I91L unfolding trajectories, $T$ = 1.3, steps in units of $10^6$.

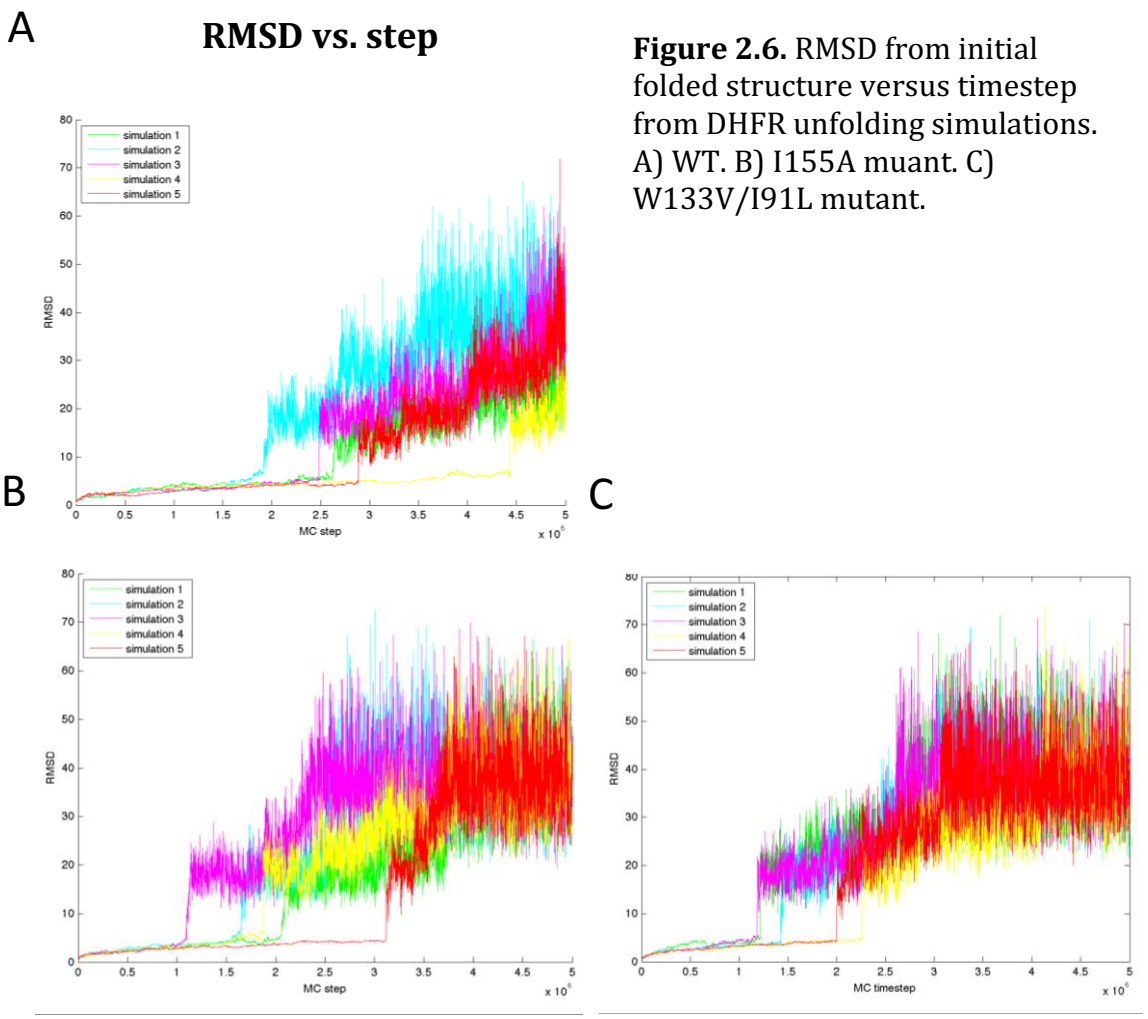



**Figure 2.6.** RMSD from initial folded structure versus timestep from DHFR unfolding simulations. A) WT. B) I155A muant. C) W133V/I91L mutant.

**Table 2.1.** Estimated time of first major unfolding event (units of 10^6 MC steps)

| simulation | 1 | 2 | 3 | 4 | 5 | mean < WT | mean < I155A |
|---|---|---|---|---|---|---|---|
| **WT** | 2.6 | 1.9 | 2.5 | 4.5 | 2.9 | | |
| **I155A** | 2.1 | 1.6 | 1.1 | 1.9 | 3.1 | p = 0.07 | |
| **W133V/I91L** | 1.2 | 1.5 | 1.2 | 2.8 | 2.0 | p = 0.03 | p = 0.32 |

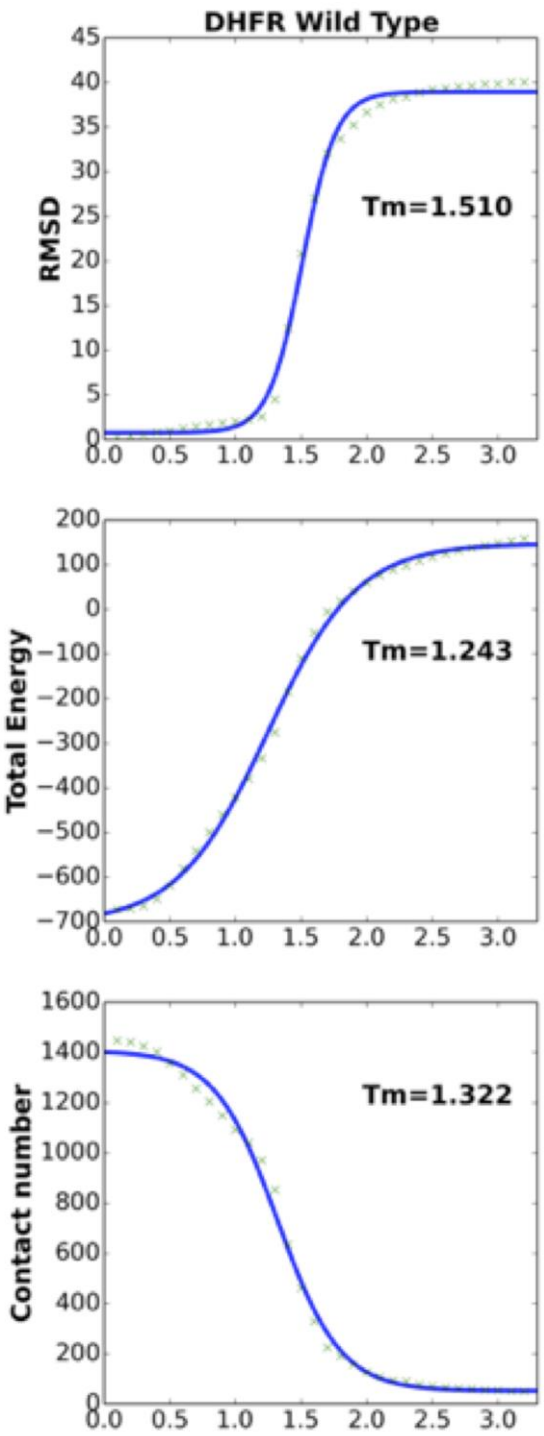



**Figure 2.7.** Plots of RMSD, total energy, and number of contacts as a function of temperature. Data is averaged over the final 1,000,000 steps of the 2,000,000 step simulation and over 50 separate runs. Green x's show data points, and the blue line shows the sigmoidal fit.

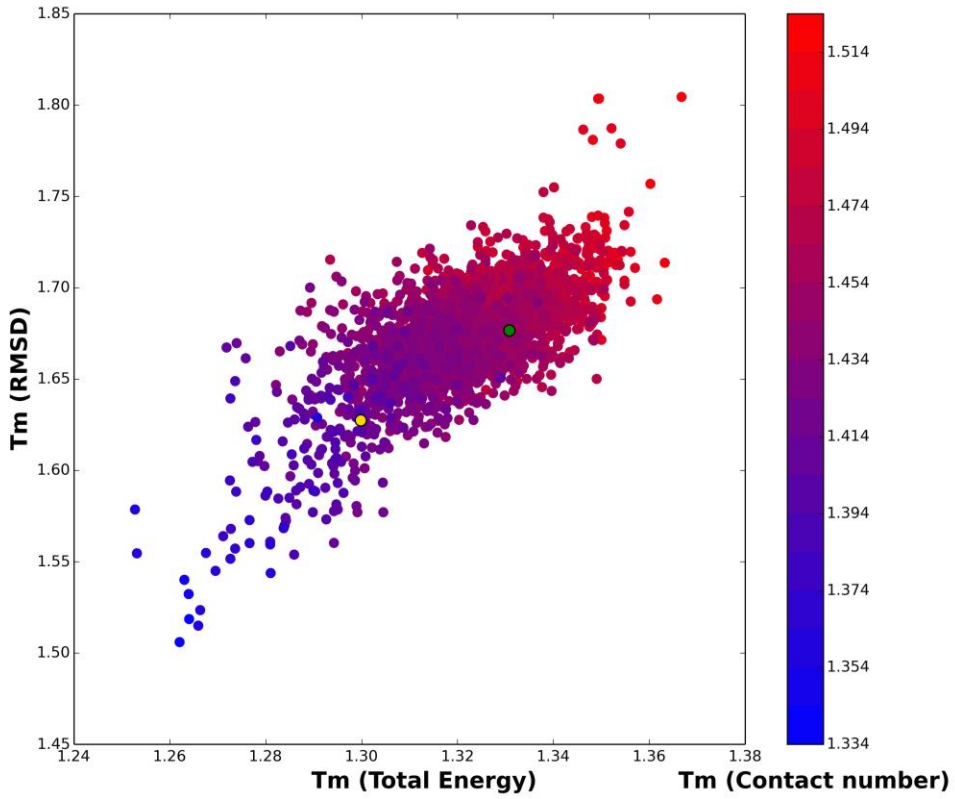



**Figure 2.8.** Scatter plot of Tm values determined by energy (x-axis), RMSD (y-axis) and number of contacts (see color bar to right of plot). The green ball shows WT and the gold ball, I155A. The correlation coefficients of simulated Tm between RMSD and total energy, RMSD and Contact number, and Contact number and total energy were 0.68, 0.79 and 0.84, respectively.

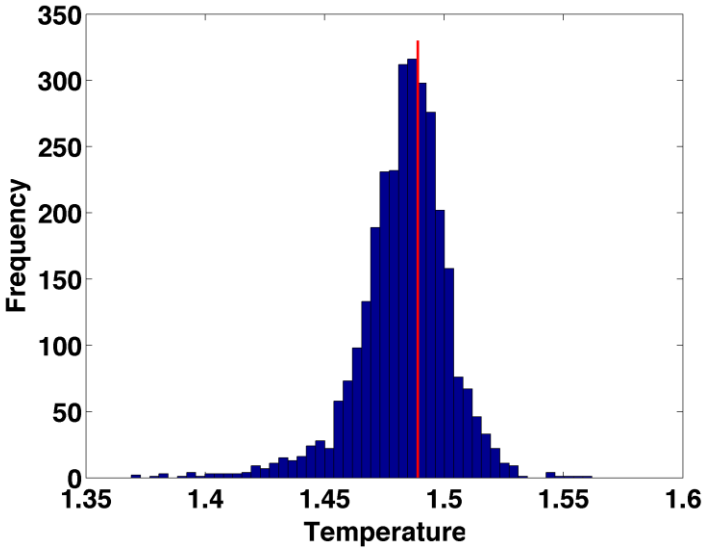



**Figure 2.9.** Histogram of $T_m$ values, determined by averaging $T_m$ values from energy, RMSD, and number of contacts. The vertical red line denotes WT $T_m$.

We selected a subset of the mutations that were predicted to be stabilizing by all three metrics for in depth computational and experimental analysis. To achieve this, we first selected residue positions at which multiple mutations were predicted to be stabilizing. We then selected one of the most stabilizing mutations at each of these residue positions. This process yielded 23 point mutations predicted to be highly stabilizing, shown in Table 2.2. In addition, five stabilizing mutations at different sites in DHFR, shown in Figure 2.10, were combined to form the five multiple mutants shown in Table 2.3, which were also subjected to further computational and experimental analysis. We reasoned that combination of stabilizing mutations could result in a further stabilized enzyme.

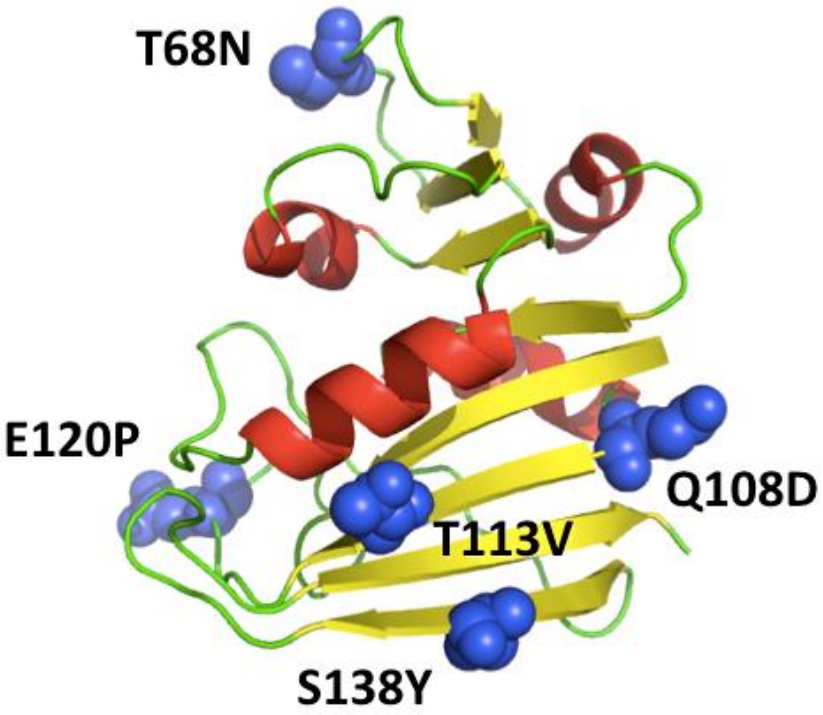



**Figure 2.10.** The structure of *E. coli* DHFR (PDB ID 4DFR), with residues that were altered in the stabilized quintuple mutant shown in blue.

**Table 2.2.** Simulated $T_m$ values of selected point mutants and WT DHFR.

| Mutations | $T_m$ (RMSD) | $T_m$ (Total Energy) | $T_m$ (Contact Number) | Average $T_m$ |
|---|---|---|---|---|
| WT | 1.507 | 1.243 | 1.323 | 1.358 |
| D27F | 1.525 | 1.261 | 1.351 | 1.379 |
| T113V | 1.551 | 1.259 | 1.358 | 1.389 |
| Q108D | 1.510 | 1.244 | 1.329 | 1.361 |
| S138Y | 1.518 | 1.247 | 1.333 | 1.366 |
| D116F | 1.525 | 1.248 | 1.334 | 1.369 |
| T68N | 1.516 | 1.250 | 1.336 | 1.367 |
| E120P | 1.519 | 1.257 | 1.337 | 1.371 |
| V119F | 1.519 | 1.252 | 1.344 | 1.372 |
| S135I | 1.527 | 1.248 | 1.340 | 1.371 |
| C152I | 1.534 | 1.253 | 1.345 | 1.377 |
| H114R | 1.506 | 1.249 | 1.342 | 1.366 |
| S49E | 1.509 | 1.260 | 1.340 | 1.370 |
| H141F | 1.536 | 1.264 | 1.351 | 1.384 |
| E157F | 1.536 | 1.268 | 1.352 | 1.385 |
| G15W | 1.513 | 1.261 | 1.342 | 1.372 |
| E154V | 1.607 | 1.273 | 1.372 | 1.417 |
| L156Y | 1.510 | 1.247 | 1.334 | 1.364 |
| E139V | 1.548 | 1.271 | 1.355 | 1.391 |
| D87P | 1.510 | 1.251 | 1.339 | 1.367 |
| G43P | 1.510 | 1.263 | 1.336 | 1.370 |
| W74F | 1.512 | 1.252 | 1.334 | 1.366 |
| G67H | 1.515 | 1.254 | 1.339 | 1.369 |
| A6I | 1.542 | 1.260 | 1.347 | 1.383 |

Note: The data were simulated with 50 replications, for a total of 2,000,000 MC. The last 1,000,000 steps were used to calculate $T_m$.

*2.3.3. Computational test of theoretical analysis*

We test computationally two predictions that emerge from the theoretical analysis of unfolding simulations. First, the apparent unfolding temperature should decrease as the length of the unfolding simulation increases. Second, the mutational change in relative apparent unfolding temperature (i.e., normalized to WT) should be robust with respect to simulation time, provided simulations have equilibrated in the native basin.

**Table 2.3.** Simulated and experimental results of selected mutants and WT DHFR.

| Mutations | $T_m$ (DSC) | $C_m$ (CD) | $k_{cat}$ | $k_{cat}/K_m$ | Simulated $T_m$ |
|---|---|---|---|---|---|
| WT | 54.1 | 3.09 | 24.60 | 14.07 | 1.358 |
| T113V | 58.0 | 3.28 | 13.67 | 10.86 | 1.389 |
| Q108D | 55.7 | 3.18 | 24.60 | 10.35 | 1.361 |
| S138Y | 55.6 | 3.33 | 24.51 | 9.33 | 1.366 |
| D116F | 55.5 | 3.43 | 24.80 | 9.53 | 1.369 |
| T68N | 55.5 | 3.26 | 29.36 | 13.32 | 1.367 |
| E120P | 55.3 | 3.25 | 30.02 | 13.91 | 1.371 |
| T68N,Q108D,T113V,E120P,S138Y | 61.3 | 3.52 | 32.63 | 12.20 | 1.400 |
| T113V,E120P,S138Y | 58.5 | 3.49 | 31.13 | 13.10 | 1.384 |
| T68N,Q108D,E120P,S138Y | 56.4 | 3.47 | 22.80 | 10.94 | 1.377 |
| T68N,Q108D | 55.8 | 3.14 | 17.99 | 15.24 | 1.366 |
| E120P,S138Y | 55.6 | 3.29 | 16.01 | 10.81 | 1.371 |

Note: The data were averaged over 50 replications. 2,000,000 MC steps were simulated in total, and the last 1,000,000 steps were used to calculate $T_m$.
Units: $T_m$: °C , $C_m$: M, $k_{cat}$: $s^{-1}$, $k_{cat}/K_M$ : $s^{-1}$ $\mu M^{-1}$

As a test of the theoretical predictions, we carried out two sets of simulations: 2,000,000 steps and 20,000,000 steps in length for the 23 predicted stabilizing mutants, 15 mutants studied previously by experiment, and the 5 stabilizing multiple mutants in Table 2.3. We compared the predicted absolute and relative simulated unfolding temperatures from these simulations. We found that, consistent with predictions, apparent unfolding temperature decreases with simulation time (Figure 2.11) while the relative unfolding temperature is remarkably independent of simulation time (Figure 2.12). In what follows, we use relative melting temperature when comparing simulation results with experimental results, where $T_m$ is averaged over that obtained using energy, number of contacts, and RMSD.

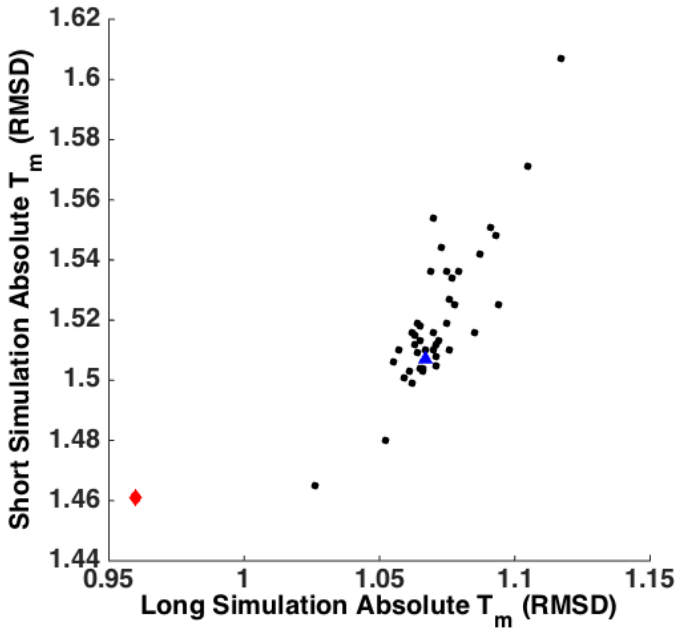



**Figure 2.11.** $T_m$ calculated from simulation RMSD, for short (2,000,000-step) and long (20,000,000-step) simulations. Simulation $T_m$ is smaller for long simulations, in which the protein has more time to unfold.

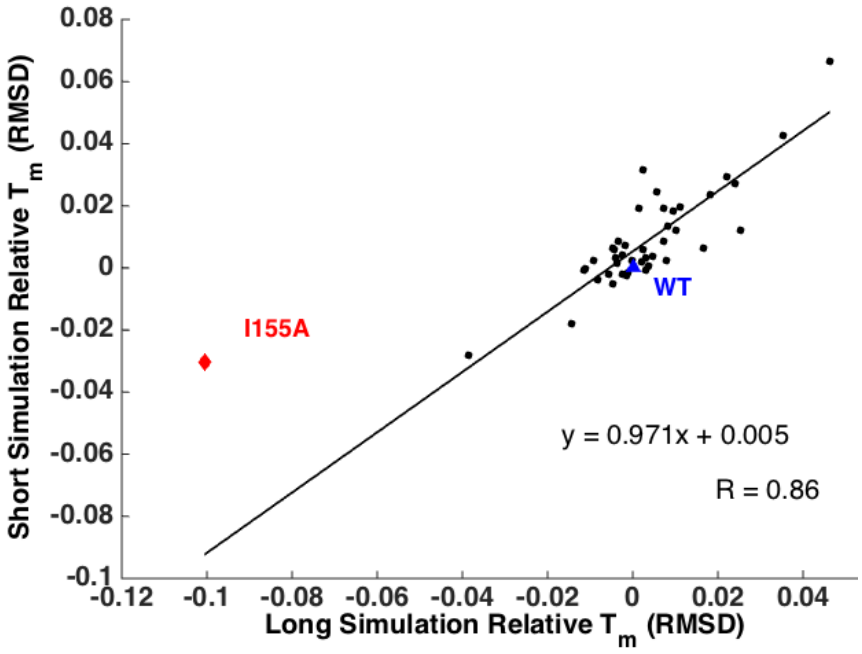



**Figure 2.12.** Relative $T_m$ (normalized to WT), for short and long simulations. Remarkably, the points fall nearly on the line y = x, with a correlation of 0.86, with the distinct outlier I155A.

*2.3.4. Experimental characterization of mutants*

Experimental details are given in (Tian et al., 2015). Briefly, we found that the mutant E154V was aggregation prone and excluded it from subsequent analysis. All other mutants were catalytically active except for D27F; D27 is a key catalytic

residue of DHFR (Ohmae et al., 2013). 10 of 22 predicted point mutations were found to be stabilizing according to $T_m$ values measured by DSC (Table 2.4). Since less than 18% of random mutations will be stabilizing (Tokuriki et al., 2007; Zeldovich et al., 2007), this result indicates that our method is effective in predicting stabilizing mutations (p = 0.002 under the null hypothesis that mutations are random). Predicted stabilizing mutations, with true stabilizing mutations colored orange, are shown in Figure 2.13.

**Table 2.4.** Experimental results for point mutants of DHFR

| Mutation | $T_m$(DSC) | $C_m$(CD) | $k_{cat}$ | $k_{cat}/K_m$ |
|---|---|---|---|---|
| WT | 54.1 | 3.09 | 24.60 | 14.07 |
| D27F | 61.7 | 4.55 | N.D. | N.D. |
| T113V | 58.0 | 3.28 | 13.67 | 10.86 |
| Q108D | 55.7 | 3.18 | 24.60 | 10.35 |
| S138Y | 55.6 | 3.33 | 24.51 | 9.33 |
| D116F | 55.5 | 3.43 | 24.80 | 9.53 |
| T68N | 55.5 | 3.26 | 29.36 | 13.32 |
| E120P | 55.3 | 3.25 | 30.02 | 13.91 |
| V119F | 54.9 | 3.12 | 28.50 | 12.57 |
| S135I | 54.8 | 3.33 | 33.35 | 16.66 |
| C152I | 54.2 | 3.15 | 22.99 | 11.44 |
| H114R | 54.1 | 3.07 | 28.31 | 14.06 |
| S49E | 53.5 | 2.89 | 10.55 | 5.24 |
| H141F | 53.0 | 2.94 | 12.07 | 6.00 |
| E157F | 52.4 | 3.07 | 29.07 | 14.45 |
| G15W | 52.3 | 3.07 | 10.55 | 5.24 |
| L156Y | 51.3 | 2.62 | 6.02 | 3.00 |
| E139V | 51.3 | 2.73 | 24.80 | 12.31 |
| D87P | 51.0 | 2.88 | 25.73 | 13.18 |
| G43P | 51.0 | 2.83 | 10.07 | 5.02 |
| W74F | 50.5 | 2.96 | 3.44 | 1.71 |
| G67H | 48.1 | 2.65 | 17.20 | 8.57 |
| A6I | 47.2 | 3.05 | 19.66 | 9.79 |

Units: $T_m$: °C , $C_m$: M, $k_{cat}$: $s^{-1}$, $k_{cat}/K_M$ : $s^{-1}$ $\mu M^{-1}$

Combination of single stabilizing mutations led to more stable multiple mutants, as predicted by simulation. In particular, $T_m$ of the quintuple mutant (T68N, Q108D, T113V, E120P, S138Y) was found to be 7.2°C higher than WT. All multiple mutants were catalytically active, and the quintuple mutant and triple mutant (T113V, E120P, S138Y) were found to be more catalytically active than WT. Figure 2.14 shows that the stability effects of mutations are less than additive.

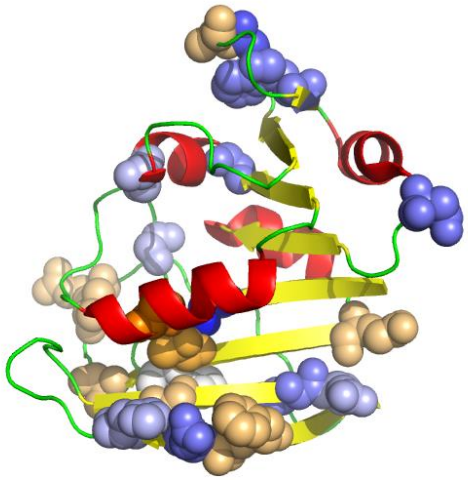


**Figure 2.13.** Locations of the 22 predicted stabilizing mutations (sphere representation), with stabilizing mutations colored orange and destabilizing mutations colored blue.

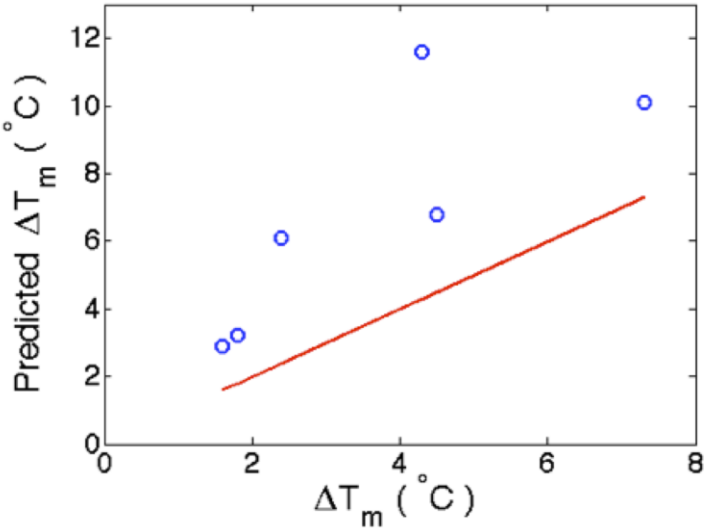



**Figure 2.14.** Change in experimental melting temperature relative to WT is predicted by summing melting temperature changes of individual mutants. This predicted $\Delta T_m$ is plotted relative to the observed $\Delta T_m$ (blue circles). r = 0.80, p = 0.06. Red line denotes predicted $\Delta T_m$ = observed $\Delta T_m$.

We compared computationally predicted and experimental unfolding temperatures. The correlation coefficient between experimental relative $T_m$ and simulated relative $T_m$ was 0.65 (Figure 2.15). Noting the theoretical prediction that simulated relative $T_m$ should be proportional to equilibrium stability, we plotted the relation between simulated relative $T_m$ and the equilibrium measurement of stability by urea denaturation (quantified by the mid-transition urea concentration, $C_m$). We observed a slightly higher correlation of $r$ = 0.68 (Figure 2.16).

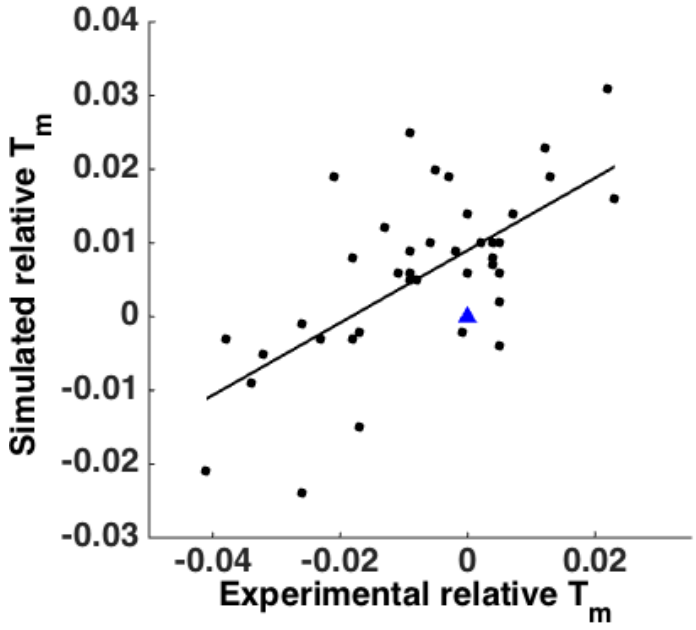



**Figure 2.15.** Correlation between simulated and experimental relative $T_m$ values. Values are from this study and from Bershtein et al. (Bershtein et al., 2012). Relative $T_m$ was calculated by normalizing to WT: ($T_m$(mutant)-$T_m$(wild type))/$T_m$(WT). WT is shown as a blue triangle. $r$ = 0.65, $p$ = 3 x $10^{-6}$.

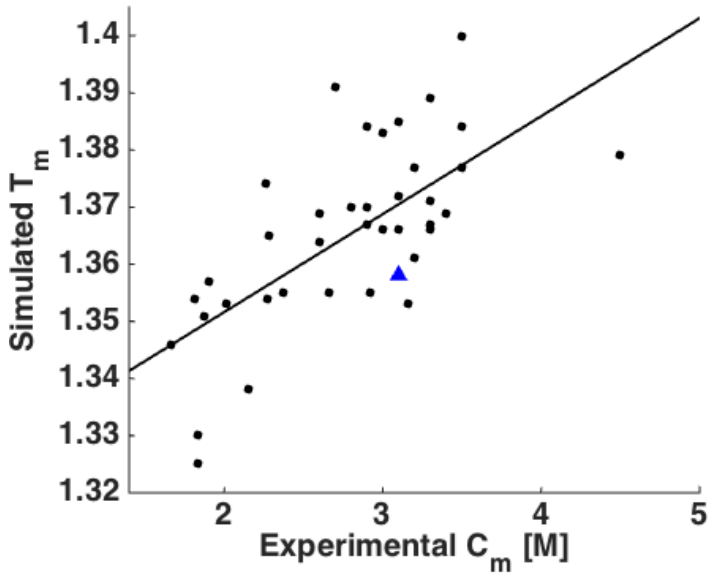



**Figure 2.16.** Correlation between simulated $T_m$ and experimental $C_m$. $r$ = 0.68. $p$ = 6 x $10^{-7}$.

We evaluated the effect of the number of replications and the number of MC steps on the performance of the method. Figure 2.17-A shows that prediction accuracy is sensitive to the number of replications, and we estimate that to achieve a reliable prediction of unfolding temperature, and least 20 replications should be used. We estimate from Figure 2.17-B that simulations should be run for at least 200,000 steps; this may allow time for simulations to equilibrate in the native basin. We note that after 200,000 steps, longer simulations do not yield increased accuracy.

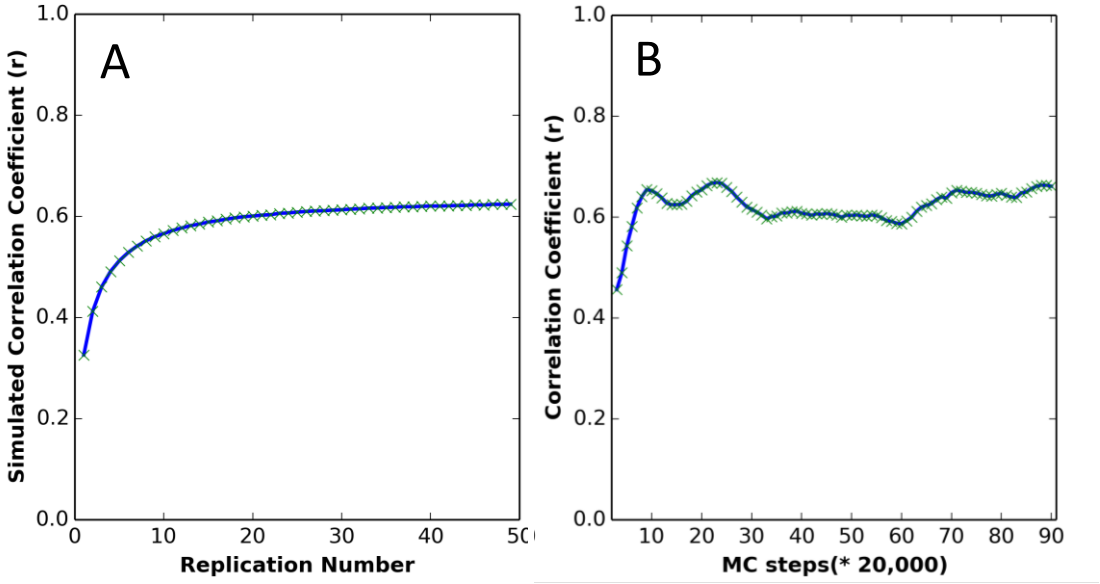
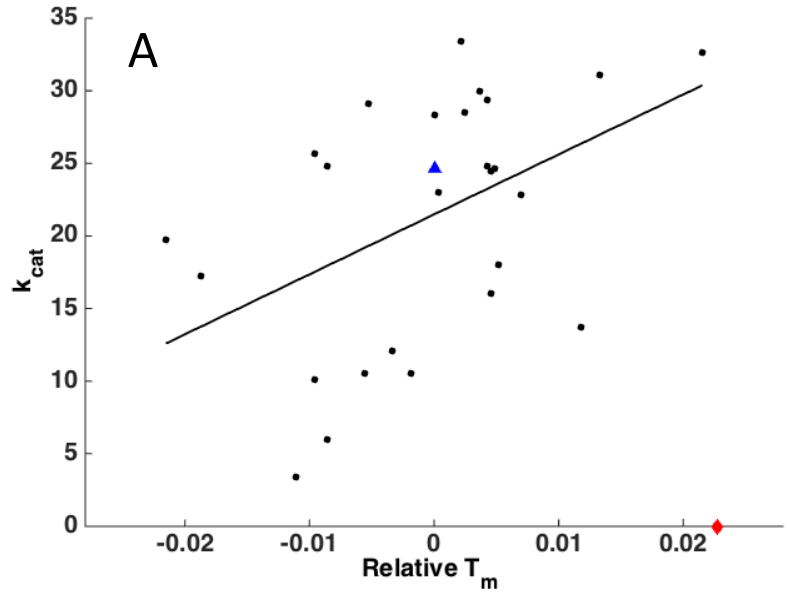



**Figure 2.17.** The effect of number of replications and number of MC steps on simulation accuracy. A) Correlation between simulate $T_m$ and experimental Tm, averaging over different numbers of replications. Each protein was simulated for 2,000,000 MC steps. B) Correlation between simulated $T_m$ and experimental Tm with 50 replications and different numbers of MC steps. Each protein was first simulated for the number of steps given on the x-axis, and the next 100,000 steps were averaged in determining the simulated Tm.

*2.3.5. Stability and activity do not trade off for DHFR*

It has been proposed that stability-activity tradeoffs prevent proteins from becoming overly stable (Beadle and Shoichet, 2002; DePristo et al., 2005; Studer et al., 2014). However, for DHFR we see a weak *positive* correlation between stability and activity, with the notable outlier D27F, where the mutation is made in the active site, rendering the protein inactive. Figure 2.18 shows experimental correlations between stability ($T_m$) and activity ($k_{cat}$ and $k_{cat}/K_M$).



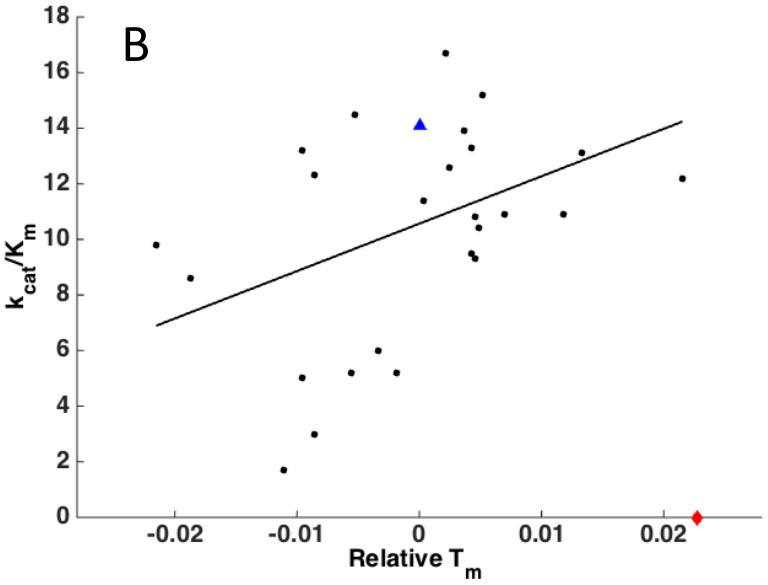




**Figure 2.18.** Correlation between DHFR stability and activity. WT is shown as a blue triangle, and D27F is shown as a red diamond at zero activity. A) Plot of $k_{cat}$ vs. experimental relative $T_m$. $r$ = 0.46, $p$ = 0.02 (excluding outlier D27F). B) Plot of $k_{cat}/K_m$ vs. experimental relative Tm. $r$ = 0.41, $p$ = 0.03 (excluding outlier D27F).

*2.3.6. Evolutionary analysis*

We determined the DHFR consensus sequence from an alignment of 290 bacterial DHFRs (Figure 2.19). In 4/16 of the experimentally stabilizing mutations, a

residue was changed to the consensus residue, while only 2/29 destabilizing mutations resulted from a change to consensus. Likewise, in 18/29 destabilizing mutations, a residue was changed away from the consensus residue, while this was true for only 5/16 of stabilizing mutations.

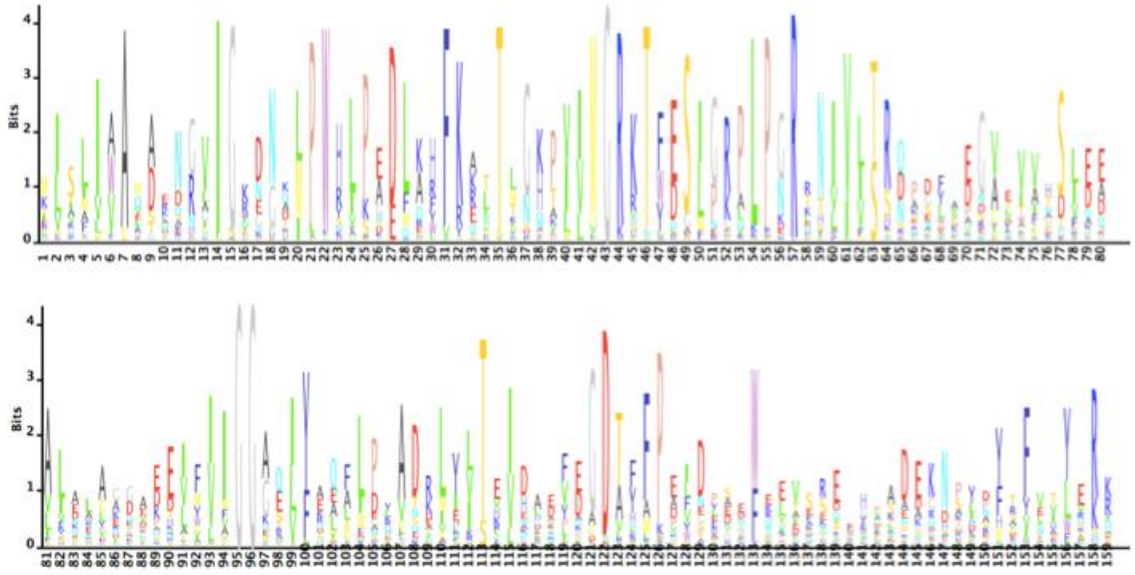



**Figure 2.19.** Sequence alignment and sequence entropy for 290 bacterial DHFRs.

*2.3.7. Simulated melting temperature by residue*

We compared the minimum and maximum simulated $T_m$ values obtainable by mutating a single residue to any other amino acid (Figure 2.20-A). There is a weak positive correlation between minimum and maximum melting temperatures. This might be expected, since, for instance, a residue that is already near its most stabilizing amino acid variant cannot be stabilized much further by mutation. Outliers in the plot correspond to the loci with the strongest stabilizing and destabilizing effects of mutations. Interestingly, these outliers tend to fall on the

interface connecting the C-terminal hairpin with the rest of the protein (Figure 2.20-B), which is the interface that is first to dissociate in simulations.

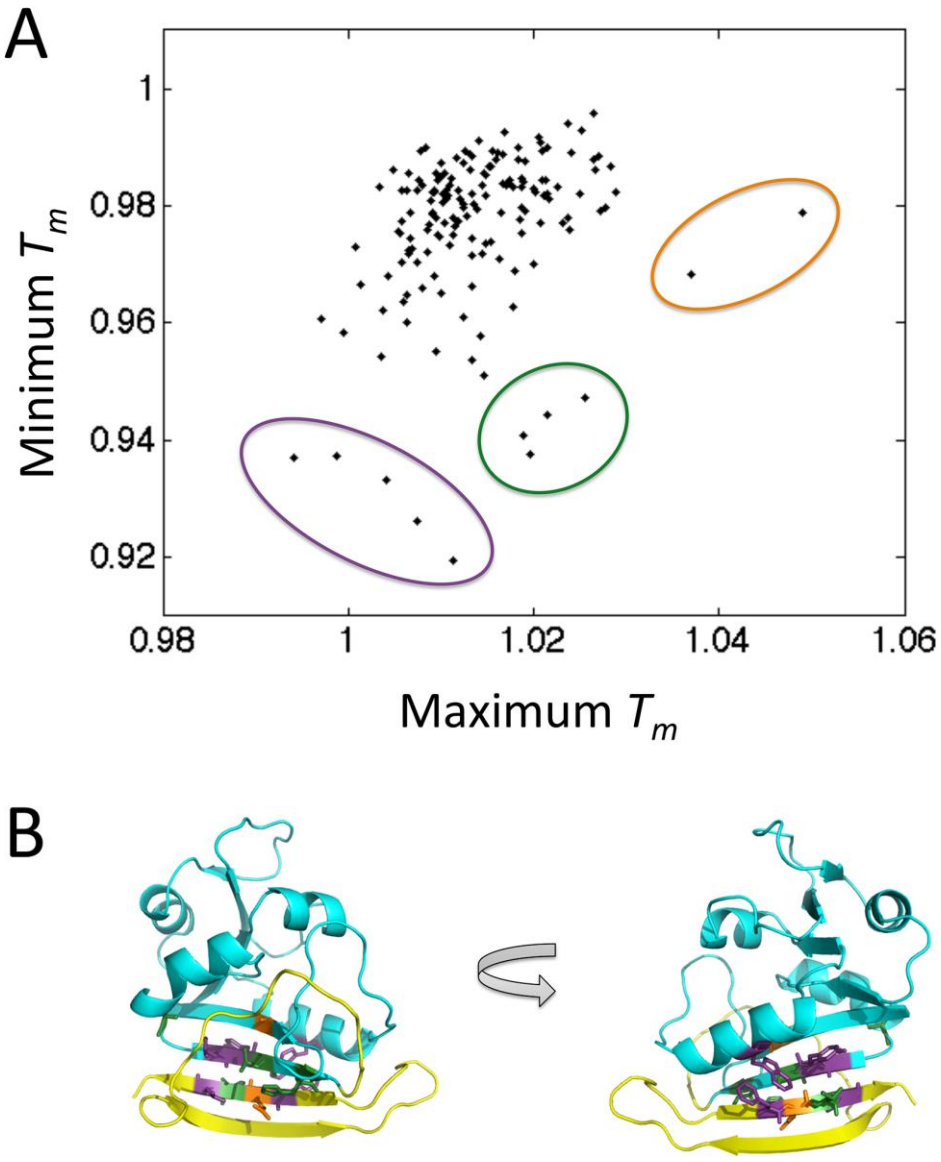



**Figure 2.20.** Maximum stabilization and destabilization induced by mutations at each residue position. A) Plot comparing the minimum and maximum simulated Tm values for each residue across all 19 simulated mutants. Tm is normalized to WT by dividing each $T_m$ value by the simulated WT value. Outliers are circled in purple (left), green (middle) and orange (right). B) DHFR with outlier residues colored according to the color scheme from A. Purple residues: F153, W30, Y111, L156, L110. Green: A107, I115, L112, H114. Orange: A6, E154. Excluding outlier residues, the C-terminal hairpin is colored yellow, and the rest of the protein is colored cyan.

*2.4. Discussion*

Ideally, estimates of protein stability could be obtained from long Molecular Dynamics (MD) simulations, allowing for multiple rounds of folding and unfolding of the protein. However, given the amount of computational resources required, this method would currently be prohibitive for all but the smallest proteins. We present a method for estimating the effect of a mutation on the equilibrium stability of a protein, using short unfolding simulations. While we use a Monte Carlo (MC) program to simulate unfolding, the method could be modified for use with MD simulations. It would be interesting to compare the accuracy of an MD-based approach with our MC-based results.

Protein stabilization can be achieved by slowing the rate of unfolding and/or accelerating the rate of folding. Our method based on unfolding simulations can be used to identify mutations that change the unfolding rate, which according to a recent study constitutes the majority of mutations (Naganathan and Muñoz, 2010), with $\varphi$ roughly constant around 0.24. These mutations are located at residue positions that are unfolded in the transition state, meaning that they are relatively early to unfold. In fact, many of the experimentally verified stabilizing mutations in DHFR predicted by our method are found in the C-terminal beta hairpin region, which is the first to unfold in simulations. Interestingly, the source of ultra-stability in hyperthermophiles generally arises from slowing the unfolding rate, rather than increasing the folding rate (Luke et al., 2007).

It has been hypothesized that a trade-off occurs between enzyme stability and activity due to a requirement that the enzyme must be sufficiently flexible to promote catalysis. This conclusion was based on exploration of the stability effects of mutations in the active sites of beta-lactamase, rubisco, and barnase (Beadle and Shoichet, 2002; DePristo et al., 2005; Studer et al., 2014). We observe a similar effect for the mutation D27F, which is located in the active site of DHFR and renders the protein inactive. Carving of an active site requires the specific selection of catalytic amino acids, which would be expected to have a destabilizing effect. However, we find that exploring only mutations in the active site provides a biased view of the relationship between stability and activity. Instead, we find that most mutations exhibit an opposite trend: a *positive* correlation between stability and activity. The lack of a global dynamics-driven tradeoff between stability and activity is an idea supported by other authors (Adamczyk et al., 2011; Bloom et al., 2006; Taverna and Goldstein, 2002).

A possible explanation for the *positive* correlation between stability and activity is that highly stable proteins have a greater effective concentration of protein available in the active form. We note that the correlation is only revealed when stabilizing mutants are included in the analysis; our earlier study (Bershtein et al., 2012) analyzed a smaller set of primarily destabilizing mutations and did not find any statistically significant trend between stability and activity for DHFR.

In conclusion, we developed a method to determine stability effects of mutations and to search for stabilized mutants using short Monte Carlo simulations. Our method shows good performance for the enzyme DHFR and for other proteins

(see (Tian et al., 2015)). We propose that this method could be useful as a technique for discovering the stability effects of mutations and for rationalizing these effects based on the protein structure and unfolding pathway.

*Contributions*

This work is described in a recent publication (Tian et al., 2015). Jaie Woodard carried out preliminary simulations of DHFR and several mutants. Jian Tian carried out simulations on all mutants and performed experiments. Anna Whitney ran longer simulations on selected mutants. Jaie Woodard and Jian Tian analyzed computational and experimental data. Eugene Shakhnovich proposed the theory relating simulated melting temperature and experimental stability. Jaie Woodard, Jian Tian, and Eugene Shakhnovich wrote the paper.

# 3. A Simple Model of Protein Domain Swapping in Crowded Environments

*Abstract*

Protein domain swapping is an intriguing structural phenomenon with possible relevance to protein aggregation and dimer evolution. However, the mechanism of domain swapping and its relevance within the crowded cellular environment are still not well understood. We propose a simple Monte Carlo model of domain swapping in two dimensions. The model allows for functional and non-functional interactions between proteins, for partial unfolding of the protein, and for motion in continuous space. We find that domain swapping occurs at intermediate temperatures, and that torsional strain can promote domain swapping, consistent with experimental observations. In addition, we predict that non-specific interactions between unfolded proteins occur at intermediate temperature and high concentration, consistent with the Flory theorem for polymer chains. For folded proteins, we predict that functional interactions are strongest at intermediate temperature, while non-specific interactions become more common at low temperatures, matching previous computational results for 3D lattice proteins.

## *3.1. Introduction*

Domain swapping is a type of protein-protein interaction that requires at least partial unfolding of the protein. In this way, domain swapping is similar to

most forms of protein aggregation, and in fact some types of aggregates may be formed through the process of domain swapping (Rousseau et al., 2003). In a domain-swapped structure, a structural element such as a beta strand, an alpha helix, or an entire protein domain is exchanged between two proteins, such that each reconstituted monomer contains contributions from two separate chains (Bennett et al., 1995; Gronenborn, 2009; Liu and Eisenberg, 2002). An example is shown in Figure 1.4 in the Introduction section. The protein segment separating the two "domains" is called the hinge loop. This region of the protein changes conformation in the transition between the closed native state and the open form required for domain swapping.

Protein sequences and cellular abundances must evolve to promote folding and functional interactions while avoiding non-specific interactions. For instance, proteins that are highly abundant tend to have less sticky surfaces (Levy et al., 2012), due to increased pressure to prevent non-specific interactions. Protein sequences also tend to evolve to avoid aggregation. However, the details of these processes have not yet been fully elucidated.

Lattice models and other coarse-grained models of proteins have been used extensively to study protein folding and protein-protein interactions (Abeln and Frenkel, 2008; Deeds et al., 2007; Ding et al., 2002; Lobkovsky et al., 2010; Mirny and Shakhnovich, 2001; Sali et al., 1994; Shakhnovich and Gutin, 1993). These simplified models allow for a greater sampling of the accessible conformational space within a given amount of computation time. In addition to simplifying the protein representation, many models contain fewer than the natural 20 amino acid

types (Li et al., 2008; Shakhnovich and Gutin, 1993; Straub and Thirumalai, 2011), reducing the number of possible sequences.

Here, we present a simple model of domain swapping in crowded environments. In this model, two-dimensional proteins are allowed to "unfold" partially by rotation of each the two domains about a hinge. With four residue types and one bead per residue, the model is designed to be a minimalistic model with the potential to reproduce the temperature dependence and sequence specificity of the domain swap interaction while allowing for specific and non-specific interactions between proteins. We find that domain swapping is promoted by a strong domain-domain interaction combined with torsional strain favoring the open conformation. Domain swapping occurs at intermediate temperature and intermediate concentration, consistent with experiment, while non-specific interactions between unfolded proteins occur at intermediate temperature and high concentration. For folded proteins, functional interactions are most common at intermediate temperature, while promiscuous interactions become more common at low temperatures.

## *3.2. Methods*

### *3.2.1. Model*

Figure 3.1-A shows a single model protein in the folded state. The protein consists of two domains (residues 1-6 and 7-12). Each domain can individually rotate about the hinge, shown as a black +. Note that there is not a residue at the hinge position. A functional dimerization interface is defined as the four-residue

surface opposite the hinge (residues 3, 6, 9, and 12). The interaction potential is a step function centered at each residue, with a hard-sphere radius of 0.75 units and an interaction radius of 1.80 units. The spacing between residues is 2.0 units, as shown in Figure 3. 1-B.

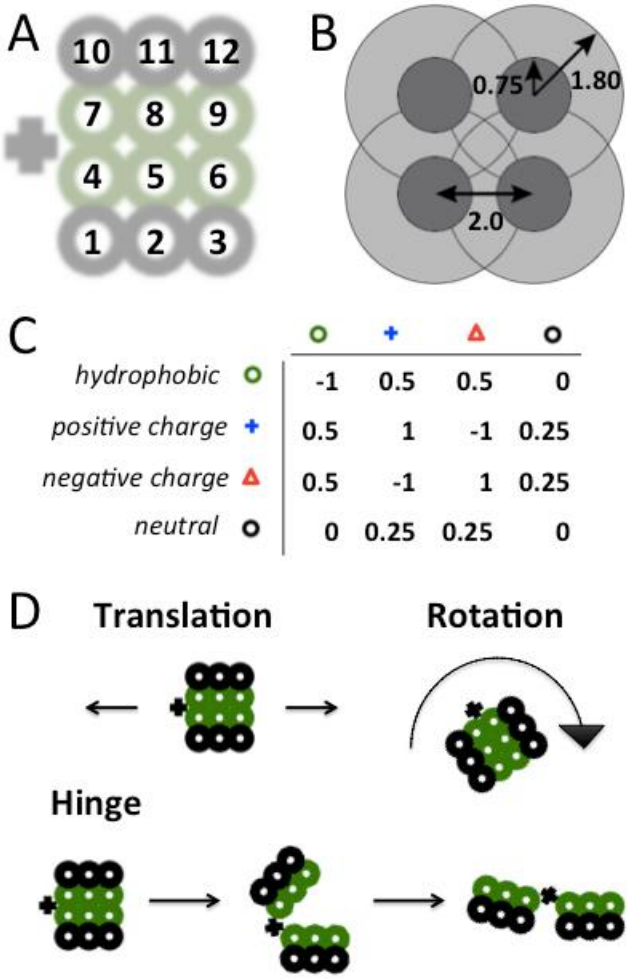



**Figure 3.1.** Model definition. A) A single protein in the folded state. Residues are numbered 1-12. The hinge is shown as a black +. The hinge position does not contain a residue. B) Interaction radii. C) Matrix of interaction energies between contacting residues. D) The three move types in the Monte Carlo move set: translation in any direction in two dimensions, clockwise or counter-clockwise rotation of the whole protein, and clockwise or counter-clockwise rotation of a single domain.

There are four residue types: hydrophobic, positively charged, negatively charged, and neutral. The interaction energy matrix is shown in Figure 3.1-C. Opposite charges attract, like charges repel, and hydrophobic residues attract. Units

of energy are defined such that the magnitude of these interactions equals 1. Hydrophobic and neutral residues repel charged residues by a smaller amount, reflecting phase separation. The electrostatic interaction is short-range, to represent screening by salt. Solvent is not explicitly included in this model.

An additional energy term biases the two domains toward an open conformation. This bias reflects torsional strange within the hinge loop, which is present in many domain-swapping proteins. In our model, the energy is proportional to the angle between domains, where lowest energy occurs at 180° (the open, domain-swap-prone state).

To represent the crowded cellular environment, which contains densely interacting proteins, multiple proteins are simulated at a range of concentrations, within a square cell. Periodic boundary conditions are employed. At the start of simulations, proteins begin in the folded state, evenly spaced within the cell. Protein concentration is adjusted by varying the cell size.

*3.2.2. Move set*

Monte Carlo simulations are carried out on model proteins moving in two-dimensions. Three possible moves are allowed: translation of the protein in a random direction, rotation of the protein either clockwise or counter-clockwise, and conformational change by rotation of a single domain about the hinge (Figure 3.1-D). An additional move is included to allow two proteins to translate or rotate simultaneously (Deeds et al., 2007). The magnitude of the move is chosen randomly, according to a Gaussian distribution centered at 0 and with standard deviation 0.5

for translation, 0.3 for rotation of the full protein, and 0.2 for rotation of a domain about the hinge. The probabilities of each move are 0.2 for translation, 0.2 for rotation, and 0.6 for rotation of a domain about the hinge. If two proteins are interacting and a translation or rotation move is rejected for one of the proteins (i.e., the dimer does not dissociate), then there is a probably of 0.5 of attempting a two-protein move. Moves are accepted or rejected according to the Metropolis criterion:

$$P_{accept} = \min\left(e^{-\frac{\Delta E}{kT}}, 1\right) \qquad \text{(Eq. 3.1)}$$

where $\Delta E$ is the change in energy as a result of the proposed move, and a move is always rejected if interaction between inner radii occurs.

*3.2.3. Categories of protein-protein interaction*

Numbers of folded monomers, unfolded monomers, folded functional dimers, domain-swapped dimers, unfolded proteins involved in non-specific interactions, and folded proteins involved in non-specific interactions were tracked over the course of the simulation. Figure 2 shows representative structures for each of these categories. The folded monomer contains interactions between residues 4 and 7, 5 and 8, and 6 and 9, referring to the numbering shown in Figure 3.1-A. The unfolded monomer lacks at least one pair of folded-state interactions. We refer to this as the unfolded state, although in a fully unfolded biological protein individual domains would also unfold. The folded functional dimer contains two proteins in the folded state, with the interfaces opposite the hinge in contact, so that residue 9 of one protein contacts residue 6 of the other protein. The domain-swapped dimer

incorporates the same contacts as two folded monomers, but with domains exchanged between proteins (all six contacts must be present). Non-specific dimers contain at least four contacts between proteins but do not fall into the functional dimer or domain swapped categories.

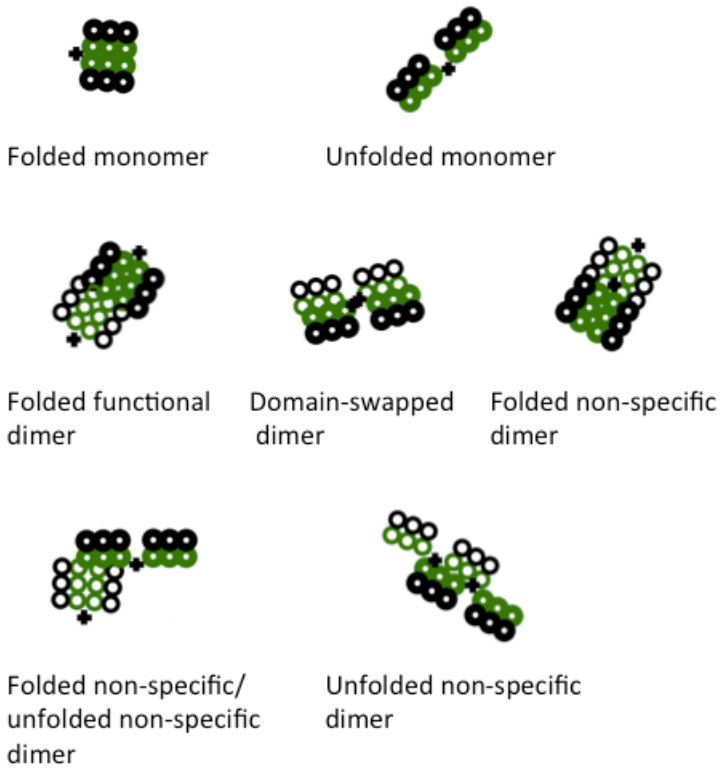



**Figure 3.2.** Categories of protein folding and interaction.

The total number of proteins involved in each type of interaction was tabulated. For instance, an interaction between a folded protein and an unfolded protein would count as one non-specific folded interaction and one non-specific unfolded interaction. An unfolded protein bound to a functional dimer would count as two functional dimer interactions and one non-specific unfolded interaction. The average number of proteins in each state was computed to generate 2D histograms.

A smoothing function was applied to each histogram, and histograms were combined to form phase diagrams.

*3.2.4. Sequence selection*

Six sequences were chosen for simulation and analysis (Figure 3.3, residue types defined in Figure 3.1-C). These sequences were chosen to span a range of protein stabilities and protein-protein interaction propensities. Sequences 0, 1, 2, and 3 contain hydrophobic residues at the domain-domain interface. Sequence 0 contains neutral residues elsewhere, so that the protein surface is partly neutral and partly hydrophobic. Sequence 1 contains a hydrophobic residue at the center of each 3-residue surface, making the surface more hydrophobic. Sequence 2 contains four hydrophobic residues at the functional interface. Sequence 3 contains charged residues along the 3-residue surfaces, allowing for specific interactions between charges. Sequence 4 is similar to sequence 0, but with one neutral residue at the domain-domain interface, destabilizing the protein. Sequence 5 is like sequence 2, but with charged residues added outside of the functional interface, which also destabilizes the protein. The energy difference between folded and unfolded states (without the additional energy term biasing towards the unfolded state) is -7 for proteins 0, 1, 2, and 3; -5 for protein 4, and -4 for protein 5.


**Figure 3.3.** The six protein sequences analyzed in this study. Residue types are defined in Figure 3.1 C.

*3.2.5. Simulation protocol*

Monte Carlo simulations were performed, starting from a square grid of 16 equally spaced folded proteins. Periodic boundary conditions were employed, and concentration was adjusted by varying the cell length from 80 to 320 units. 2,000,000 Monte Carlo steps were attempted per run, and statistics were averaged over the last 200,000 steps. The temperature ranged from kT = 0.2 to 2.0, in increments of 0.1 (we did not attempt a mapping of our simulation units to real temperatures). Simulations were carried out with and without a hinge energy biasing the protein toward the unfolded state. The biasing term had a magnitude of 2 times the angle between domains, in radians. Results were averaged over 20 separate runs.

*3.2.6. Energy diagrams*

Plots of energy versus hinge angle for singe proteins were generated by sampling the angle at increments of 0.01 radians and calculating: (energy between domains) + (hinge energy). Plots of folded fraction versus kT were generated by calculating $e^{-E/kT}$ at each angle and then calculating the sum over folded states divided by the sum over all states.

*3.2.7. Code*

The complete code for our model can be found on the Shakhnovich group's website: http://faculty.chemistry.harvard.edu/shakhnovich/software.

Additional analysis was performed in MATLAB (The MathWorks, Natick, MA). A smoothing function was applied to 2D histograms for phase diagrams and energy diagrams using gridfit.m by John D'Errico (available on the MATLAB Central File Exchange, http://www.mathworks.com/matlabcentral/fileexchange/), using a smoothness of 5.

## *3.3. Results*

### *3.3.1. Simulation trajectories*

In simulation trajectories, proteins begin in the folded monomeric state. As the simulation proceeds, proteins unfold, refold, and form interactions with other proteins. The total energy versus simulation step is shown in Figure 3.4, for a sample trajectory. We note that energy equilibrates by the end of the simulation, although there is still oscillation about the equilibrated value. Individual simulation frames are shown in Figure 3.5-A. The number of proteins in each interaction category as a function of Monte Carlo step are shown in Figure 3.5-B. Equilibrium between folded and unfolded monomers is established in the first 500,000 steps. Non-specific interactions involving folded and partially unfolded proteins appear early in the trajectory, while domain-swapped dimers appear later in the trajectory, becoming common after about 500,000 steps.

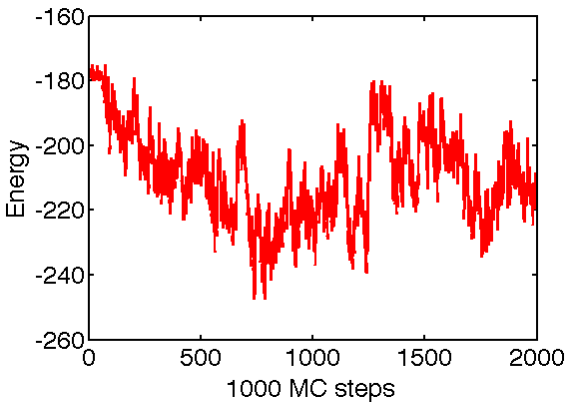



**Figure 3.4.** Total simulation energy vs. Monte Carlo step from a single sample trajectory. Sequence: 0, hinge energy = -2 times the angle between domains, cell length = 113 units, kT = 0.7.

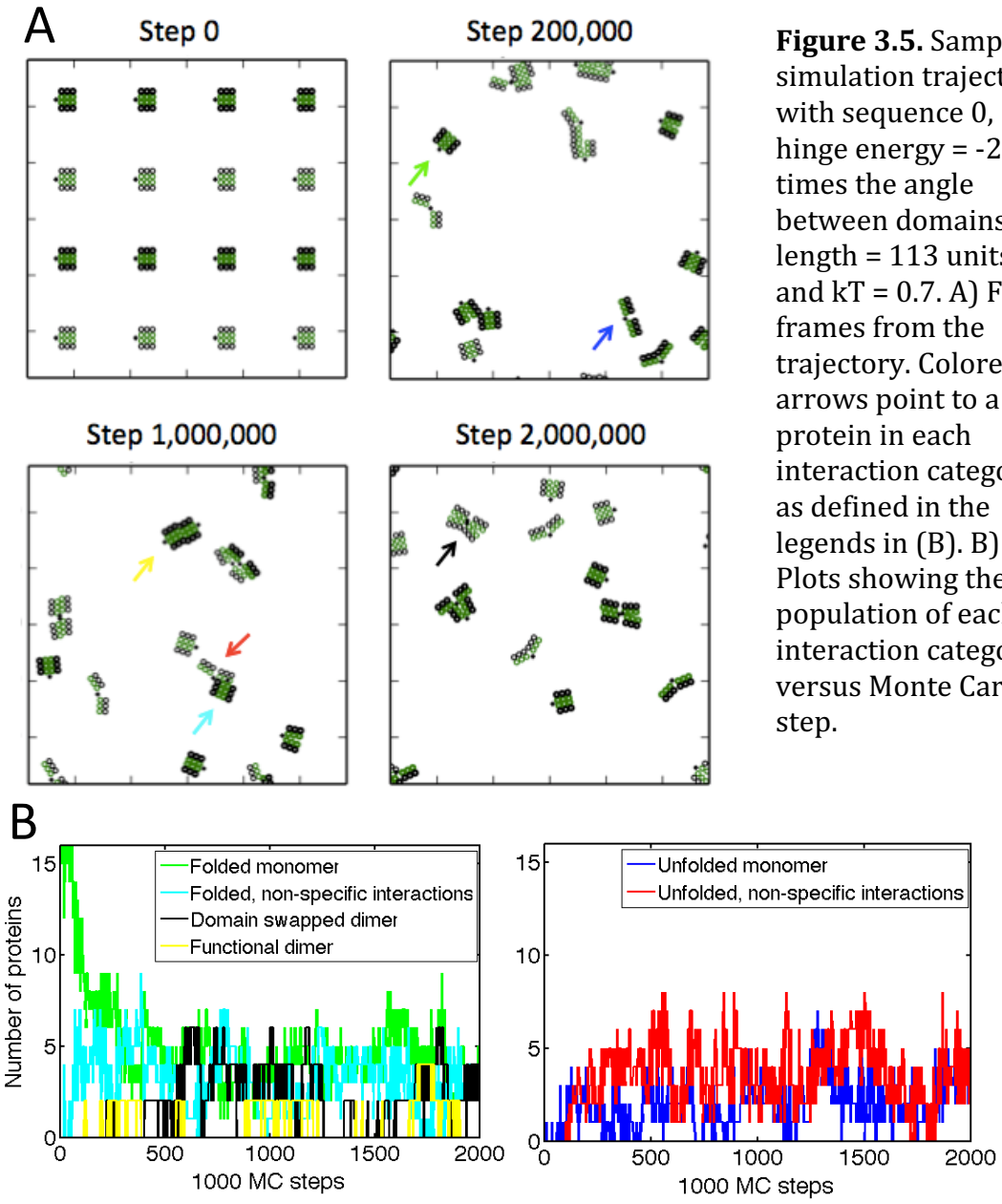



**Figure 3.5.** Sample simulation trajectory, with sequence 0, hinge energy = -2 times the angle between domains, cell length = 113 units, and kT = 0.7. A) Four frames from the trajectory. Colored arrows point to a protein in each interaction category, as defined in the legends in (B). B) Plots showing the population of each interaction category versus Monte Carlo step.

*3.3.2. Temperature and concentration dependence of oligomeric state*

Statistics from trajectories were averaged over the last 200,000 steps and over 20 runs. Figure 3.6 shows results as a function of temperature for sequence 0. At high concentration and low temperature, most of the proteins are in folded dimeric states. As temperature increases, proteins dissociate into monomers and begin to unfold. A sample frame from simulations at high temperature and high concentration is shown in Figure 3.7-A. At lower concentrations, dimers dissociate more abruptly with increasing temperature, and protein-protein interactions are not seen at high temperatures.

Hinge strain causes unfolding to occur at lower temperatures. In the presence of hinge strain, domain swapping occurs at intermediate temperature. Domain swapping is most prevalent at intermediate concentrations, while non-specific interactions involving unfolded proteins increase at high concentrations. The domain-swap interaction falls off more rapidly with increasing temperature than the non-specific unfolded interactions, most likely due to the low entropy of the domain swapped state relative to the non-specific unfolded state.

Results as a function of temperature for sequence 1 are shown in Figure 3.8. As expected, non-specific interactions between folded proteins are more common than for sequence 0. The domain swap interaction at high concentration is also less for this protein, while non-specific interactions involving unfolded proteins are increased. As shown in Figure 3.7-B, non-specific interactions include a variety of interaction types involving both the protein surface and domain-domain interface residues. Figure 3.9 shows results for sequence 2. Specific interactions are common

at low to intermediate temperatures. For proteins without hinge energy at intermediate concentration (Figure 3.9 C), the amount of functional dimer depends non-monotonically on temperature. At intermediate temperatures, functional dimers are the most common species, while at low temperatures non-specific interactions between folded proteins become more common. As for sequence 1, for proteins with hinge energy at high concentration (Figure 3.9 B), there are few domain swapped dimers and many non-specific interactions involving unfolded proteins.

Results for sequence 3 are shown in Figure 3.10. Relative to sequence 0, for proteins without hinge energy at high concentration (Figure 3.10-A), there are fewer protein-protein interactions. At intermediate and high concentrations (Figure 3.10-C, E), non-specific interactions between folded proteins are more common than functional interactions at low temperature, while functional interactions become more common at intermediate temperatures. Functional interactions are less common than for sequence 2 (note that residues interact on the diagonal, so the functional interface is not as strong). Domain swapping behavior is similar to sequence 0.

Results for sequences 4 and 5 are shown in Figures 3.11 and 3.12, respectively. Domain swapping occurs for both proteins at lower temperatures than for sequence 0, and sequence 4 exhibits more domain swapping than sequence 5. However, for both sequences, non-specific interactions between unfolded proteins are more common than domain swapping. For sequence 4 without hinge energy, functional interactions are less common at low temperature relative to sequence 0,

since the functional interface is weakened. Also, as expected, proteins unfold at a lower temperature, since the domain-domain interface is weakened. Sequence 5 exhibits even more functional interaction at low temperature than sequence 2, since the sequence is designed to inhibit non-functional interaction between folded protiens.

Plots of energy versus angle between domains were constructed to better understand the temperature dependence of folding and oligomerization state (Figure 3.13). Note that the unfolded state (shown in black in Figure 3.13-A) is more entropically favored than the folded state (shown in red), since most angles correspond to the unfolded state. However, for all sequences with hinge energy equal to zero and for sequences 0, 1, and 2 with the hinge energy bias, the folded state is lower in energy. Therefore, for these proteins, the folded state will be favored at low temperature, while the unfolded state will become more common at high temperature (Figure 3.13-B, C). This is in fact what we see in simulations, for these sequences. Boltzmann weighting predicts that for sequences 3, 4, and 5, with hinge energy biasing towards the open state, the unfolded state will actually be more populated at low temperature. However, due to the fact that simulations begin with proteins in the folded state, kinetics and possibly interactions between folded proteins cause the folded state to be preferred at low temperature in simulations for these proteins as well.

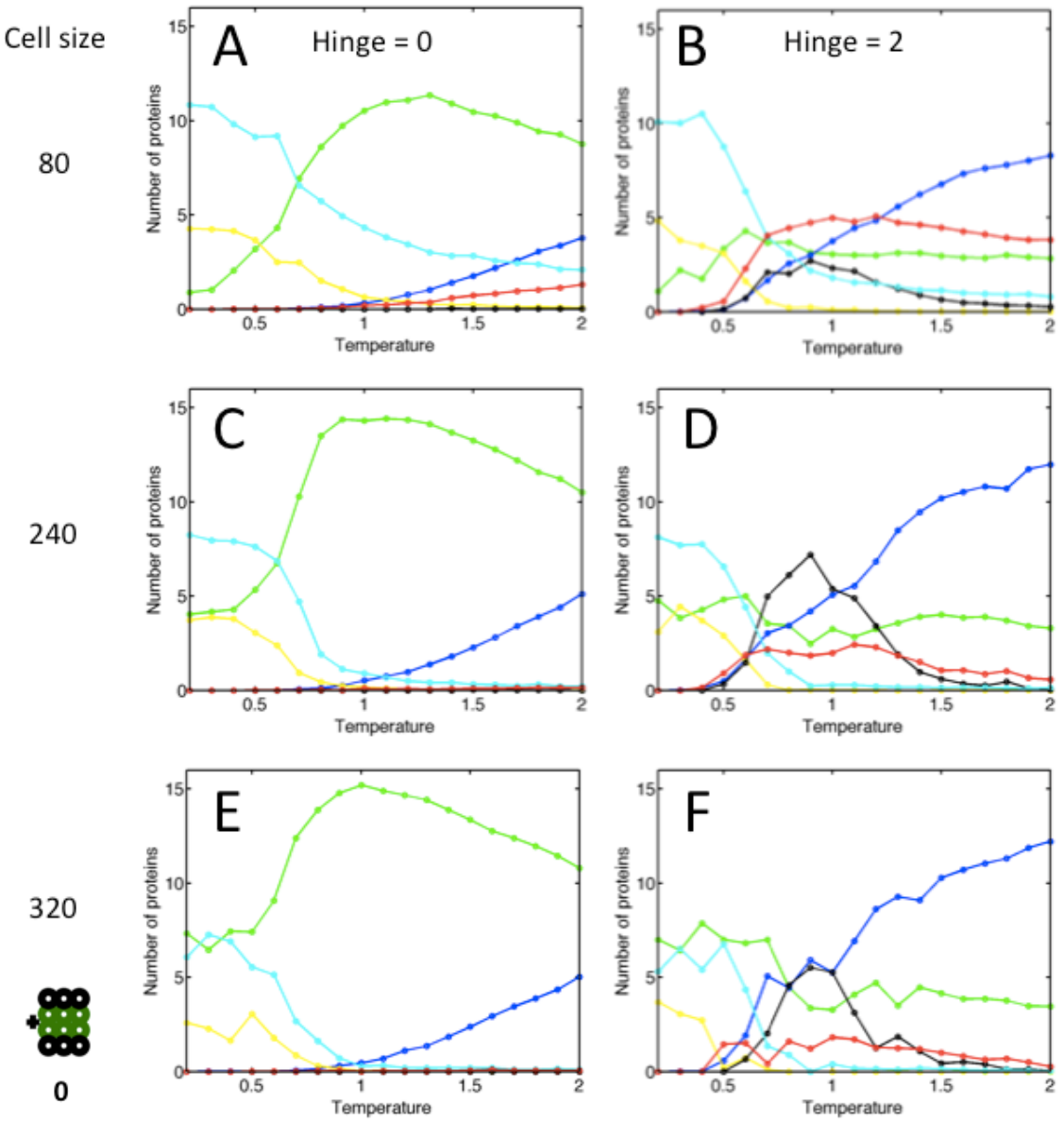



**Figure 3.6.** Simulation statistics as a function of temperature for sequence 0. Colors represent interaction categories, as defined in Figure 3.5: green: folded monomers, cyan: folded proteins exhibiting non-specific interactions, black: domain swapped dimers, yellow: functional dimers, blue: unfolded monomers, red: unfolded proteins exhibiting non-specific interactions. Results are averaged over the final 200,000 frames of a 2,000,000 step simulation and over 20 individual runs. Cell size = 80 units for (A,B), 240 units for (C,D), and 320 units for (E,F). Hing energy = 0 for (A, C, E) and 2 times the angle between domains for (B, D, F).

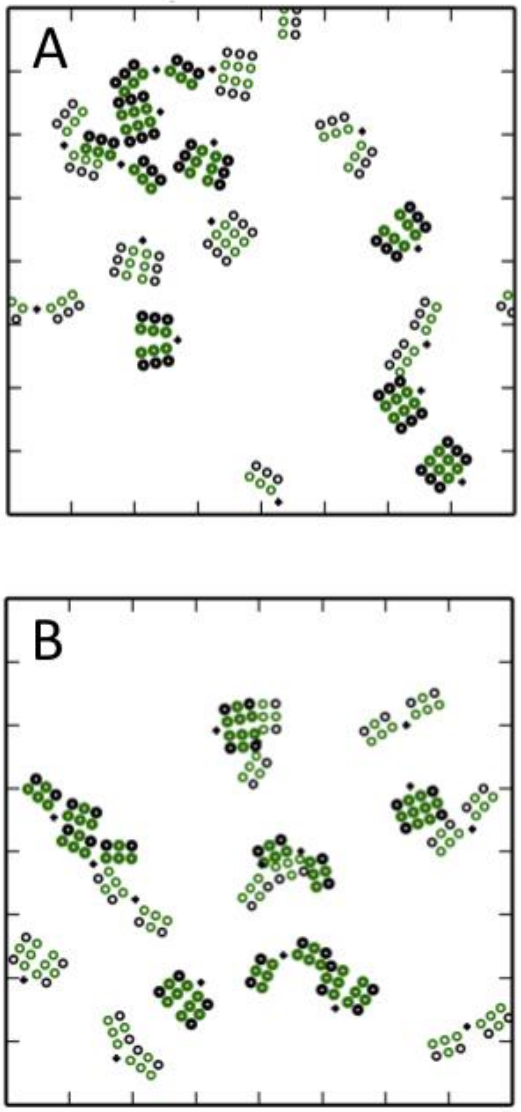



**Figure 3.7.** Representative frames from simulations at high concentration. A) sequence = 0, hinge = 0, cell size = 80, temperature = 2.0. B) sequence = 1, hinge = 2 times the angle between domains, cell size = 80, temperature = 1.0.

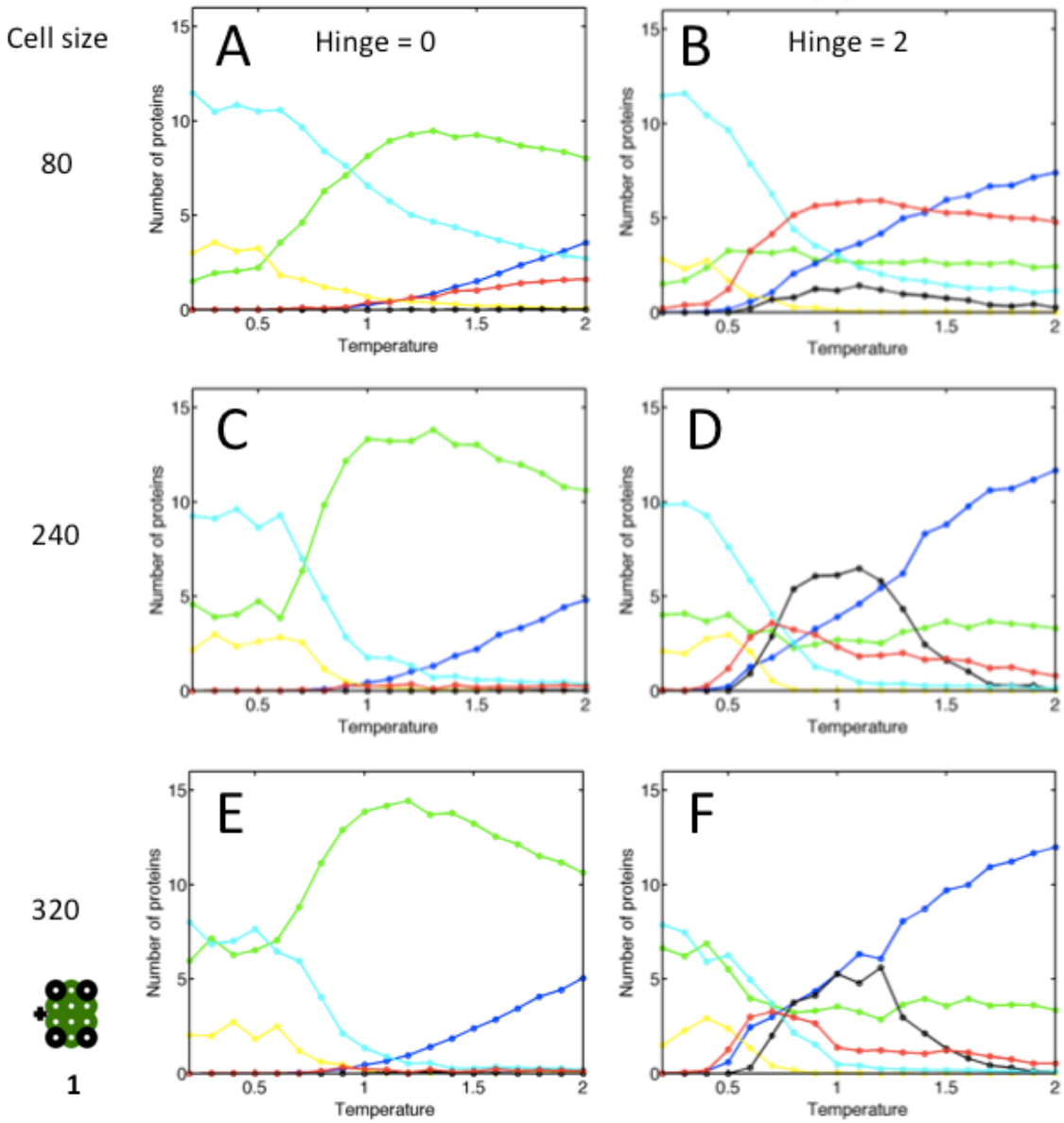



**Figure 3.8.** Simulation statistics as a function of temperature for sequence 1. (See Figure 3.6 caption.)

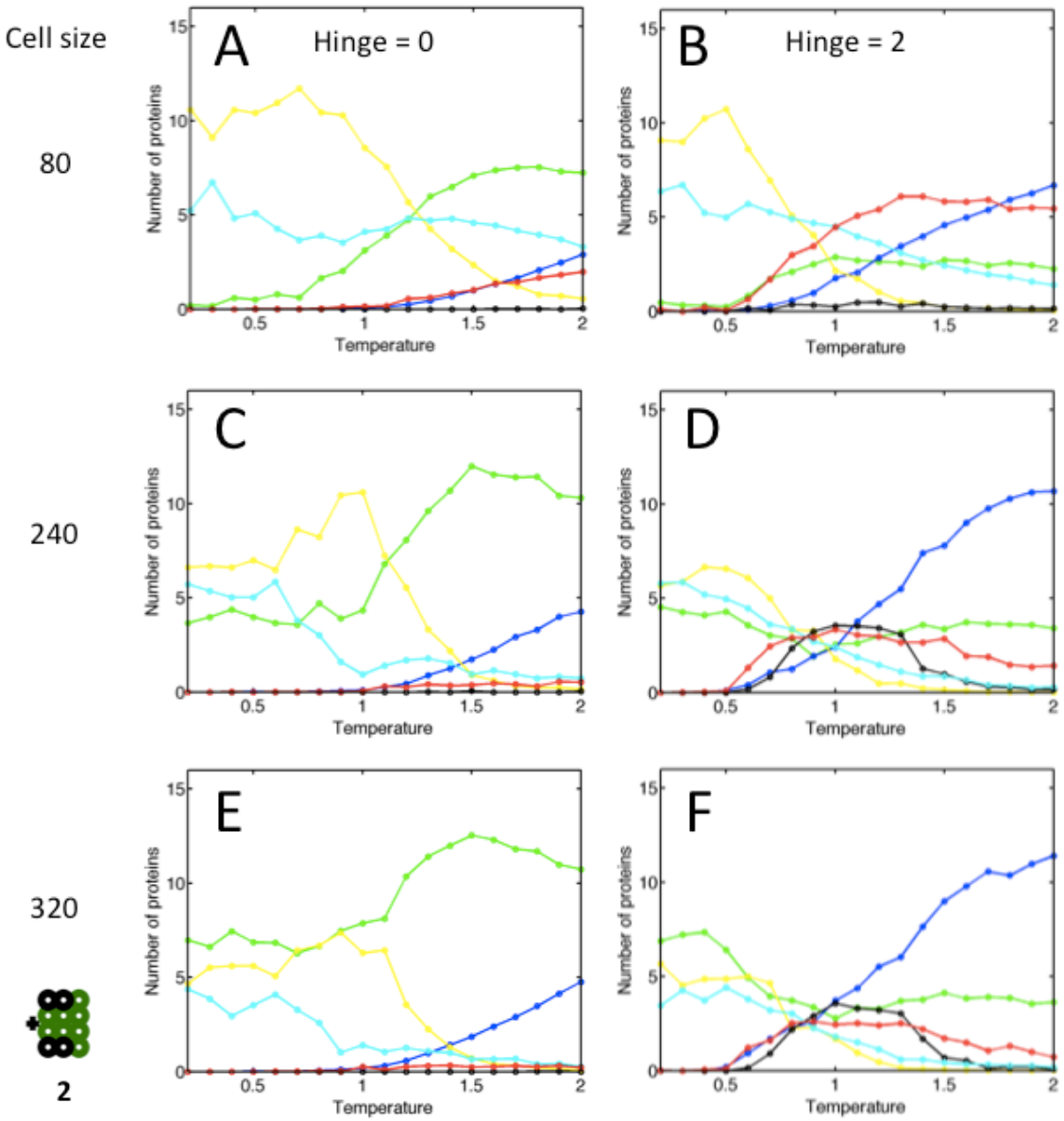



**Figure 3.9.** Simulation statistics as a function of temperature for sequence 2. (See Figure 3.6 caption.)

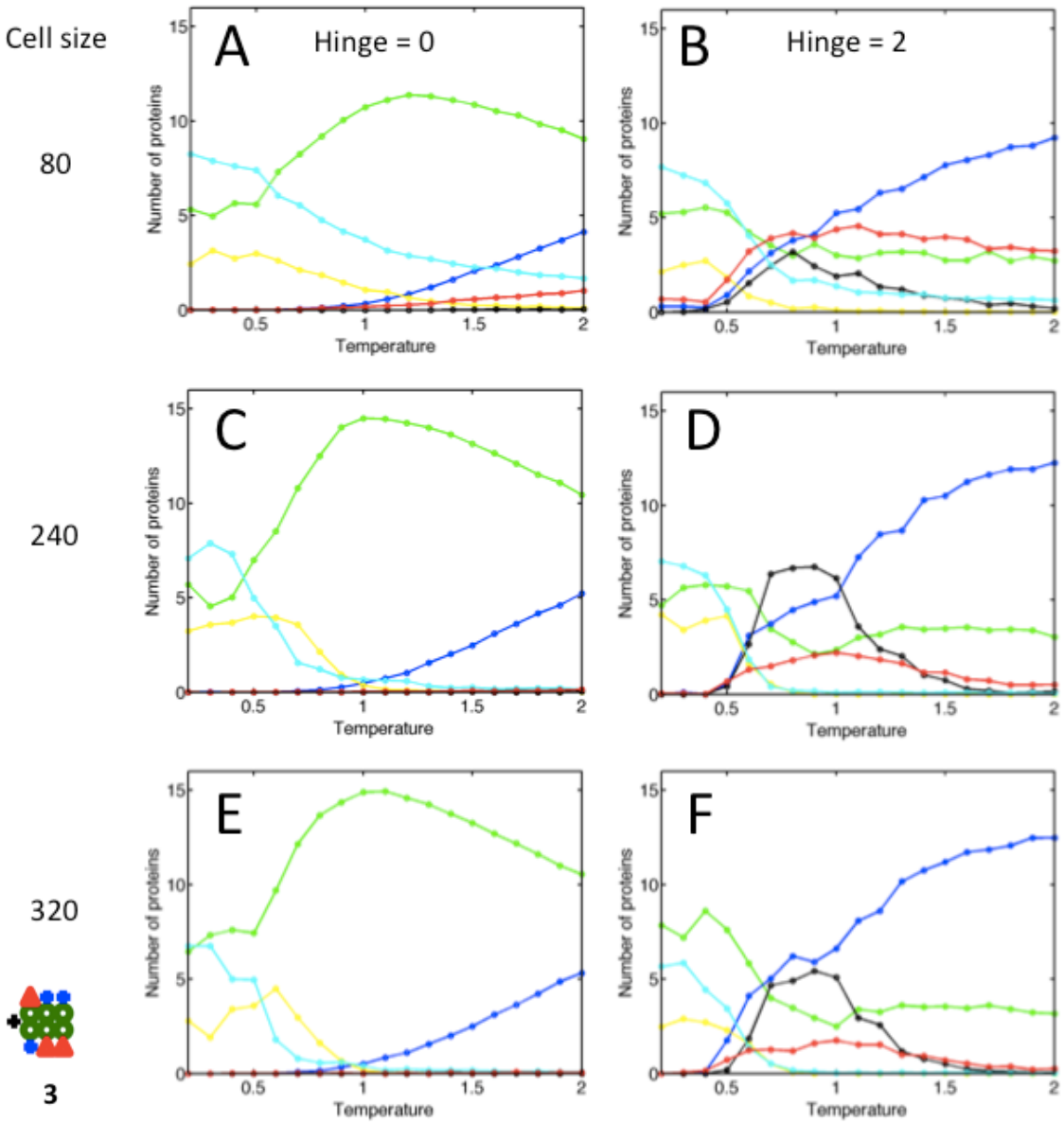



**Figure 3.10.** Simulation statistics as a function of temperature for sequence 3. (See Figure 3.6 caption.)

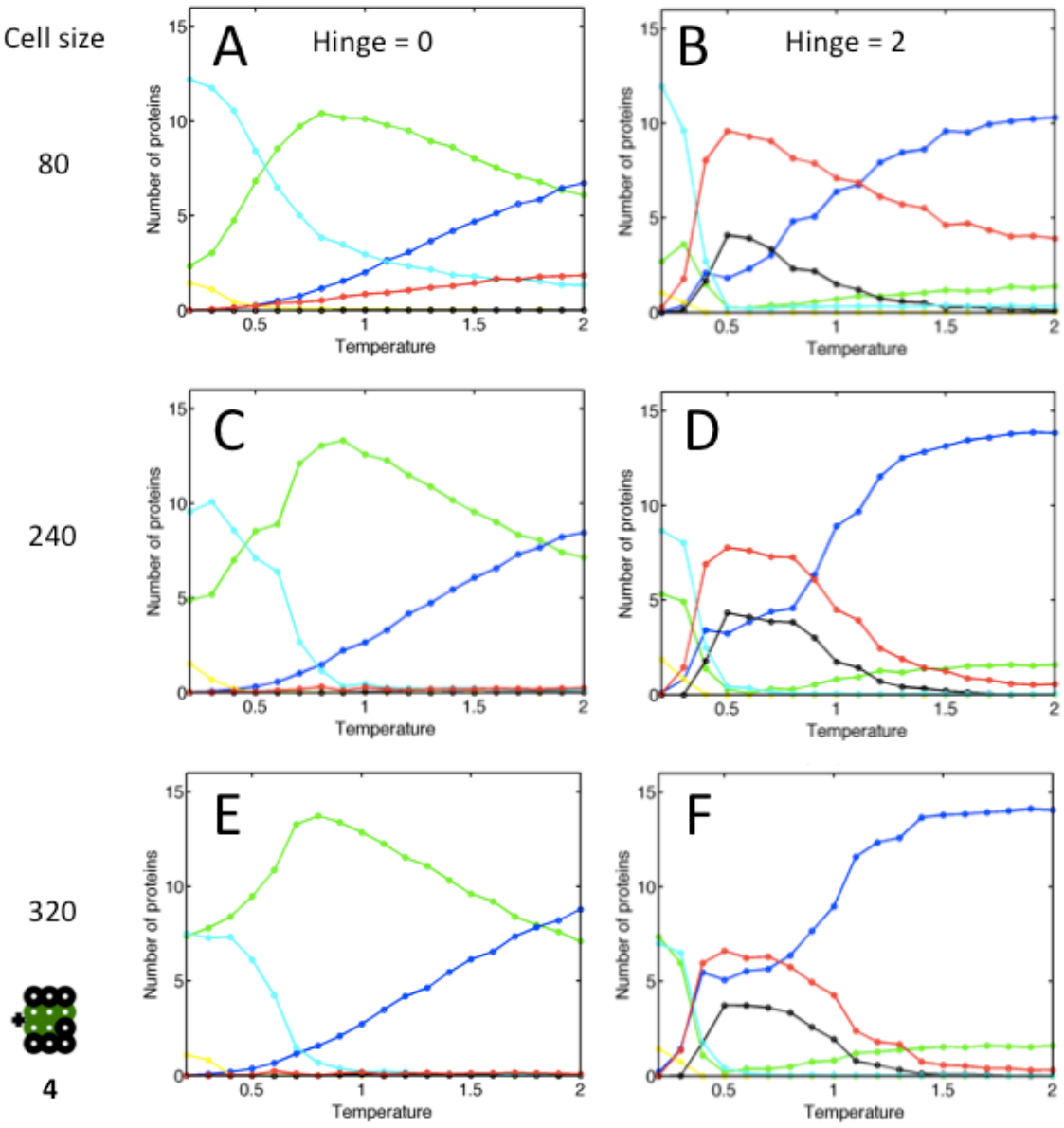



**Figure 3.11.** Simulation statistics as a function of temperature for sequence 4. (See Figure 3.6 caption.)

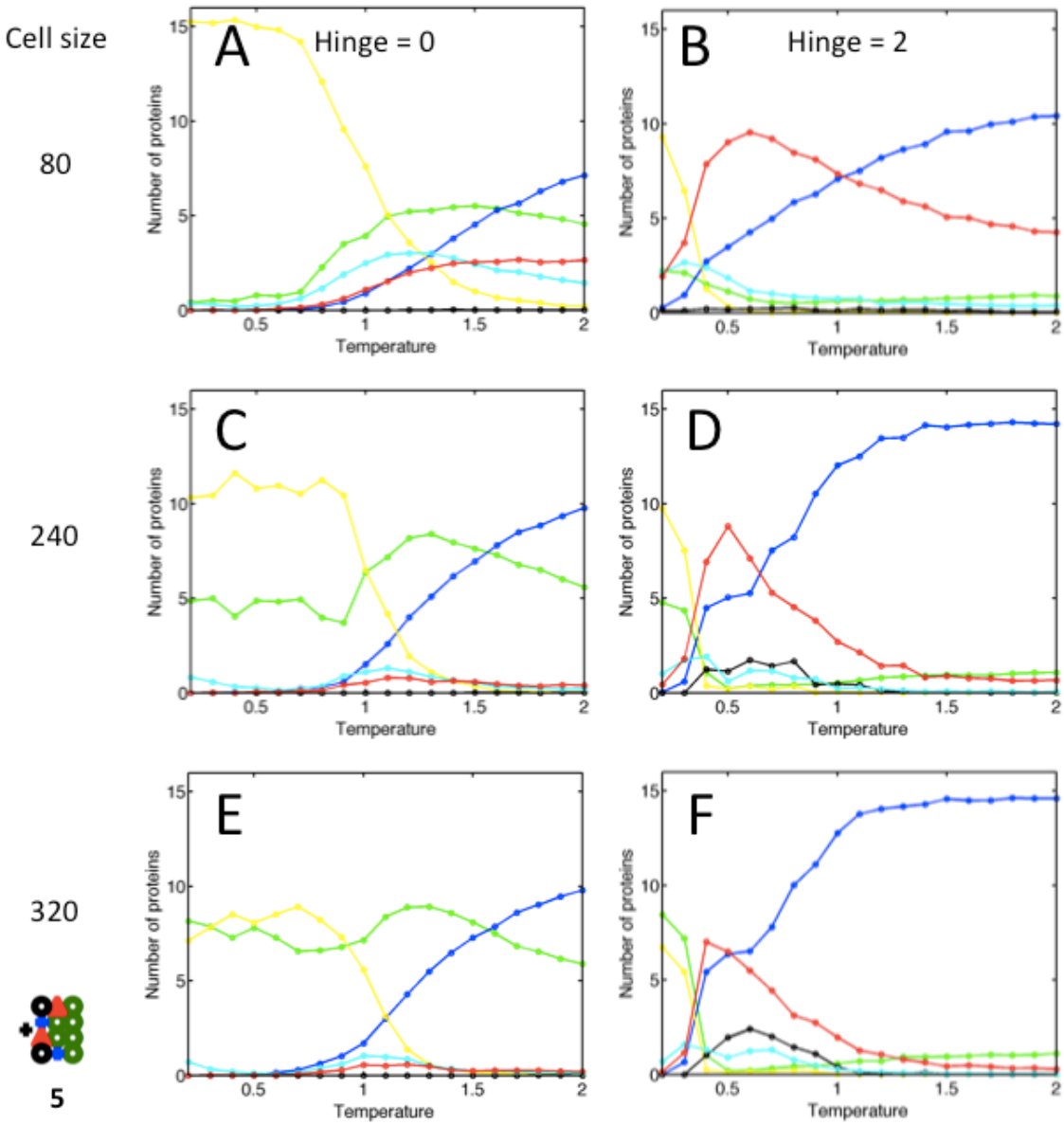



**Figure 3.12.** Simulation statistics as a function of temperature for sequence 5. (See Figure 3.6 caption.)

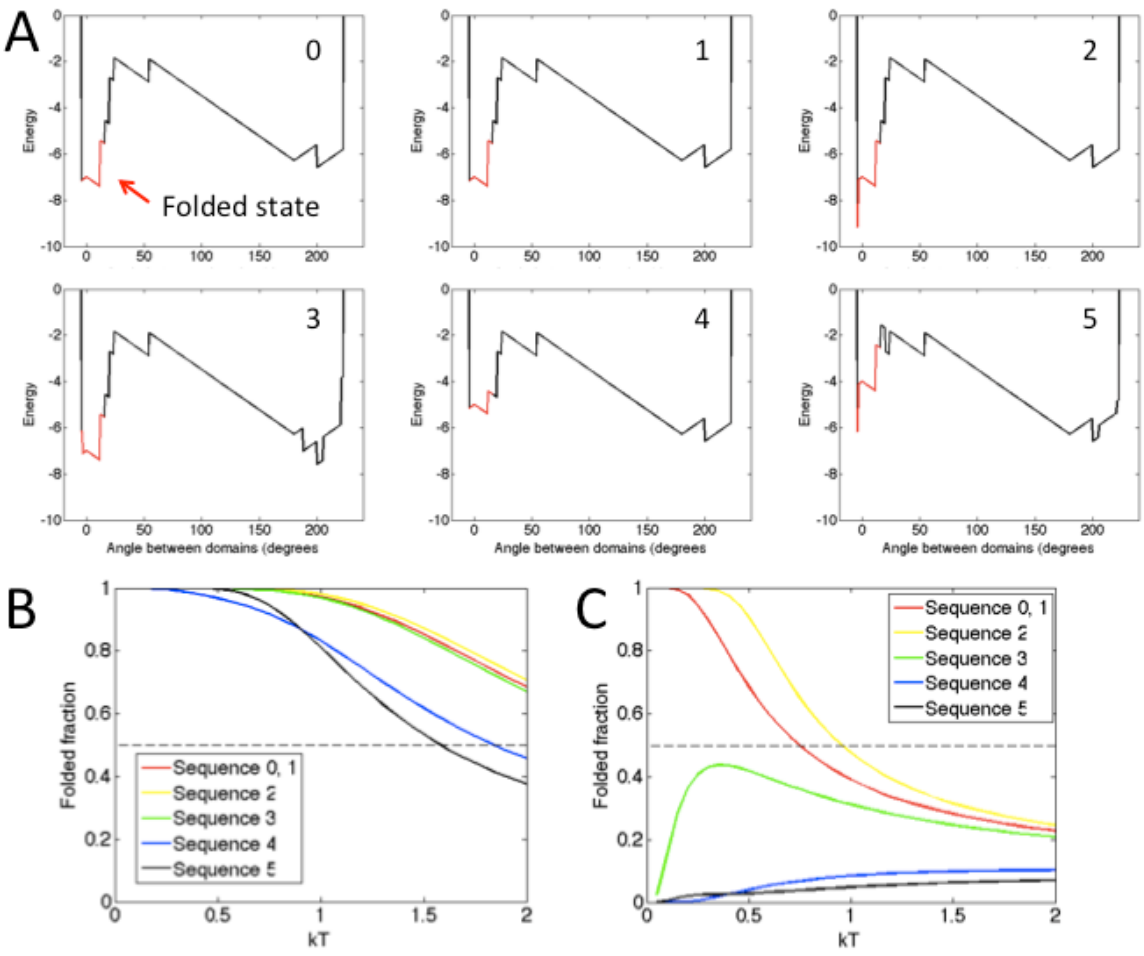



**Figure 3.13.** Single protein energy landscapes and temperature dependence of folded fraction. A) Domain-domain interaction energy as a function of angle between domains (in degrees), for sequences 0-5. Hinge energy = 2 times angle between domains (in radians). The region defined as the folded state is colored red. B) Population of the folded state, calculated from intra-protein interaction diagrams with hinge energy = 0. The dotted line indicates an equal number of folded and unfolded proteins. C) Population of the folded state, calculated from intra-protein interaction diagrams with hinge energy = 2 times the angle between domains (shown in (A)).

*3.3.3. Phase diagrams*

Phase diagrams showing the most prevalent protein species as a function of temperature and concentration are shown in Figure 3.14, for proteins without the hinge energy term. Non-specific interactions between folded proteins are most prevalent for sequence 1, which has hydrophobic residues at the top and bottom interfaces, while functional dimerization is most prevalent for sequences 2 and 5, for which all of the residues at the functional interface are hydrophobic. Note that functional dimerization, which was designed to be stronger than non-functional dimerization, persists out to higher temperatures. We see unfolding at high temperatures for sequences 4 and 5, consistent with predictions from single molecule energies shown in Figure 3.13 B. In general, folded dimers occur at low temperature and low concentration, while folded monomers occur at low concentration and higher temperatures. For destabilized proteins (those with a weaker domain-domain interaction interface), unfolded monomers occur at high temperature. Figure 3.15 shows that the lowest energy occurs in the folded dimeric region of the phase diagram for all six sequences.

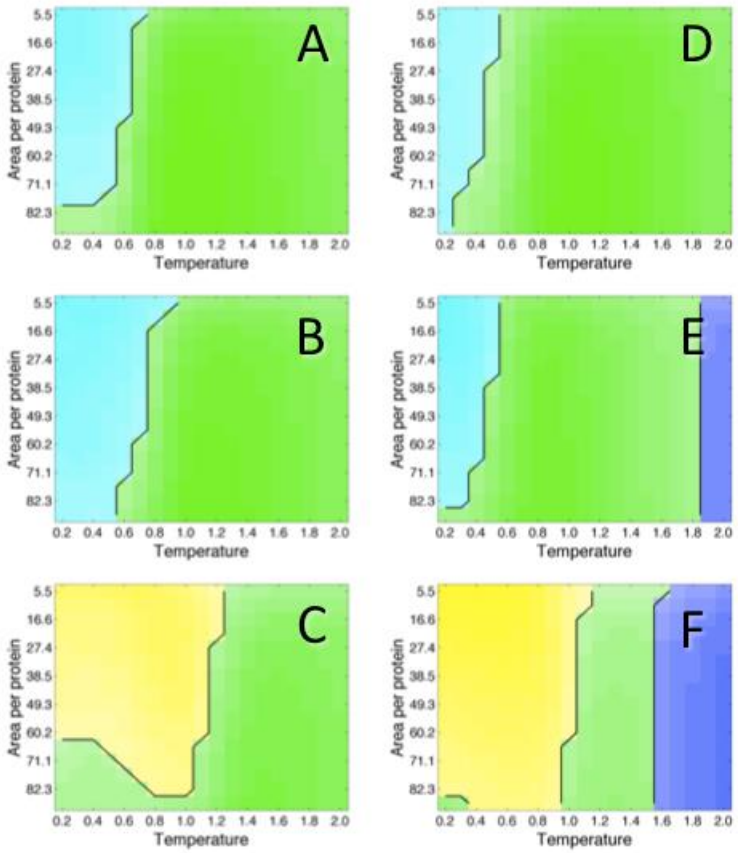



**Figure 3.14.** Phase diagrams showing the most prevalent interaction category as a function of temperature and concentration. Hinge energy = 0. Temperature is given in simulation units of kT, and concentration is given in terms of area per protein, normalized by the length times the width of a single protein (72.96), with cell length ranging from 60 to 320 in simulation units. High concentration corresponds to low area per protein. Color (see Figure 3.6) denotes the most populated category, and shade indicates the population of this category, with darker shades corresponding to a greater number of proteins. Each plot represents a single protein. A) Sequence 0. B) Sequence 1. C) Sequence 2. D) Sequence 3. E) Sequence 4. F) Sequence 5.

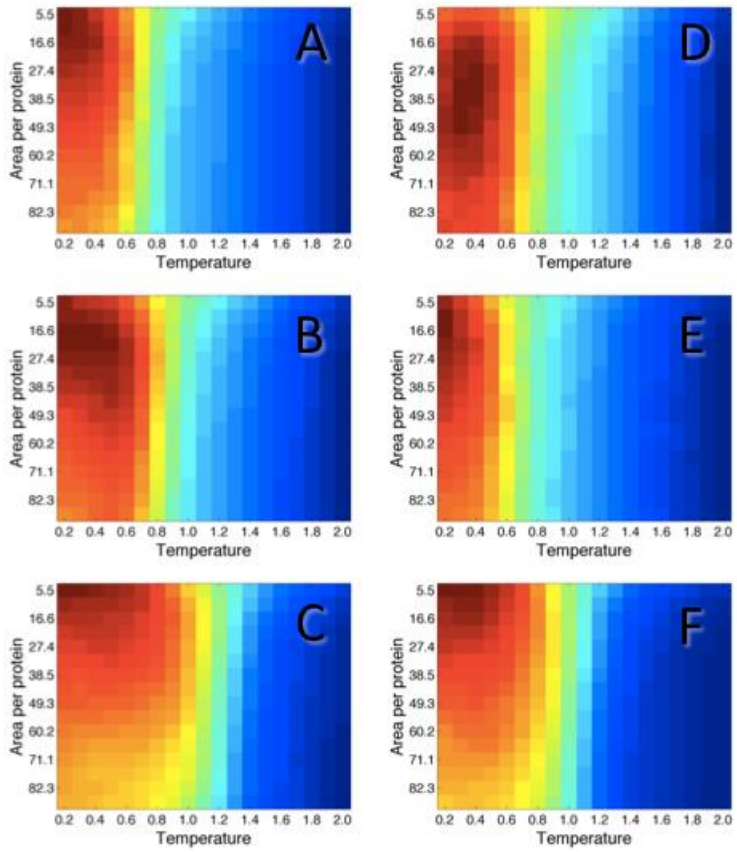



**Figure 3.15.** Total interaction energy from simulations as a function of cell area and temperature. Hinge = 0. Results are averaged over the final 200,000 frames and over 20 individual runs. Dark red indicates the lowest energy, and dark blue indicates the highest energy. Color scales are normalized for each plot. A smoothing function was applied to each plot in two dimensions. A) Sequence 0. B) Sequence 1. C) Sequence 2. D) Sequence 3. E) Sequence 4. F) Sequence 5.

Phase diagrams for proteins with a hinge energy term biasing the domains towards an open state are shown in Figure 3.16. Domain swapping is most common at intermediate temperatures and intermediate concentrations and is seen in the phase diagram for sequences 0, 1, 2, and 3. At intermediate temperature and high concentration, non-specific interactions between unfolded proteins become more common than domain swapping. Which region of phase space has lowest energy depends on the sequence. Figure 3.17 shows that the region favoring folded

dimerization has lowest energy for sequences 0 and 2, while the region favoring domain swapped dimerization has lowest energy for sequences 1 and 3. However, domain swaps do not occur at low temperature, most likely due to the kinetic barrier to unfolding. Similarly, the region that favors non-specific interactions between unfolded proteins has the lowest energy for sequence 4; however, folded dimers are seen at low temperature due to kinetic trapping.

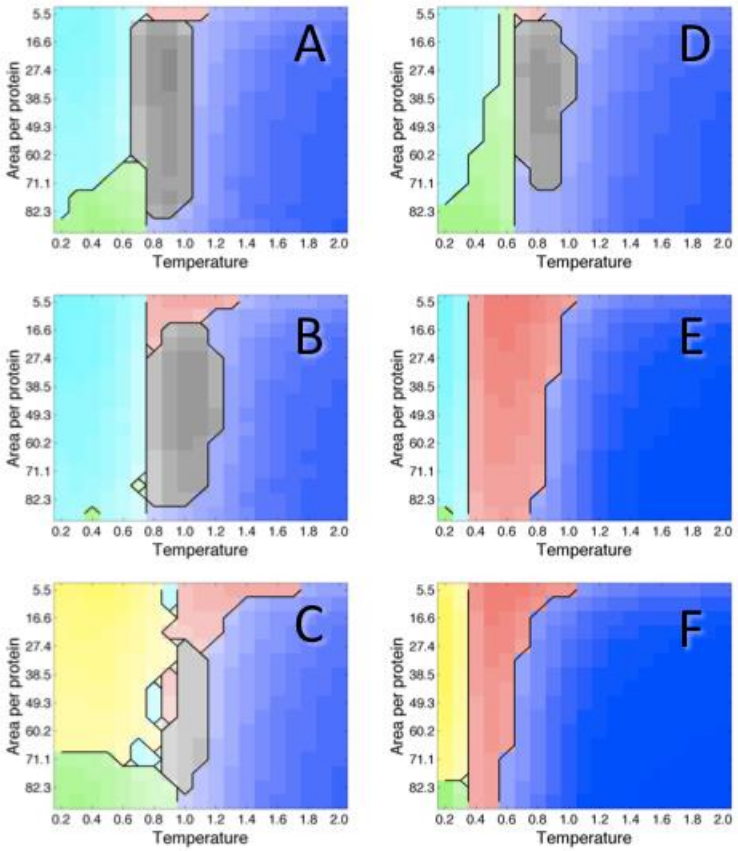



**Figure 3.16.** Phase diagrams showing the most prevalent interaction category as a function of temperature and concentration. Hinge energy = 2 times angle between domains. A) Sequence 0. B) Sequence 1. C) Sequence 2. D) Sequence 3. E) Sequence 4. F) Sequence 5.

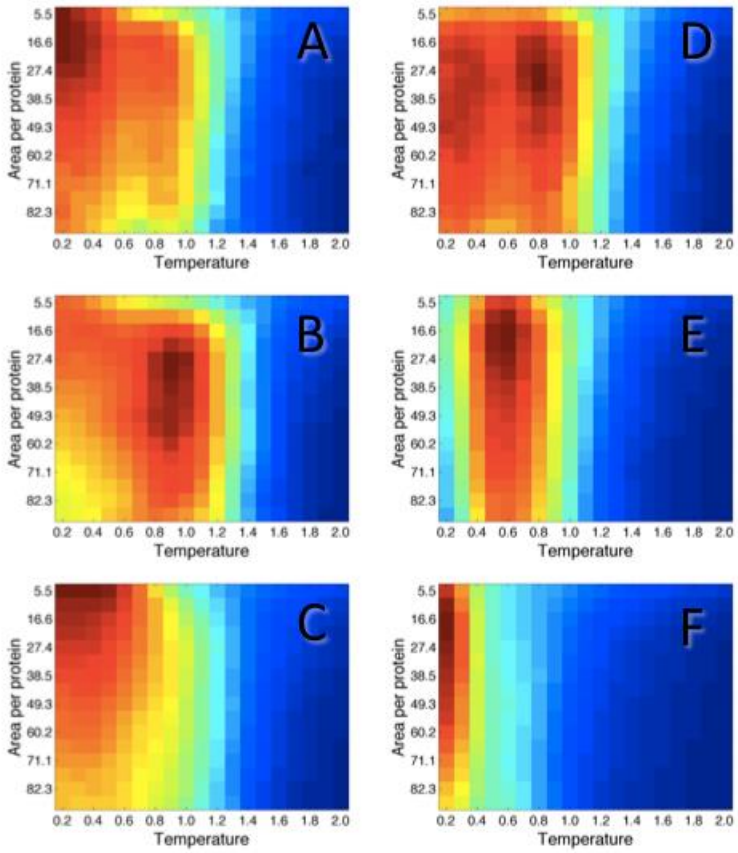



**Figure 3.17.** Total interaction energy from simulations as a function of cell area and temperature. Hinge = 2 times the energy between domains.

In the plots shown above, a smoothing function was applied for ease of visualization (see Methods section 3.2). Raw phase diagrams, without smoothing or contour lines, are shown in Figure 3.18.

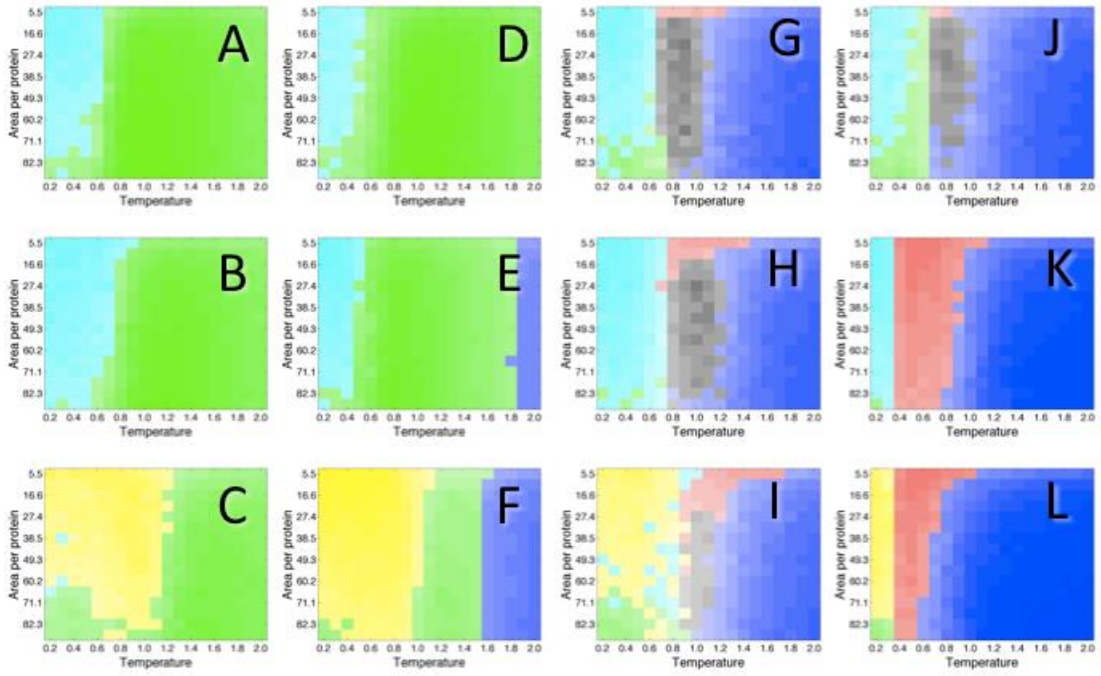



**Figure 3.18.** Raw phase diagrams, prior to applying the smoothing function to generate Figures 3.14 and 3.16. Hinge energy = 0 for (A-F). A) Sequence 0. B) Sequence 1. C) Sequence 2. D) Sequence 3. E) Sequence 4. F) Sequence 5. Hinge energy = 2 times angle between domains for (G-L). G) Sequence 0. H) Sequence 1.. I) Sequence 2. J) Sequence 3. K) Sequence 4. L) Sequence 5.

### *3.4. Discussion*

A key result of simulations that is consistent with experimental observations is the non-monotonic temperature dependence of domain swapping. Domain swapping occurs at intermediate temperatures. At low temperature, proteins are in the folded state, so they do not have the opportunity to domain swap. At intermediate temperatures, proteins begin to unfold due to the higher entropy of the unfolded and partially unfolded states. Proteins then populate the open, domain-swap-prone state and are able to form domain swap interactions. At high temperatures, domain swapped dimers dissociate into unfolded monomers.

Another observation from simulations is that domain swapped dimers persist out to higher temperatures than monomers do at low concentration, creating a folded-like state at temperatures beyond where the monomer would unfold. This could explain the experimental observation that a mutant of the enzyme DHFR that exhibits dimerization at high temperatures shows a beneficial fitness effect when replacing the WT protein in *E coli* (Bershtein et al., 2012). Domain swapped dimerization in this case could rescue the folded state and prevent aggregation. In general, intertwining of protein chains has been proposed as a mechanism to increase protein stability (MacKinnon and Wodak, 2015; Wodak et al., 2015).

In many proteins, a single mutation is sufficient to induce the transition from monomer to domain swapped dimer (Chirgadze et al., 2004; O'Neill et al., 2001; Szymańska et al., 2012; Vottariello et al., 2011). We hypothesized that mutation at the domain-domain interface could increase the propensity for domain swapping. While such a mutation (e.g., protein 4) caused domain swapping to occur at lower temperatures, it increased the propensity for non-specific interactions at intermediate temperatures more than it increased domain swapping. Mutations within the hinge loop that increase hinge strain, or shortening or lengthening of the loop, can also lead to domain swapping in real proteins (Rousseau et al., 2003). We modeled hinge loop torsional strain as a bias favoring the open state, and we found that inclusion of hinge strain causes domain swapping in all six sequences. Our model suggests that modifying the hinge loop to favor domain swapping, while maintaining the primary interface, is the most effective strategy to promote domain swapping.

For proteins without hinge bias, our model shows that the dimer dissociation temperature is highest for proteins with a large hydrophobic surface. In the model, the drop in dimeric protein with increasing temperature is most abrupt at lower concentrations. These dependencies can be predicted by considering a partition function accounting for the interaction between two folded proteins at each surface. For protein 2, which has a strong functional interface and the propensity to form non-functional interactions, functional interactions are most prevalent at intermediate temperatures, while non-functional interactions become more common at low temperatures and monomers dominate at high temperatures. This effect was noted previously for lattice proteins in three dimensions (Deeds et al., 2007).

Another prediction of our model involves the concentration dependence of protein-protein interaction at intermediate temperatures. At low concentrations, monomers are most common, whereas at high concentrations non-specific interactions between unfolded proteins are most common; it is only at intermediate concentrations that domain swapped dimers are the most prevalent species. We hypothesize that this observation is an instance of the Flory theorem for polymer chains (Flory, 1953), which states that unfolded states become common at high concentrations due to the ability of unfolded polymers to form interchain interactions to replace intrachain ones while achieving high entropy. It will be interesting to test experimentally whether domain swapping and/or amyloid formation is decreased relative to amorphous aggregation and non-specific interactions at high concentrations or in crowded environments.

We note that while we observe dimerization at low temperatures for all of our sequences, real proteins do not always dimerize at low temperature. The high surface area to volume ratios of our proteins, and our choice of sequences, could contribute to this bias. The lowest temperatures may also be said to occur below the physiological temperature range. Another potential caveat is that domain swapping may in some cases require complete unfolding of the protein, while only separation of domains is possible in our model. In the cell, where proteins are generally degraded before they achieve full unfolding (Bershtein et al., 2013), the mechanism of domain swapping is likely to involve only partially unfolded states.

In future work, it will be interesting to explore the phase behavior of additional protein sequences and to develop algorithms for the evolution of proteins towards desired interaction states. The results of such studies could suggest design strategies for producing proteins that are folded, dimeric, or domain swapped and may provide insight into dimer evolution. While our simplistic model allows for a large sampling of phase space and of different sequences in a short amount of computational time, it will also be important to study the mechanism of domain swapping with more detailed simulations. For instance, domain-swapped structures have been reproduced in simulations using a Go-like model (Ding et al., 2002; Ding et al., 2006; Yang et al., 2004). As another approach, the Shakhnovich group is currently developing a multichain all-atom Monte Carlo method which accounts for specific and non-specific interactions between protein molecules.

*Contributions*

This work is described in a recent publication (Woodard et al., 2016). Jaie Woodard developed the model and wrote a preliminary version of the simulation program in MATLAB. Sachith Dunatunga wrote the C++ version of the program. Jaie Woodard carried out simulations and analyzed data. Eugene Shakhnovich suggested the relevance of the Flory Theorem to simulation results. Jaie Woodard and Eugene Shakhnovich wrote the paper.

# 4. Stability, disulfide bonding, and aggregation in cataract-associated mutants of γD-crystallin

*Abstract*

γD-crystallin is a highly stable two-domain protein that aggregates in human cataracts. Mutations at tryptophan residues within the protein core are known to accelerate aggregation. However, the mechanism of non-amyloid aggregation of these mutants is unknown. Also unknown is the mechanism of an experimentally observed "inverse prion" behavior of γD-crystallin, in which the WT protein promotes aggregation of the mutant. Here we use Monte Carlo simulations to predict the unfolding pathway and aggregate structures of γD-crystallin. We find that extrusion of the N-terminal hairpin is an early event in protein unfolding. We find in two-molecule simulations that this hairpin interacts with the C-terminal domain of the other protein to form an extended beta sheet in an interaction that resembles domain swapping, which could lead to aggregate formation. Cataract associated mutations and an experimentally observed disulfide bond promote unfolding and aggregation. A domain-domain interaction observed in simulations of WT and mutant proteins could explain how WT accelerates mutant aggregation.

## *4.1. Introduction*

Protein aggregation is implicated in several human diseases, including amyotrophic lateral sclerosis, Parkinson's, and cataracts (Aguzzi and Calella, 2009;

Bartels et al., 2011; Bloemendal et al., 2004; Chiti and Dobson, 2009; Horwich, 2002). Often, a mutation in protein sequence or a covalent modification such as a disulfide bond will increase aggregation propensity and speed onset of a disease. Such modifications can serve to increase the population of an aggregation-prone folding state relative to the native state.

Aggregation of crystallin proteins in the lens leads to human cataract (Michael and Bron, 2011; Serebryany and King, 2014). γD-crystallin has been studied extensively by experimental and computational methods. Its structure consists of two domains, each containing two Greek key motifs (Figure 4.1). Its folding has been studied *in vitro*, and it was found that the C-terminal domain helps to stabilize the less thermodynamically stable N-terminal domain (Flaugh et al., 2005a, b; Flaugh et al., 2006). Several mutations and covalent modifications are known to destabilize γD-crystallin and increase its aggregation propensity. Mutations such as W42Q and W130E were designed to mimic oxidative damage that can lead to age onset cataract. Interestingly, the related mutation W42R was found to cause cataracts in humans (Wang et al., 2011).

While non-amyloid aggregation is generally termed "amorphous," evidence suggests that this type of aggregation may involve specific interactions between proteins (Horwich, 2002; Speed et al., 1996). One hypothesis is that aggregation may occur by run-away domain swapping (Guo and Eisenberg, 2006; Rousseau et al., 2003). However, the structure of amorphous aggregates is difficult to study experimentally, and few computational studies have sought to determine aggregate structure. One exception is a computational study by Das et al. (Das et al., 2011) of

γD-crystallin, in which molecular dynamics simulations were first used to unfold the N-terminal domain and then to simulate interaction of partially unfolded molecules.

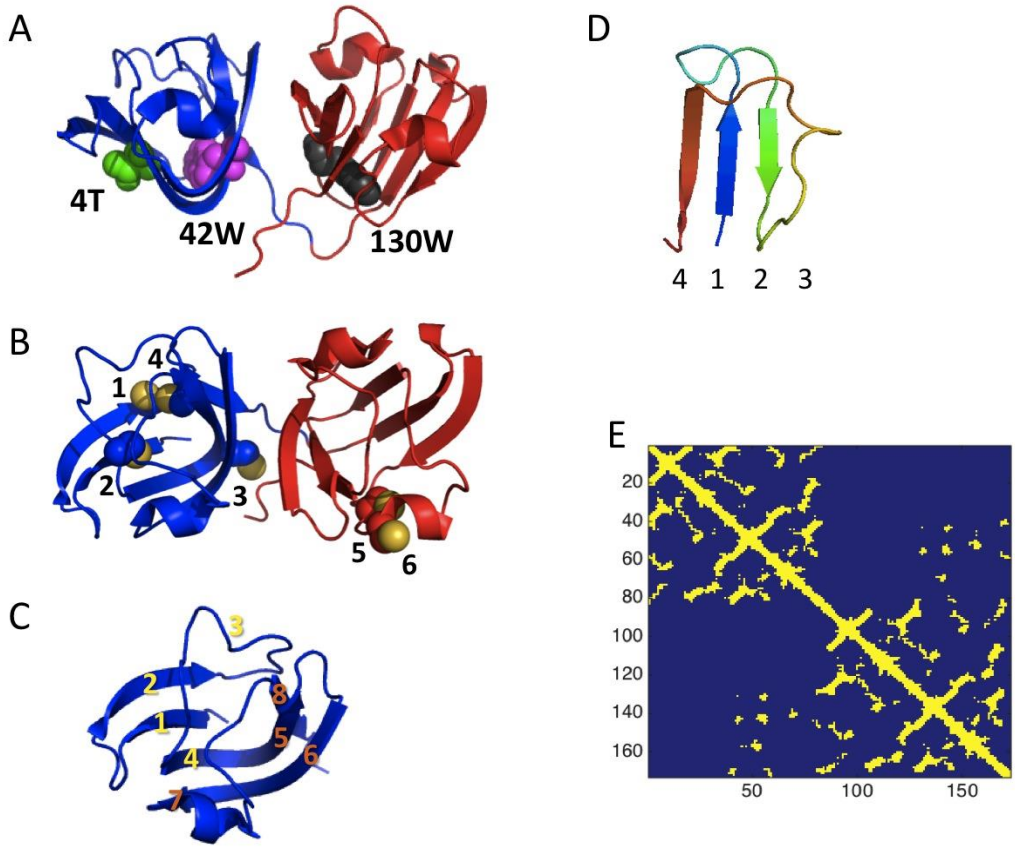



**Figure 4.1**. Structure of γD-crystallin. N-terminal domain is shown in blue, C-terminal domain in red. A) Resdiues mutated in mutations are shown in sphere representation. B) Cysteine residues (sphere representation) are labeled. C) Strands within the N-terminal domain are numbered. D) Strands within the first Greek key motif. E) Contact map, where residues are in contact if their alpha carbons are within 10 Angstroms.

Working closely with experimental collaborators, we simulated unfolding of human γD-crystallin and several mutants using Monte Carlo simulations. We then simulated interaction of γD-crystallin molecules, starting from the native state but allowing for unfolding, by connecting two molecules by a flexible linker. We found

that cataract associated mutants and an experimentally observed disulfide bond promote non-amyloid aggregation involving interactions between specific beta strands. Based on simulations, we propose a mechanism for mutant aggregation, where the N-terminal hairpin binds to the C-terminal domain of the next protein. In addition, we propose a mechanism for the experimental observation that WT proteins can accelerate mutant aggregation, based on domain-domain interactions observed in simulations, where a half-domain-swap interaction promotes unfolding of the free domain.

## *4.2. Methods*

### *4.2.1. Monte Carlo Simulation Program*

Simulations were performed using the Shakhnovich group's all-non-hydrogen atom Monte Carlo program, described in the Introduction section of this thesis. To simulate disulfide bonding, a term proportional to $(d - 2)^2$, where d is the distance in Angstroms between sulfur atoms, was added to the energy function.

### *4.2.2. Unfolding simulations*

The initial simulation structure for single molecule simulations was human γD-crystallin (PDB ID IHK0). For mutants W42Q, W42R, T4P, and W130E, the amino acid mutation was introduced using PyMOL. An initial 2,000,000-step simulation was run at low temperature (T = 0.150 in simulation units). The final structure of this simulation was used as input for a 60,000,000-step simulation, at each of 32 temperatures ranging from T = 0.1 – 1.65, with 50 runs performed at each

temperature. The final frame was extracted from each run, and RMSD from the native structure was calculated separately for the N-terminal domain and C-terminal domain and averaged over the 50 runs. Curves of RMSD vs. temperature were fit to a sigmoidal function, corresponding to a two-state model of unfolding. Contact maps were generated from final simulation frames for W42R, where a contact occurred if the residue-residue alpha-carbon distance was less than 10 Angstroms. Maps were clustered using the MATLAB clusterdata function with a cutoff of 0.9, and representative images were generated for representative structures from the two most populated clusters. For W42R, 300 additional simulations were performed at T = 0.800, and these simulations were visually mined for misfolded structures.

*4.2.3. Two-molecule tethered simulations*

Simulations were carried out in which two γD-crystallin molecules (PDB ID 1HK0) were connected by a 12-residue linker (GSGSGSGSGSGS). The linker was generated in ModLoop (Fiser and Sali, 2003). 300 separate 80,000,000-step Monte Carlo simulations were run for WT, W42R, W42Q, and W42R with each possible disulfide bond (6 total) between cysteines in the N-terminal domain. Frames were extracted every 5,000,000 simulation steps. Contact maps were generated, where residues were said to be in contact of their alpha carbons were within 10 Angstroms. A protein-protein interaction was said to occur if more than 50 pairwise residue-residue contacts occurred between proteins. Beta strand interactions were found by looking for diagonal lines on the contact map. An interaction was labeled

an antiparallel beta strand interaction if it included more than 6 residue pairs in a row in an antiparallel configuration. An interaction was labeled a parallel beta strand interaction if it included more than 6 residue pairs in a row in a parallel configuration. The interaction was labeled as native-like if more than 10 of the interactions were present between the domains of the native structure. For parallel and antiparallel beta strand interactions, we charted which beta strands were involved in the interaction.

## *4.3. Results*

### *4.3.1. Unfolding pathway and folding intermediates*

Unfolding of human γD-crystallin was simulated using the Shakhnovich group's All-Atom Monte Carlo simulation program. Figure 4.2 shows a sample unfolding trajectory. The first major unfolding event is separation of the two domains. Next, the N-terminal hairpin of the N-terminal domain separates from the rest of the protein. In some trajectories, splitting of the N-terminal domain into two Greek keys is observed prior to detachment of the N-terminal hairpin. The N-terminal domain then continues to unfold while the C-terminal domain begins unfolding, generally starting with detachment of the N-terminal hairpin within the C-terminal domain.

Unfolding simulations were also carried out for the aggregation-prone mutant W42R, to determine partially-unfolded states that might be involved in aggregation. Two unfolding intermediates were identified using a clustering approach (Figure 4.3). In structure 4.3-A, the N-terminal hairpin is detached from

the rest of the protein, while the C-terminal domain and the second Greek key remain intact. We note that cysteines at residues 32 and 41 (i.e., cysteines 2 and 3) could easily come together to form a disulfide bond starting from this structure. In the second structure (4.3-B), the two Greek keys of the N-terminal domain are separated, while each Greek key individual remains intact, and the C-terminal domain remains folded.

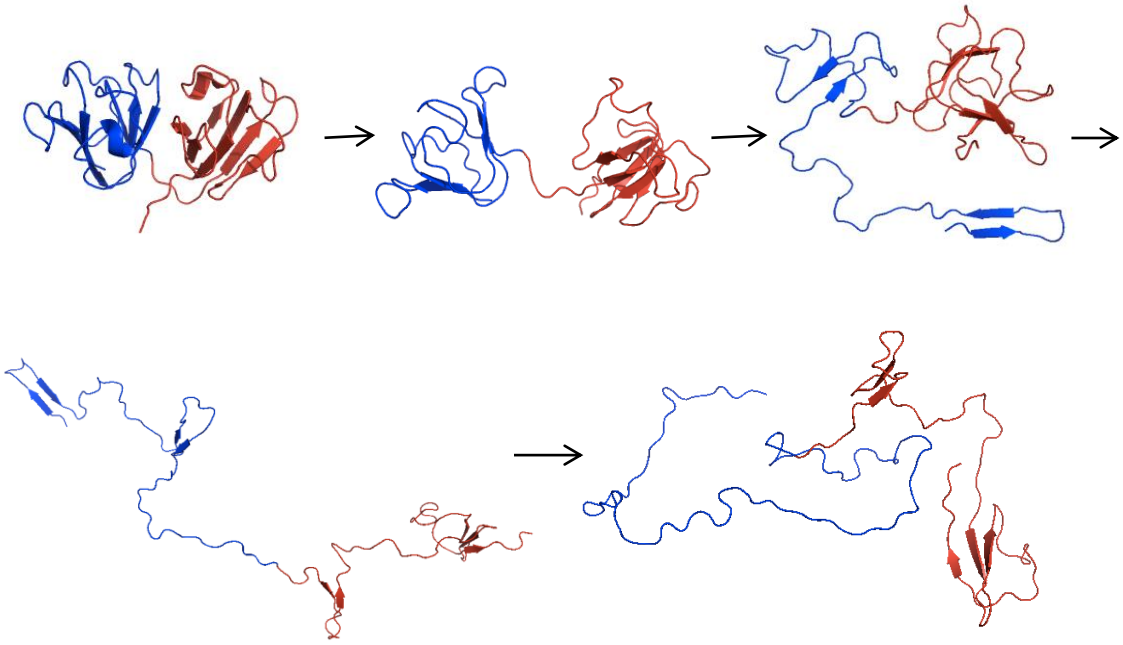


**Figure 4.2**. Simulated unfolding pathway of WT γD-crystallin. Separation of the two domains and detachment of the N-terminal hairpin are early events in the unfolding pathway.

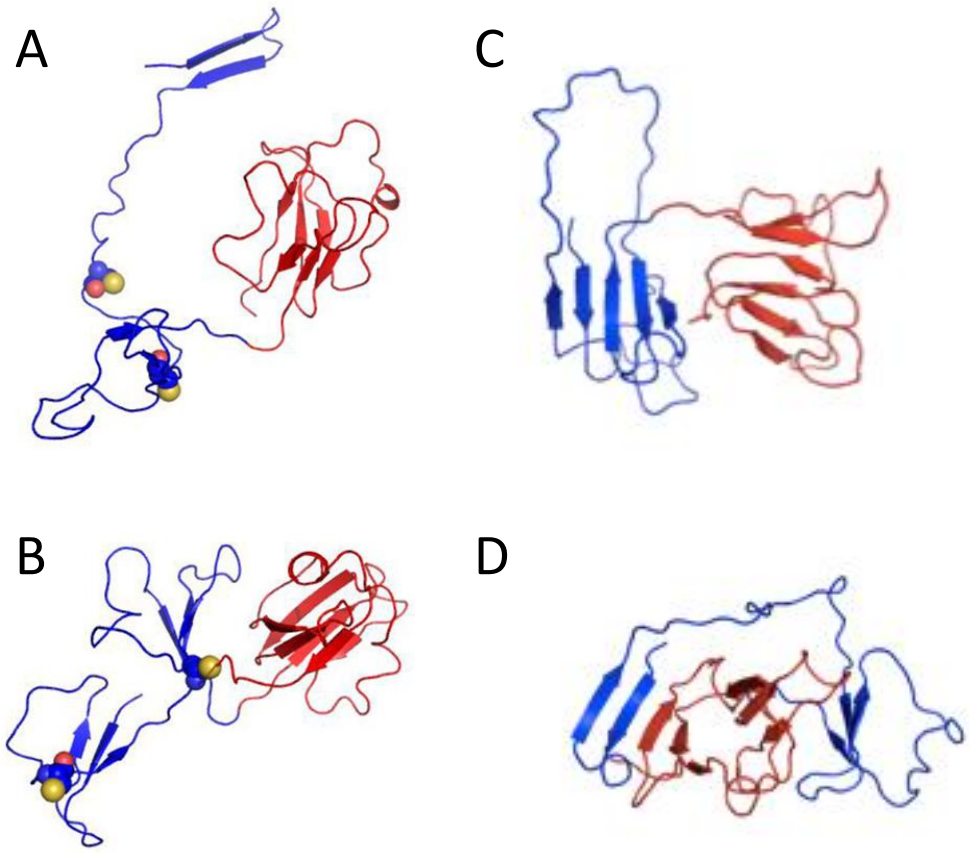



**Figure 4.3.** Representative structures from W42R simulations at T = 0.800. A-B) Highly populated partially-unfolded structures. Cysteines 2 and 3 are shown in sphere representation. C-D) Misfolded structures.

Off-pathway rearranged structures, which include contacts not found in the native structure, were also observed for W42R. Structure 4.3-C shows antiparallel hydrogen bonding of strand 1 to strand 8 (see labeling scheme in Figure 4.1-C) within the N-terminal domain, to form an extended beta sheet. Structure 4.3-D shows antiparallel hydrogen bonding between strand 1 and strand 14 (i.e., the 6$^{th}$ strand of the C-terminal domain). Such misfolded structures may be precursors in aggregate formation, or they may suggest intermolecular interactions that could be involved in aggregation.

*4.3.2. Effects of mutation on stability*

WT and mutants of γD-crystallin (W42Q, W42R, T4P, and W130E) were each simulated at several temperatures, to determine how mutation affects simulated melting temperature. RMSD from the folded state was plotted for the N-terminal domain and C-terminal domain separately, using the last step of simulations, averaged over 50 runs (Figure 4.4). In the WT protein, the N-terminal domain was found to unfold at a lower temperature than the C-terminal domain. Mutations W42Q and W42R destabilize the N-terminal domain, showing very similar apparent melting curves. T4P increases the N-terminal domain RMSD primarily at temperatures below the melting temperature for WT. This is likely due to detachment of the N-terminal hairpin from the rest of the N-terminal domain at relatively low temperatures. W130 is located within the C-terminal domain, and W130E in fact decreases the melting temperature of this domain. In general, mutants are found to destabilize the domain in which they are located, consistent with expectations and with experimental results (Serebryany et al., 2016a; Serebryany et al., 2016b). We do not find in our simulations that a mutation in one domain significantly affects the melting temperature of the other domain.

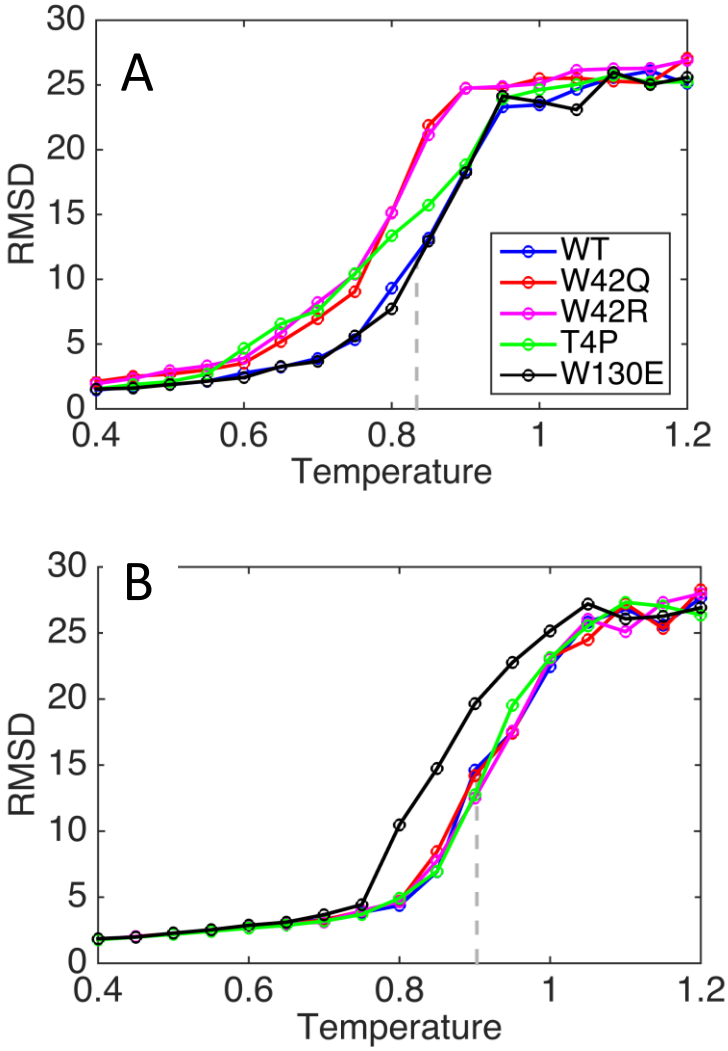



**Figure 4.4.** Simulated unfolding curves for WT and mutants. A) N-terminal domain. B) C-terminal domain.

*4.3.3. Two-molecule simulations predict the dependence of aggregation propensity on mutation and disulfide bonding*

In order to simulate the interaction of two protein molecules, two molecules were connected by a 12-residue glycine-serine linker. The linker served two purposes: first, it limited the distance between molecules to facilitate interaction, and second, it allowed us to simulate two γD-crystallin molecules using the singe-chain Monte Carlo program. The initial structure for two-molecule simulations is shown in Fig. 4.5. It was found that the number of simulation frames that exhibited a

substantial number of protein-protein interactions was larger for mutants W42R and W42Q than for WT (Figure 4.6).

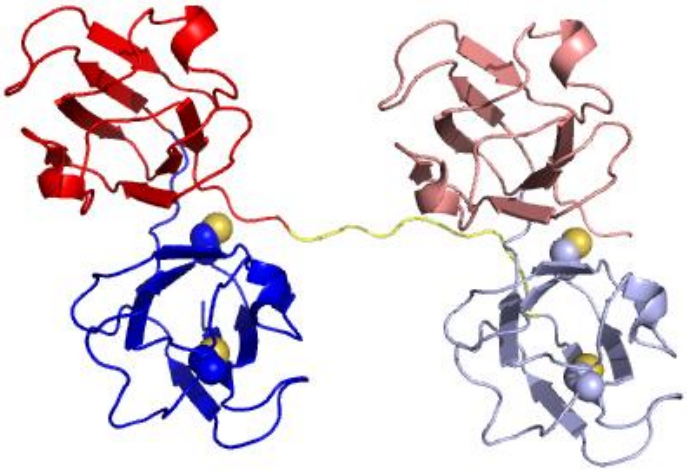


**Figure 4.5.** Initial structure for two-molecule single chain simulations. N-terminal domains are shown in blue and C-terminal domains are shown in red. Molecule 2 is shown in darker colors than molecule 1. Cysteines 2 and 3 are shown in sphere representation.

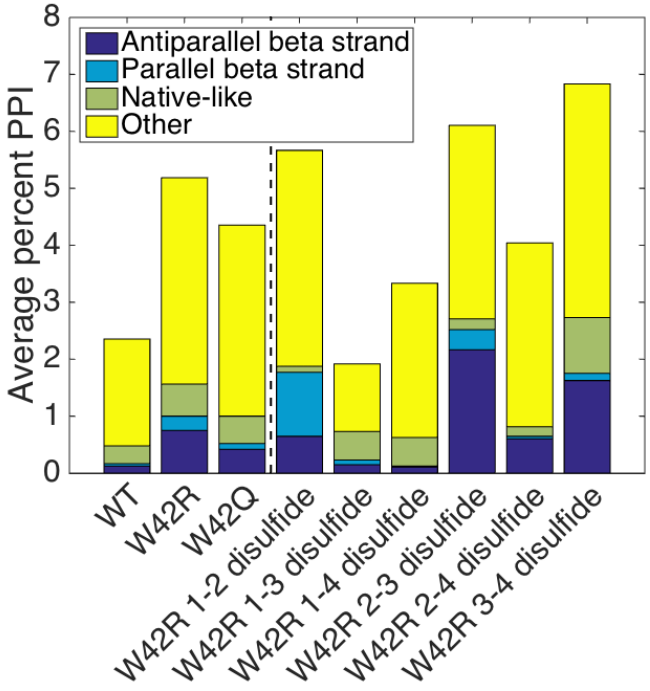



**Figure 4.6.** The percentage of frames showing protein-protein interactions, averaged over 300 trajectories.

Intramolecular disulfide bonding was modeled by introducing a harmonic energy term dependent on the distance between sulfur atoms within cysteine residues. For the W42R mutant, we carried out simulations with each possible disulfide bond between cysteine residues in the N-terminal domain. The disulfide bond formed early in simulations, due to the strength of the energy term. Disulfide bonding was found to increase the amount of protein-protein interaction for disulfide bonds between adjacent cysteines (Figure 4.6). However, for non-adjacent cysteines, disulfide bonding actually decreased the amount of protein-protein interaction. Figure 4.7 shows structures of proteins exhibiting each possible disulfide bond. The adjacent disulfides tend to promote detachment of the N-terminal hairpin from the rest of the N-terminal domain, leading to a more open structure with more solvent exposed residues. The non-adjacent disulfides, however, tend to promote more compact structures. For instance, the 1-4 disulfide creates a bond between cysteines that are nearby in the native structure, enforcing a structure that is native-like. In this case, the disulfide helps to lock the protein in a native-like conformation, decreasing the amount of protein-protein interaction.

*4.3.4. Two-molecule simulations predict aggregate structure*

We noted three categories of protein-protein interaction in two-molecule simulations. First is a native-like interaction, in which the N-terminal domain of one protein forms native-like contacts with the C-terminal domain of the other protein (Fig. 4.8 A). Such an interaction does not require unfolding of either domain. Second is a parallel interaction between beta strands, which often occurs between strand 1

of each protein (Fig 4.8-B). Third is an antiparallel interaction between beta strands (Fig. 4.8 C-E). Which category is most prevalent depends on the mutant and the disulfide bond present (see Figure 4.6).

A disulfide bond between the second and third cysteines was observed *in vitro*, as a requirement for aggregation of the W42Q mutant, and *in vivo* (Fan et al., 2015; Serebryany et al., 2016b). Notably, the 2-3 disulfide led to the greatest number of antiparallel beta strand interactions in our simulations. For molecules containing the 2-3 disulfide, antiparallel beta strand interactions were far more common than native-like interactions or parallel beta strand interactions.

For the W42R mutant with the 2-3 disulfide bond, we generated strand-strand contact maps for beta strand interaction, in order to determine which strands are most often involved in interactions (Figure 4.9). The most common interaction was an antiparallel interaction between beta strands 1 and 14. Next was an antiparallel interaction between strand 1 and strand 6, followed by an antiparallel interaction between strand 1 from each protein. Note that all three structures require dissociation of the N-terminal hairpin from the rest of the N-terminal domain, although the hairpin itself can remain intact. The 1-6 and 1-14 structures can be propagated (molecule A binds to molecule B, which binds to molecule C, etc.), while the 1-1 structure cannot.

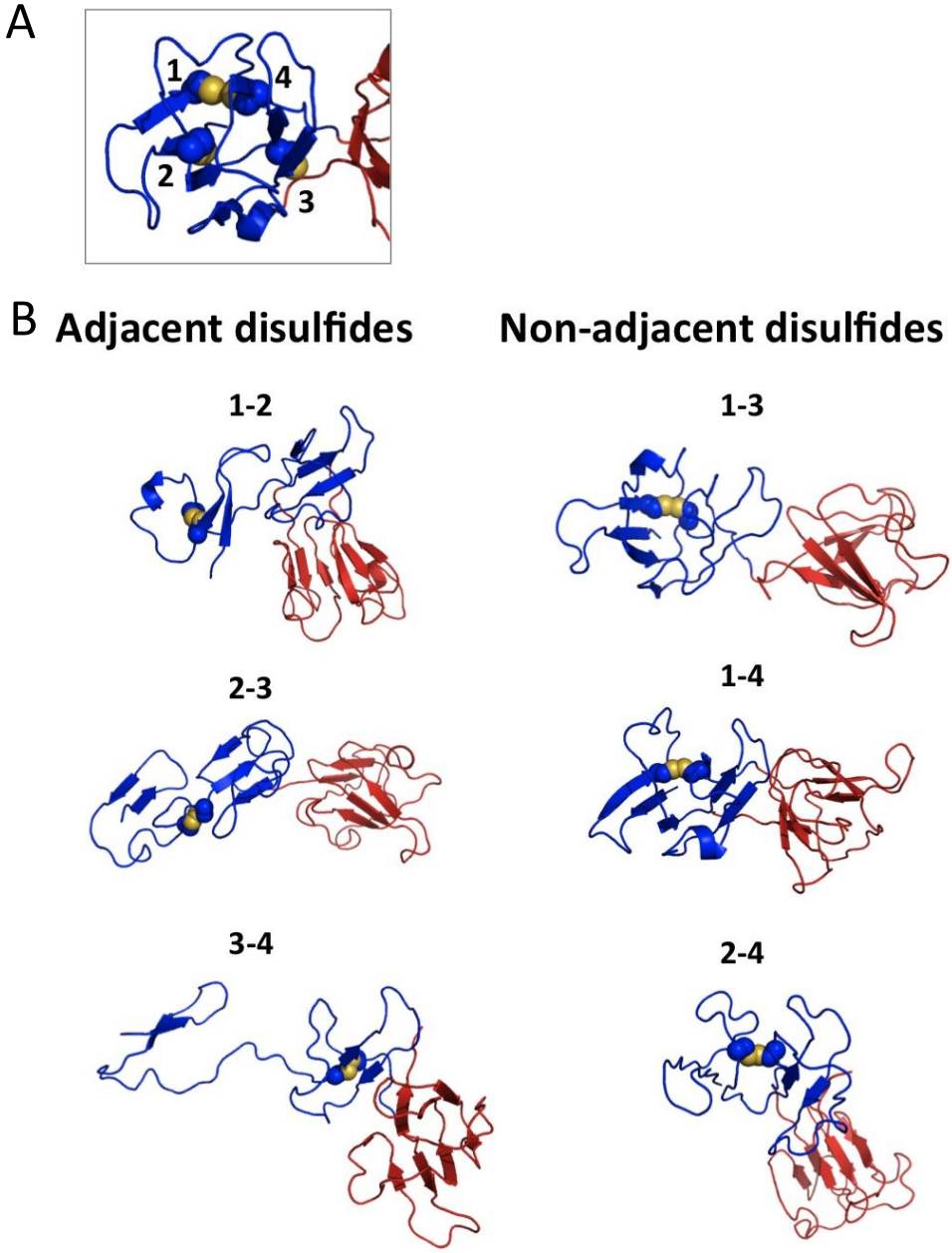



**Figure 4.7.** Disulfide bonds affect protein conformation and propensity to aggregate. A) N-terminal domain with its four cysteines labeled and cysteine alpha carbons and sidechain heavy atoms shown in sphere representation. B) Sample structures from step 40,000,000 of 2-molecule simulations, for each possible disulfide bond.

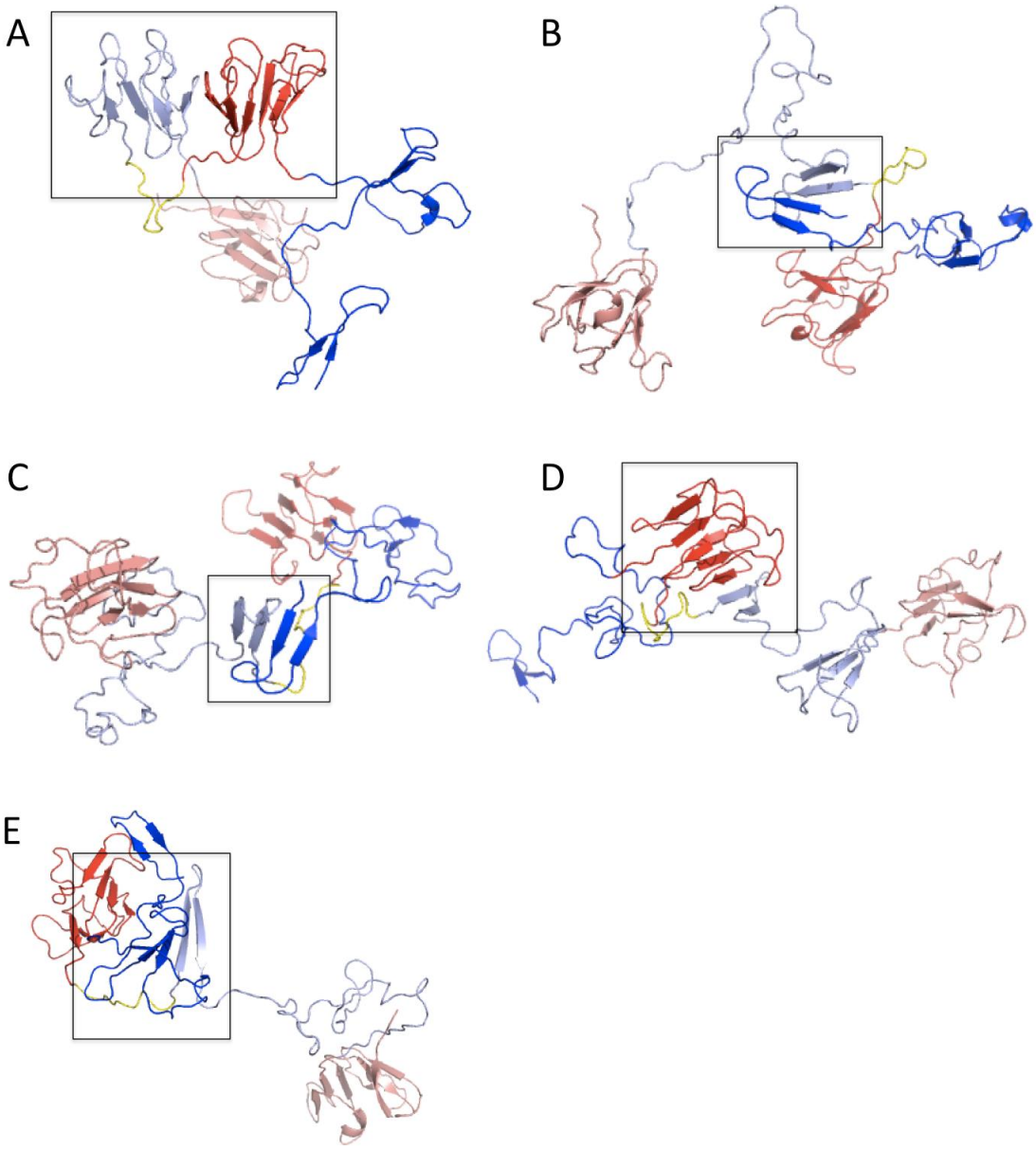
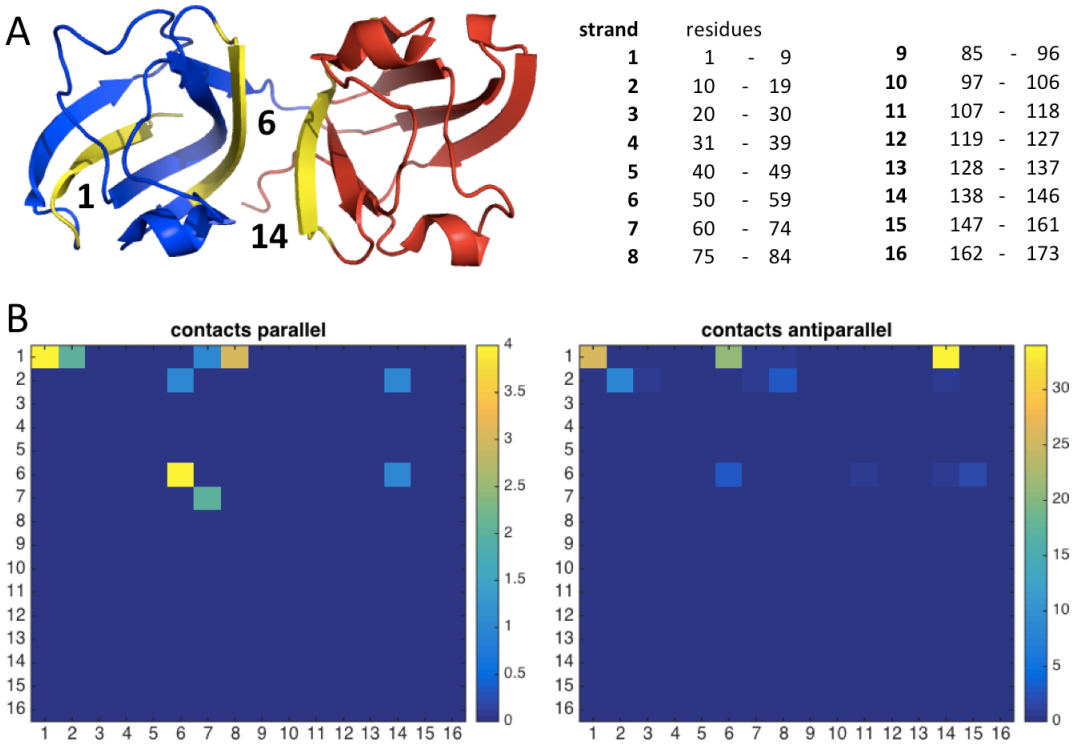



**Figure 4.8.** Examples of protein-protein interaction from simulations. A) Native-like interaction between domains. B) Parallel interaction between strand 1 from each protein. C) Antiparallel interaction between strand 1 from each protein. D) Antiparallel interaction between strand 1 and strand 14. E) Antiparallel interaction between strand 1 and strand 6.




| strand | residues | | | | | |
|---|---|---|---|---|---|---|
| **1** | 1 | - | 9 | **9** | 85 | - 96 |
| **2** | 10 | - | 19 | **10** | 97 | - 106 |
| **3** | 20 | - | 30 | **11** | 107 | - 118 |
| **4** | 31 | - | 39 | **12** | 119 | - 127 |
| **5** | 40 | - | 49 | **13** | 128 | - 137 |
| **6** | 50 | - | 59 | **14** | 138 | - 146 |
| **7** | 60 | - | 74 | **15** | 147 | - 161 |
| **8** | 75 | - | 84 | **16** | 162 | - 173 |

**Figure 4.9.** Strand-strand contacts from two-molecule simulations of W42R with 2-3 disulfide. A) Strands 1, 6, and 14 are colored yellow. Residues contained in each strand are listed. B) Strand-strand contact maps for parallel and antiparallel interactions.

Based on the strand 1-strand 14 simulation structure, we propose a possible aggregation mechanism (Figure 4.10). First, the N-terminal hairpin detaches from the rest of the N-terminal domain. Next, a disulfide bond forms between the second and third cysteines of the N-terminal domain, stabilizing a partially unfolded intermediate state with the N-terminal hairpin accessible. Finally, strand 1 from the hairpin hydrogen bonds to strand 14 of the next protein, extending a beta sheet within the C-terminal domain. This structure can propagate indefinitely to form a non-amyloid aggregate.

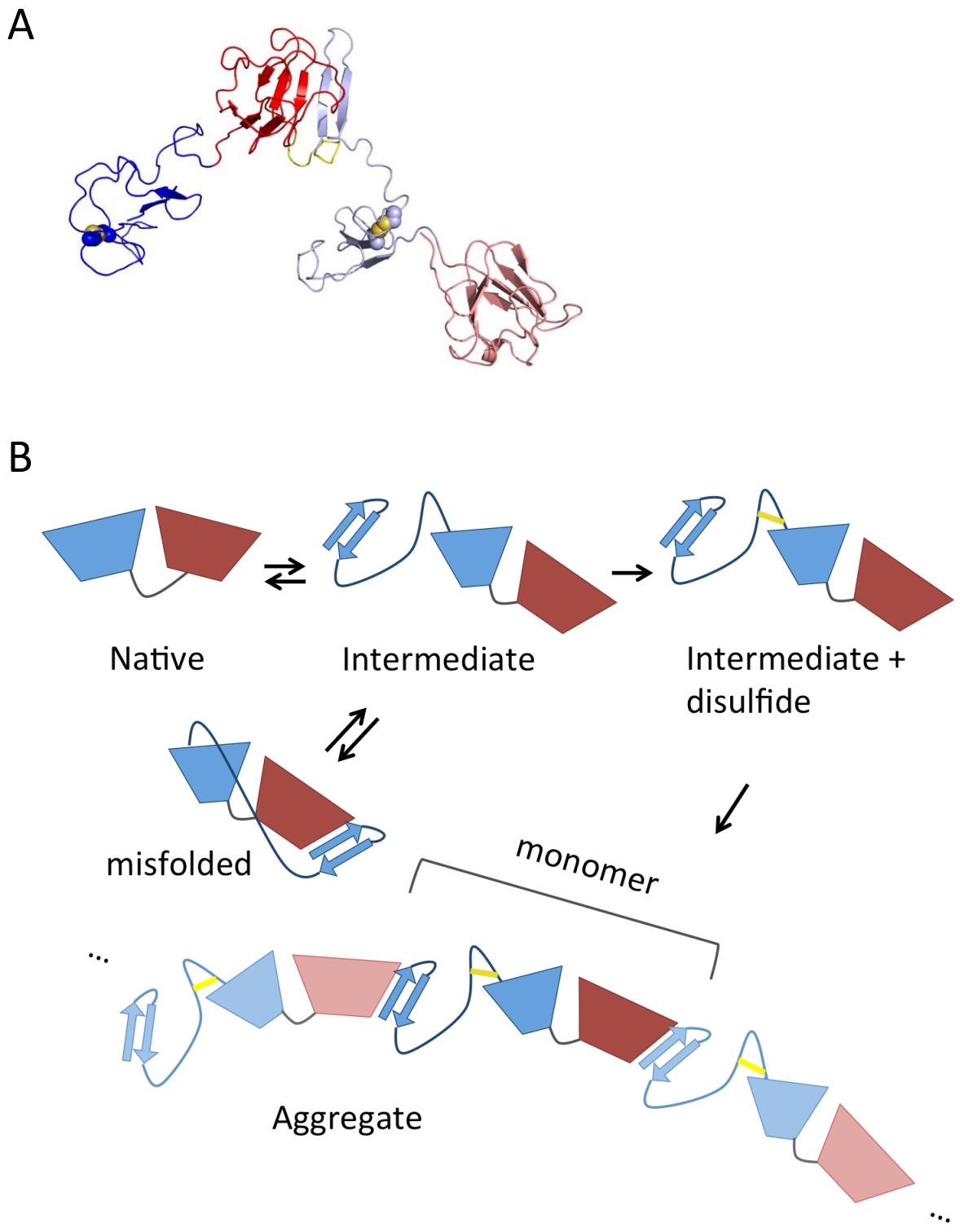



**Figure 4.10.** Model of mutant aggregation. A) Simulation structure of W42R with 2-3 disulfide, showing an antiparallel interaction between strand 1 and strand 14. B) Model of aggregation based on simulation structure.

The major unfolding events that must take place in order for strand 1 to bind to strand 14 are the separation of the N-terminal domain and C-terminal domain and extrusion of the N-terminal hairpin. These are the first events observed in unfolding simulations (Figure 4.2). Figure 4.11 A-C show that other conformational changes that take place within the C-terminal domain upon binding to the N-terminal hairpin are fairly subtle. Depicted in 4.11-D is a functional protein-protein interaction observed in nature, the PDZ domain binding to its ligand (Maisonneuve et al., 2016), which shows similarities to our proposed mechanism of γ-crystallin aggregation.

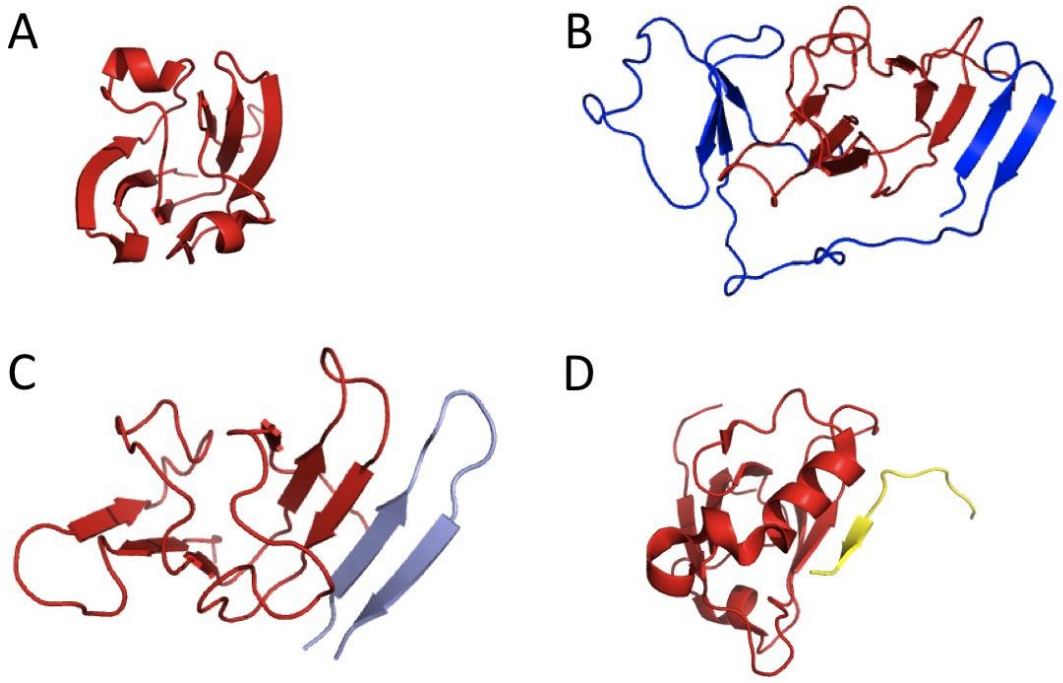



**Figure 4.11.** Mechanism of protein-protein binding for the proposed aggregation mechanism. A) Isolated C-terminal domain in the native state. B) Misfolded intermediate, with strand 1 bound to strand 14 in an intramolecular interaction. C) Protein-protein interaction with strand 1 bound to strand 14 in an intermolecular

interaction. D) PDZ domain bound to its ligand: a functional example of beta-sheet completion to facilitate protein-protein interaction.

Figure 4.12 shows the interface between strand 1 and strand 14 in the predicted oligomerizing structure, in a sample frame from simulations. R141 contacts E7, an interaction between a positively charged residue and a negatively charged residue. W156 occupies a hydrophobic surface created by residues within the two strands. The loop that contains W156 is distorted relative to the native structure.

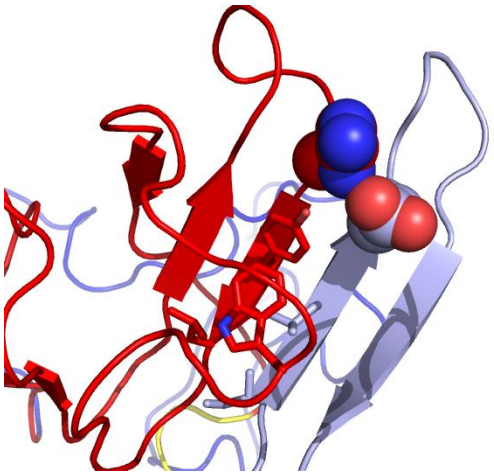


**Figure 4.12.** Interactions between strand 1 and strand 14 in a sample frame from simulations containing the predicted oligomer forming structure. E7 and R141 are shown in sphere representation, and other residues at the interface are shown in stick representation.

We further analyzed simulations to determine which residues are most often involved in protein-protein interactions, for each of the two proteins. In this analysis, we kept track of all protein-protein interactions, including those in the "native-like" and "other" categories. Results are shown in Figure 4.13. Results depend somewhat on which molecule is considered, an artifact of the two-molecule tethered approach. Many of the residues exhibiting protein-protein interactions are near the interface between the N-terminal domain and C-terminal domain, which contains many hydrophobic residues that become solvent exposed upon separation

of the two domains. The C-terminal tail, which contains a phenylalanine residue, also exhibits many interactions.

Residues within the N-terminal hairpin interact often with the other protein, particularly for the destabilized mutant. Residues F11 and L5 show many interactions, especially for molecule 2 within the W42R mutant with the 2-3 disulfide bond imposed. These may be key residues driving beta sheet completion interactions involving the N-terminal hairpin. For molecule 1, the WT protein shows few interactions involving the N-terminal hairpin, while the mutant protein and the mutant protein with the disulfide bond show more such interactions. In all, it is clear that mutation and disulfide bonding increase the amount of protein-protein interaction, and that the propensity to form interactions is residue-dependent for both the WT protein and the mutant.

A residue-residue contact map was generated for the W42R mutant containing the 2-3 disulfide, using the final frames of the 300 simulations (Figure 4.14). Strand-strand interactions are visible as short diagonal stretches within the map. Many interactions are between the N-terminal hairpin of molecule 2 and either the N-terminal hairpin of molecule 1 or regions of molecule 1 that are at the domain-domain interface of the folded molecule. In addition, several interactions are seen between the C-terminal tail of molecule 1 and residues within the N-terminal hairpin of molecule 2 are seen.

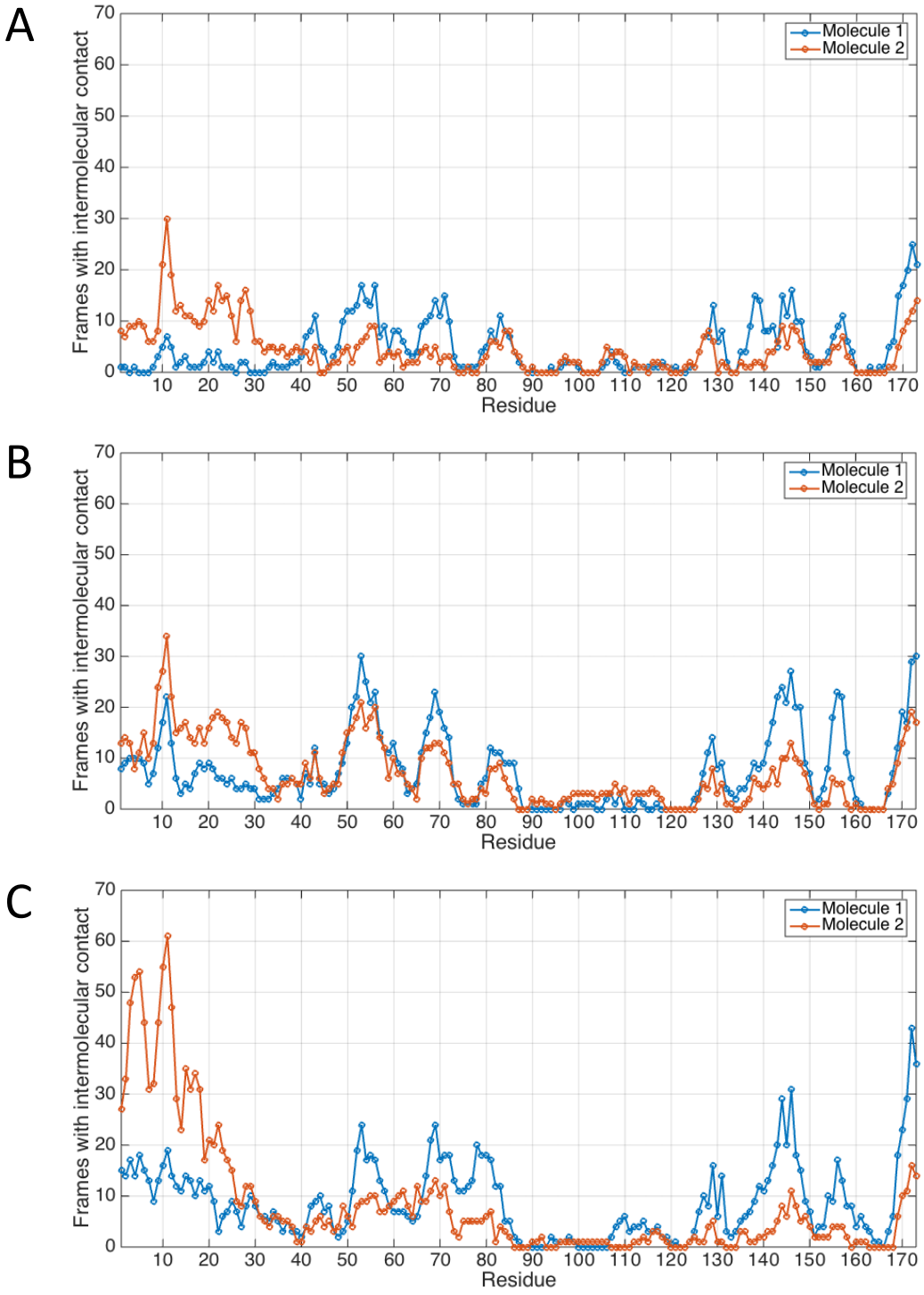



**Figure 4.13.** The propensity of each residue to exhibit intermolecular interactions. Plots show the number of frames from step 80,000,000 of 300 simulations at T = 0.8 that exhibit contacts between a given residue and any residue of the other protein. A) WT. B) W42R, no disulfide bonds. C) W42R with the 2-3 disulfide bond.

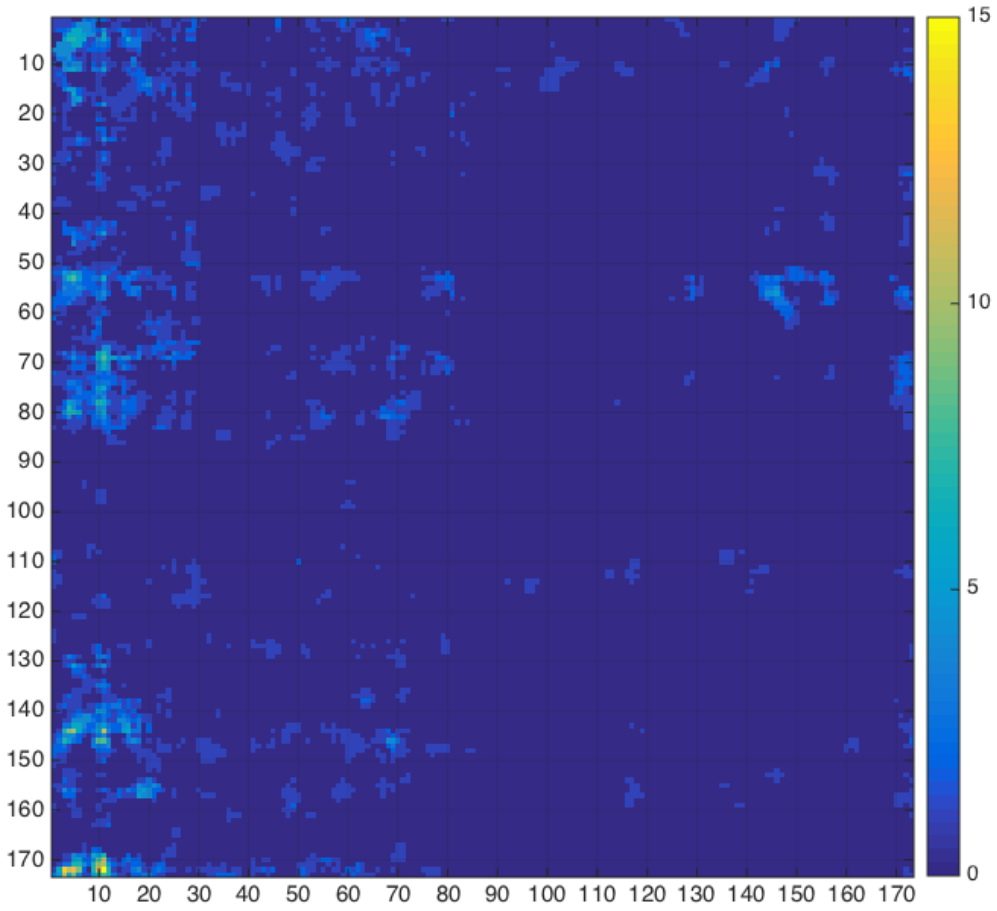



**Figure 4.14.** Residue-residue contact map for W42R 2-3 disulfide, generated from step 80,000,000 of 300 simulations.

*4.3.5. Mechanism for WT acceleration of mutant aggregation*

The WT protein was shown to promote aggregation of the W42Q mutant *in vitro* (Serebryany et al., 2016b). To investigate the mechanism of WT acceleration of mutant aggregation, we carried out simulations of one W42Q molecule linked to one WT molecule. To mimic experimental conditions, we included the 2-3 disulfide in the mutant and 5-6 disulfide (within the C-terminal domain) in the WT protein. In several simulations, we observed a structure in which the folded WT N-terminal

domain forms native-like interactions with the mutant C-terminal domain (Figure 4.15). The mutant N-terminal domain is then no longer stabilized by its C-terminal domain and is more likely to unfold. We propose that partial or full unfolding of the mutant N-terminal domain, which may induce the formation of intramolecular disulfide bonds, could then accelerate aggregation of the mutant.

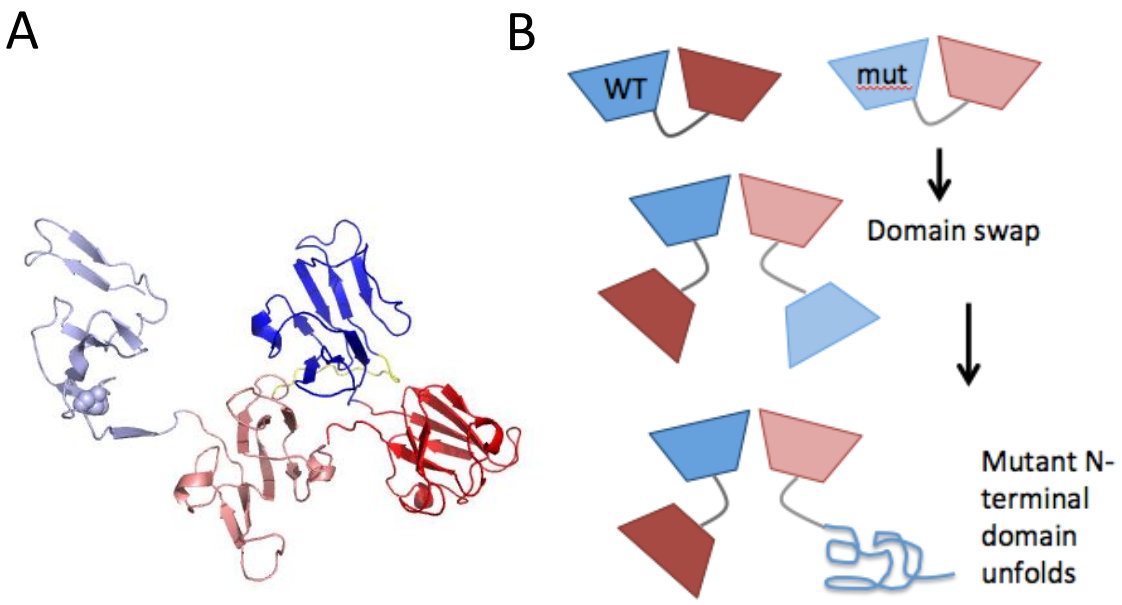



**Figure 4.15.** Model of WT acceleration of mutant aggregation. A) Structure showing interaction of WT (right, darker red/blue) with W42Q, from two-molecule simulations. B) Model of how WT might accelerate mutant N-terminal domain unfolding, based on simulation structure.

*4.4. Discussion*

Our observation that detachment of the N-terminal hairpin is an early event in unfolding simulations is interesting for several reasons. First, the N-terminal domain is thermodynamically less stable than the C-terminal domain, which may be due to a faster unfolding rate, which we see in simulations. Second, many cataract-linked mutations are at or near the interface between the N-terminal domain and

the rest of the protein (Serebryany and King, 2014). Mutation at this interface may further reduce the relatively weak interactions holding the hairpin to the rest of the domain, thereby accelerating unfolding or leading to a partially-unfolded intermediate that is prone to aggregation. Third, single molecule pulling experiments suggest that a domain swap may occur between N-terminal hairpins *in vitro* (Garcia-Manyes et al., 2016). Finally, swapping of terminal hairpins is common in domain swapped structures, and interaction between hairpins is sometimes seen in amyloid fibrils, suggesting that interactions involving terminal hairpins may be a common means of protein-protein interaction and aggregation.

Chapter 2 of this thesis describes the use of All-Atom Monte Carlo unfolding simulations to predict relative stabilities of mutants of the protein DHFR and shows that this approach yields good agreement with experiment. Our γD-crystallin simulations further validate this approach, predicting that cataract-linked mutations destabilize the domain of which they are a part. In addition to predicting stability effects of mutations, we identify folding intermediates, including misfolded structures, that may be prone to aggregation and/or promote intramolecular disulfide bonding that locks an aggregation-prone conformation into place.

We found that disulfide bonding between adjacent cysteines increased the amount of protein-protein interaction, while disulfide bonding between non-adjacent cysteines decreased the amount of protein-protein interaction. This is consistent with a previous study of disulfide bonding in lattice proteins (Abkevich and Shakhnovich, 2000). Simulations with a disulfide bond between cysteines 2 and 3, the experimentally observed disulfide bond, particularly showed a large number

of antiparallel interactions between beta strands. Analysis of simulations showed that such interactions were especially long-lived, although other types of interactions between proteins were also seen. It is interesting to note that the interfaces involved in intermolecular hydrogen bonding interactions, on strands 1, 6, and 14, are buried in the crystal structure, but become exposed early in unfolding simulations. We might expect from the unfolding simulations alone that these strands could be involved in intermolecular interactions.

The two-molecule Monte Carlo simulations used in this study contained a peptide linker (GS x6) connecting the two molecules. This linker has experimental relevance as the sequence used in single-molecule pulling experiments of γD-crystallin (Garcia-Manyes et al., 2016). Other computational studies have made use of a flexible linker to constrain the distance between interacting monomers (Levy et al., 2005; Levy et al., 2004). Alternative approaches to mimic finite protein concentration in simulations are to use periodic boundary conditions or to introduce an energy term constraining the distance between proteins. A version of the Monte Carlo program incorporating the second approach is being developed within the Shakhnovich group.

Run-away domain swapping has been proposed as a mechanism for aggregate formation (Rousseau et al., 2003). We do not observe conventional domain swapping in our simulations. However, Horwich (Horwich, 2002) hypothesized that in some cases domain swapping may act to reconstitute a kinetically-stable misfolded intermediate, rather than the native state. Given that strand 1 binds to strand 14 in some single-molecule simulations, our hypothesized

structure for the formation of aggregates is an example of this type of domain swap to reconstitute a misfolded state.

In our structure, the N-terminal hairpin extends a beta sheet already present in the C-terminal domain. A similar mechanism of beta sheet completion is seen in the binding of PDZ domains to their ligands (Maisonneuve et al., 2016). Our mechanism of aggregation is somewhat similar to that proposed by Das et al. (Das et al., 2011), in which an unfolded N-terminal domain interacts with strands 13-15. However, our mechanism involves hydrogen bonding interactions between beta strands and only partial unfolding of the N-terminal domain.

It is interesting to note the homology between strands 1 and 13, which are both the first strands of a Greek key motif. Therefore, in our predicted oligomerizing structure, strand 14 is located between two homologous strands: one from the native structure (strand 13) and one from the N-terminal hairpin of another protein (strand 1). Contacts that stabilize the aggregate are therefore similar to contacts that stabilize the native structure, a feature that is shared with domain swapping mechanisms, although in our proposed mechanism the contacts are similar but not identical to those formed in the native structure.

Future experiments will test our hypothesis of aggregate structure and the mechanism by which the WT protein accelerates mutant aggregation. Based on our studies, it may eventually be possible to design therapeutics that slow the formation of cataract. For instance, drugs may be designed that interfere with specific binding to beta strand 14 within the C-terminal domain, or which stabilize the native structure to prevent unfolding. In general, our approach may be used to study

protein aggregation involved in other diseases, leading to better understanding of disease and ultimately to novel treatments.

*Contributions*

Much of this work is described in a recent publication (Serebryany et al., 2016b). Jaie Woodard performed all Monte Carlo simulations and simulation analysis and edited the MC code to introduce an energy term dependent on the distance between sulfur atoms, representing a disulfide bond. Eugene Serebryany suggested mining single-molecule simulations to look for rearranged structures and suggested that the concept of "domain swapping" to reconstitute a misfolded intermediate may be applicable. Eugene Shakhnovich noted the relevance of disulfide bonding of adjacent and non-adjacent disulfide cysteines to a previous publication.

# 5. Conclusions

Computer simulations have become popular as a means of studying the structure, dynamics, and interactions of biological molecules. The importance of this development was recently highlighted in the awarding of the 2013 Nobel Prize in Chemistry to Martin Karplus, Michael Levitt, and Arieh Warshel, pioneers of molecular dynamics simulation and computational chemistry. As computers become more powerful, it becomes possible to simulate larger systems at increasing levels of detail, over longer time scales than ever before. However, as we develop more accurate simulation methods, there is still room to develop and to utilize more simplified models, which yield answers in a shorter amount of time and with less use of computational resources, and which may be simpler to use, to understand, and to trouble-shoot in cases where results do not agree with expectations from experimental research.

In this thesis, we pose the question of whether short unfolding simulations can be used to predict the change in stability upon mutation. Theoretically, we find that the answer is yes, given a two-state model of unfolding and the assumption of constant $\phi$ values across residues. Simulating every possible mutant of the enzyme DHFR using an all non-hydrogen atom Monte Carlo simulation program, we find a good correlation between predicted and experimental stability changes, and we are able to identify several mutations that stabilize the protein.

We take a different approach in the case of the cataract-associated protein γD-crystallin, studying selected mutants in greater detail. We find that extrusion of the N-terminal hairpin is an early event in unfolding, and that cataract-linked mutants within the N-terminal domain destabilize this domain in unfolding simulations. Current models of non-amyloid aggregation prior to our studies include: 1. aggregation by non-specific association of hydrophobic residues within unfolded protein segments, and 2. aggregation by run-away domain swapping. Based on our simulations, we propose a model distinct from these two, in which hydrogen bonding occurs between specific beta strands. The model can also be viewed as a domain swap to reconstitute a misfolded intermediate structure. We plan to work with collaborators to test this model experimentally using NMR and other techniques.

Finally, we introduce a new simplified model in which proteins can unfold and domain swap. This model allows us to predict the temperature and concentration dependence of protein-protein interactions such as functional dimerization, domain swapping, and amorphous aggregation. In particular, we recover the result that domain swapping occurs at intermediate temperature. Future theoretical and experimental studies within protein biophysics will help to reveal the relevance of domain swapping within biological systems, its contribution to the process of protein evolution, and the insights that domain swapped structures can provide into the folding pathways and folding intermediate structures of proteins.

We hope that the methods developed here will be useful in the study of other protein systems, that their application has yielded meaningful insights into the systems we have presented here, and that our *in silico* experiments, analysis, and discussion will prompt further ideas, observations, and debate within the field of protein biophysics. We believe that computational methods will continue to play an important role in the study of molecular disease, protein evolution, protein design, and the development of novel therapeutics, and we look forward to progress in the years ahead.

# References


Abeln, S., and D. Frenkel, 2008, Disordered flanks prevent peptide aggregation: PLoS Comput Biol, v. 4, p. e1000241.

Abkevich, V. I., and E. I. Shakhnovich, 2000, What can disulfide bonds tell us about protein energetics, function and folding: simulations and bioninformatics analysis: J Mol Biol, v. 300, p. 975-85.

Adamczyk, A. J., J. Cao, S. C. Kamerlin, and A. Warshel, 2011, Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions: Proc Natl Acad Sci U S A, v. 108, p. 14115-20.

Aguzzi, A., and A. M. Calella, 2009, Prions: protein aggregation and infectious diseases: Physiol Rev, v. 89, p. 1105-52.

Bartels, T., J. G. Choi, and D. J. Selkoe, 2011, α-Synuclein occurs physiologically as a helically folded tetramer that resists aggregation: Nature, v. 477, p. 107-10.

Beadle, B. M., and B. K. Shoichet, 2002, Structural bases of stability-function tradeoffs in enzymes: J Mol Biol, v. 321, p. 285-96.

Bennett, M. J., M. P. Schlunegger, and D. Eisenberg, 1995, 3D domain swapping: a mechanism for oligomer assembly: Protein Sci, v. 4, p. 2455-68.

Bershtein, S., W. Mu, A. W. Serohijos, J. Zhou, and E. I. Shakhnovich, 2013, Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness: Mol Cell, v. 49, p. 133-44.

Bershtein, S., W. Mu, W. Wu, and E. I. Shakhnovich, 2012, Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations: Proc Natl Acad Sci U S A, v. 109, p. 4857-62.

Bloemendal, H., W. de Jong, R. Jaenicke, N. H. Lubsen, C. Slingsby, and A. Tardieu, 2004, Ageing and vision: structure, stability and function of lens crystallins: Prog Biophys Mol Biol, v. 86, p. 407-85.

Bloom, J. D., S. T. Labthavikul, C. R. Otey, and F. H. Arnold, 2006, Protein stability promotes evolvability: Proc Natl Acad Sci U S A, v. 103, p. 5869-74.

Braselmann, E., J. L. Chaney, and P. L. Clark, 2013, Folding the proteome: Trends Biochem Sci, v. 38, p. 337-44.

Bryan, P. N., and J. Orban, 2010, Proteins that switch folds: Curr Opin Struct Biol, v. 20, p. 482-8.

Chirgadze, D. Y., M. Demydchuk, M. Becker, S. Moran, and M. Paoli, 2004, Snapshot of protein structure evolution reveals conservation of functional dimerization through intertwined folding: Structure, v. 12, p. 1489-94.

Chiti, F., and C. M. Dobson, 2009, Amyloid formation by globular proteins under native conditions: Nat Chem Biol, v. 5, p. 15-22.

Danielsson, J., X. Mu, L. Lang, H. Wang, A. Binolfi, F. X. Theillet, B. Bekei, D. T. Logan, P. Selenko, H. Wennerström, and M. Oliveberg, 2015, Thermodynamics of protein destabilization in live cells: Proc Natl Acad Sci U S A, v. 112, p. 12402-7.

Das, P., J. A. King, and R. Zhou, 2011, Aggregation of γ-crystallins associated with human cataracts via domain swapping at the C-terminal β-strands: Proc Natl Acad Sci U S A, v. 108, p. 10514-9.

Deeds, E. J., O. Ashenberg, J. Gerardin, and E. I. Shakhnovich, 2007, Robust protein protein interactions in crowded cellular environments: Proc Natl Acad Sci U S A, v. 104, p. 14952-7.

DePristo, M. A., D. M. Weinreich, and D. L. Hartl, 2005, Missense meanderings in sequence space: a biophysical view of protein evolution: Nat Rev Genet, v. 6, p. 678-87.

Ding, F., N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, 2002, Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism: J Mol Biol, v. 324, p. 851-7.

Ding, F., K. C. Prutzman, S. L. Campbell, and N. V. Dokholyan, 2006, Topological determinants of protein domain swapping: Structure, v. 14, p. 5-14.

Dobson, C. M., 2003, Protein folding and misfolding: Nature, v. 426, p. 884-90.

Drummond, D. A., and C. O. Wilke, 2008, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution: Cell, v. 134, p. 341-52.

Edgar, R. C., 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput: Nucleic Acids Res, v. 32, p. 1792-7.

Elcock, A. H., 2010, Models of macromolecular crowding effects and the need for quantitative comparisons with experiment: Curr Opin Struct Biol, v. 20, p. 196-206.

Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali, 2006, Comparative protein structure modeling using Modeller: Curr Protoc Bioinformatics, v. Chapter 5, p. Unit 5.6.

Fan, X., S. Zhou, B. Wang, G. Hom, M. Guo, B. Li, J. Yang, D. Vaysburg, and V. M. Monnier, 2015, Evidence of Highly Conserved β-Crystallin Disulfidome that Can be Mimicked by In Vitro Oxidation in Age-related Human Cataract and Glutathione Depleted Mouse Lens: Mol Cell Proteomics, v. 14, p. 3211-23.

Fersht, A. R., and S. Sato, 2004, Phi-value analysis and the nature of protein-folding transition states: Proc Natl Acad Sci U S A, v. 101, p. 7976-81.

Fiser, A., and A. Sali, 2003, ModLoop: automated modeling of loops in protein structures: Bioinformatics, v. 19, p. 2500-1.

Flaugh, S. L., M. S. Kosinski-Collins, and J. King, 2005a, Contributions of hydrophobic domain interface interactions to the folding and stability of human gammaD-crystallin: Protein Sci, v. 14, p. 569-81.

Flaugh, S. L., M. S. Kosinski-Collins, and J. King, 2005b, Interdomain side-chain interactions in human gammaD crystallin influencing folding and stability: Protein Sci, v. 14, p. 2030-43.

Flaugh, S. L., I. A. Mills, and J. King, 2006, Glutamine deamidation destabilizes human gammaD-crystallin and lowers the kinetic barrier to unfolding: J Biol Chem, v. 281, p. 30782-93.

Flory, P. J., 1953, Principles of Polymer Chemistry, Ithaca, Cornell University Press.

Garcia-Manyes, S., D. Giganti, C. L. Badilla, A. Lezamiz, J. Perales-Calvo, A. E. Beedle, and J. M. Fernández, 2016, Single-molecule Force Spectroscopy Predicts a Misfolded, Domain-swapped Conformation in human γD-Crystallin Protein: J Biol Chem, v. 291, p. 4226-35.

Goldstein, R. A., 2011, The evolution and evolutionary consequences of marginal thermostability in proteins: Proteins, v. 79, p. 1396-407.

Gronenborn, A. M., 2009, Protein acrobatics in pairs--dimerization via domain swapping: Curr Opin Struct Biol, v. 19, p. 39-49.

Guo, Z., and D. Eisenberg, 2006, Runaway domain swapping in amyloid-like fibrils of T7 endonuclease I: Proc Natl Acad Sci U S A, v. 103, p. 8042-7.

Horwich, A., 2002, Protein aggregation in disease: a role for folding intermediates forming specific multimeric interactions: J Clin Invest, v. 110, p. 1221-32.

Khan, S., and M. Vihinen, 2010, Performance of protein stability predictors: Hum Mutat, v. 31, p. 675-84.

Kuznetsova, I. M., B. Y. Zaslavsky, L. Breydo, K. K. Turoverov, and V. N. Uversky, 2015, Beyond the excluded volume effects: mechanistic complexity of the crowded milieu: Molecules, v. 20, p. 1377-409.

Levy, E. D., S. De, and S. A. Teichmann, 2012, Cellular crowding imposes global constraints on the chemistry and evolution of proteomes: Proc Natl Acad Sci U S A, v. 109, p. 20461-6.

Levy, Y., S. S. Cho, J. N. Onuchic, and P. G. Wolynes, 2005, A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes: J Mol Biol, v. 346, p. 1121-45.

Levy, Y., P. G. Wolynes, and J. N. Onuchic, 2004, Protein topology determines binding mechanism: Proc Natl Acad Sci U S A, v. 101, p. 511-6.

Li, M. S., D. K. Klimov, J. E. Straub, and D. Thirumalai, 2008, Probing the mechanisms of fibril formation using lattice models: J Chem Phys, v. 129, p. 175101.

Liberles, D. A., S. A. Teichmann, I. Bahar, U. Bastolla, J. Bloom, E. Bornberg-Bauer, L. J. Colwell, A. P. de Koning, N. V. Dokholyan, J. Echave, A. Elofsson, D. L. Gerloff, R. A. Goldstein, J. A. Grahnen, M. T. Holder, C. Lakner, N. Lartillot, S. C. Lovell, G. Naylor, T. Perica, D. D. Pollock, T. Pupko, L. Regan, A. Roger, N. Rubinstein, E. Shakhnovich, K. Sjölander, S. Sunyaev, A. I. Teufel, J. L. Thorne, J. W. Thornton, D. M. Weinreich, and S. Whelan, 2012, The interface of protein structure, protein biophysics, and molecular evolution: Protein Sci, v. 21, p. 769-85.

Liu, Y., and D. Eisenberg, 2002, 3D domain swapping: as domains continue to swap: Protein Sci, v. 11, p. 1285-99.

Lobkovsky, A. E., Y. I. Wolf, and E. V. Koonin, 2010, Universal distribution of protein evolution rates as a consequence of protein folding physics: Proc Natl Acad Sci U S A, v. 107, p. 2983-8.

Luke, K. A., C. L. Higgins, and P. Wittung-Stafshede, 2007, Thermodynamic stability and folding of proteins from hyperthermophilic organisms: FEBS J, v. 274, p. 4023-33.

MacKinnon, S. S., and S. J. Wodak, 2015, Landscape of intertwined associations in multi-domain homo-oligomeric proteins: J Mol Biol, v. 427, p. 350-70.

Maisonneuve, P., C. Caillet-Saguy, M. C. Vaney, E. Bibi-Zainab, K. Sawyer, B. Raynal, A. Haouz, M. Delepierre, M. Lafon, F. Cordier, and N. Wolff, 2016, Molecular Basis of the Interaction of the Human Protein Tyrosine Phosphatase Non-receptor Type 4 (PTPN4) with the Mitogen-activated Protein Kinase p38γ: J Biol Chem, v. 291, p. 16699-708.

Matouschek, A., J. T. Kellis, L. Serrano, and A. R. Fersht, 1989, Mapping the transition state and pathway of protein folding by protein engineering: Nature, v. 340, p. 122-6.

Michael, R., and A. J. Bron, 2011, The ageing lens and cataract: a model of normal and pathological ageing: Philos Trans R Soc Lond B Biol Sci, v. 366, p. 1278-92.

Mirny, L., and E. Shakhnovich, 2001, Protein folding theory: from lattice to all-atom models: Annu Rev Biophys Biomol Struct, v. 30, p. 361-96.

Naganathan, A. N., and V. Muñoz, 2010, Insights into protein folding mechanisms from large scale analysis of mutational effects: Proc Natl Acad Sci U S A, v. 107, p. 8611-6.

O'Neill, J. W., D. E. Kim, K. Johnsen, D. Baker, and K. Y. Zhang, 2001, Single-site mutations induce 3D domain swapping in the B1 domain of protein L from Peptostreptococcus magnus: Structure, v. 9, p. 1017-27.

Ohmae, E., Y. Miyashita, S. Tate, K. Gekko, S. Kitazawa, R. Kitahara, and K. Kuwajima, 2013, Solvent environments significantly affect the enzymatic function of Escherichia coli dihydrofolate reductase: comparison of wild-type protein and active-site mutant D27E: Biochim Biophys Acta, v. 1834, p. 2782-94.

Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, 2005, Scalable molecular dynamics with NAMD: J Comput Chem, v. 26, p. 1781-802.

Potapov, V., M. Cohen, and G. Schreiber, 2009, Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details: Protein Eng Des Sel, v. 22, p. 553-60.

Privalov, P. L., 1979, Stability of proteins: small globular proteins: Adv Protein Chem, v. 33, p. 167-241.

Rousseau, F., J. W. Schymkowitz, and L. S. Itzhaki, 2003, The unfolding story of three-dimensional domain swapping: Structure, v. 11, p. 243-51.

Sali, A., E. Shakhnovich, and M. Karplus, 1994, How does a protein fold?: Nature, v. 369, p. 248-51.

Serebryany, E., and J. A. King, 2014, The βγ-crystallins: native state stability and pathways to aggregation: Prog Biophys Mol Biol, v. 115, p. 32-41.

Serebryany, E., T. Takata, E. Erickson, N. Schafheimer, Y. Wang, and J. A. King, 2016a, Aggregation of Trp > Glu point mutants of human gamma-D crystallin provides a model for hereditary or UV-induced cataract: Protein Sci, v. 25, p. 1115-28.

Serebryany, E., J. C. Woodard, B. V. Adkar, M. Shabab, J. A. King, and E. I. Shakhnovich, 2016b, An Internal Disulfide Locks a Misfolded Aggregation-prone Intermediate in Cataract-linked Mutants of Human γD-Crystallin: J Biol Chem, v. 291, p. 19172-83.

Serohijos, A. W., and E. I. Shakhnovich, 2014, Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics: Curr Opin Struct Biol, v. 26, p. 84-91.

Shakhnovich, E., 2006, Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet: Chem Rev, v. 106, p. 1559-88.

Shakhnovich, E. I., and A. V. Finkelstein, 1989, Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition: Biopolymers, v. 28, p. 1667-80.

Shakhnovich, E. I., and A. M. Gutin, 1993, Engineering of stable and fast-folding sequences of model proteins: Proc Natl Acad Sci U S A, v. 90, p. 7195-9.

Speed, M. A., D. I. Wang, and J. King, 1996, Specific aggregation of partially folded polypeptide chains: the molecular basis of inclusion body composition: Nat Biotechnol, v. 14, p. 1283-7.

Straub, J. E., and D. Thirumalai, 2011, Toward a molecular theory of early and late events in monomer to amyloid fibril formation: Annu Rev Phys Chem, v. 62, p. 437-63.

Studer, R. A., P. A. Christin, M. A. Williams, and C. A. Orengo, 2014, Stability-activity tradeoffs constrain the adaptive evolution of RubisCO: Proc Natl Acad Sci U S A, v. 111, p. 2223-8.

Szymańska, A., E. Jankowska, M. Orlikowska, I. Behrendt, P. Czaplewska, and S. Rodziewicz-Motowidło, 2012, Influence of point mutations on the stability, dimerization, and oligomerization of human cystatin C and its L68Q variant: Front Mol Neurosci, v. 5, p. 82.

Taverna, D. M., and R. A. Goldstein, 2002, Why are proteins marginally stable?: Proteins, v. 46, p. 105-9.

Thiltgen, G., and R. A. Goldstein, 2012, Assessing predictors of changes in protein stability upon mutation using self-consistency: PLoS One, v. 7, p. e46084.

Tian, J., J. C. Woodard, A. Whitney, and E. I. Shakhnovich, 2015, Thermal stabilization of dihydrofolate reductase using monte carlo unfolding simulations and its functional consequences: PLoS Comput Biol, v. 11, p. e1004207.

Tokuriki, N., F. Stricher, J. Schymkowitz, L. Serrano, and D. S. Tawfik, 2007, The stability effects of protein mutations appear to be universally distributed: J Mol Biol, v. 369, p. 1318-32.

Tokuriki, N., and D. S. Tawfik, 2009, Chaperonin overexpression promotes genetic variation and enzyme evolution: Nature, v. 459, p. 668-73.

Vottariello, F., E. Giacomelli, R. Frasson, N. Pozzi, V. De Filippis, and G. Gotte, 2011, RNase A oligomerization through 3D domain swapping is favoured by a residue located far from the swapping domains: Biochimie, v. 93, p. 1846-57.

Wang, B., C. Yu, Y. B. Xi, H. C. Cai, J. Wang, S. Zhou, Y. Wu, Y. B. Yan, X. Ma, and L. Xie, 2011, A novel CRYGD mutation (p.Trp43Arg) causing autosomal dominant congenital cataract in a Chinese family: Hum Mutat, v. 32, p. E1939-47.

Wodak, S. J., A. Malevanets, and S. S. MacKinnon, 2015, The Landscape of Intertwined Associations in Homooligomeric Proteins: Biophys J, v. 109, p. 1087-100.

Woodard, J. C., S. Dunatunga, and E. I. Shakhnovich, 2016, A Simple Model of Protein Domain Swapping in Crowded Cellular Environments: Biophys J, v. 110, p. 2367-76.

Wylie, C. S., and E. I. Shakhnovich, 2011, A biophysical protein folding model accounts for most mutational fitness effects in viruses: Proc Natl Acad Sci U S A, v. 108, p. 9916-21.

Xu, J., L. Huang, and E. I. Shakhnovich, 2011, The ensemble folding kinetics of the FBP28 WW domain revealed by an all-atom Monte Carlo simulation in a knowledge-based potential: Proteins, v. 79, p. 1704-14.

Yang, J. S., W. W. Chen, J. Skolnick, and E. I. Shakhnovich, 2007, All-atom ab initio folding of a diverse set of proteins: Structure, v. 15, p. 53-63.

Yang, J. S., S. Wallin, and E. I. Shakhnovich, 2008, Universality and diversity of folding mechanics for three-helix bundle proteins: Proc Natl Acad Sci U S A, v. 105, p. 895-900.

Yang, S., S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic, 2004, Domain swapping is a consequence of minimal frustration: Proc Natl Acad Sci U S A, v. 101, p. 13786-91.

Zeldovich, K. B., P. Chen, and E. I. Shakhnovich, 2007, Protein stability imposes limits on organism complexity and speed of molecular evolution: Proc Natl Acad Sci U S A, v. 104, p. 16152-7.