Investigating the impact of small molecule ligands and the proteostasis network on protein folding inside the cell

Karan Hingorani

University of Massachusetts - Amherst
INVESTIGATING THE IMPACT OF SMALL MOLECULE LIGANDS AND THE PROTEOSTASIS NETWORK ON PROTEIN FOLDING INSIDE THE CELL

A Dissertation Presented

by

KARAN S. HINGORANI

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2016

Molecular and Cellular Biology
INVESTIGATING THE IMPACT OF SMALL MOLECULE LIGANDS AND THE PROTEOSTASIS NETWORK ON PROTEIN FOLDING INSIDE THE CELL

A Dissertation Presented

by

KARAN S. HINGORANI

Approved as to style and content by:

________________________________________
Dr. Lila M. Gierasch, Chair

________________________________________
Dr. Scott C. Garman, Member

________________________________________
Dr. Peter Chien, Member

________________________________________
Dr. Lynmarie K Thompson, Member

________________________________________
Dr. Dominique Alfandari, Chair,
Molecular and Cellular Biology Program
DEDICATION

To all the teachers who taught me how to think and not what to think, and made learning fun along the way...
ACKNOWLEDGMENTS

I owe everything in my life, which includes work done during my time in graduate school to incalculable mathematics/ arbitrary chance/luck/fate/grace or any version thereof. Without that element I would be narrating a very different story. It is because of this random probability that I have come across the fantastic people who have deeply influenced my life.

I am first and foremost very grateful for the support provided by my parents and my first teachers Neela and Sunil Hingorani who have tirelessly encouraged and continue to do so, all my pursuits regardless of outcome. I would like to thank Manju Hingorani, my aunt and my doorway into the fascinating world of molecular biology. Her love of science and exploration are infectious and my experiences with her led me to the single most critical person who has played a major formative role in my graduate career – Lila Gierasch. She has been a mentor par excellence and has been my steadfast guide from the beginning to the end. I would like to thank her (and I don’t think there is an upper limit on this) for the time, patience and the critical thinking lessons she has taught me.

I would next like to thank my families – not necessarily defined by genetics. The first family would be the one who has known me since childhood, particularly my grandparents and aunts, uncles and cousins who have been exemplary role models and people who have directly or indirectly influenced the path I have chosen for my life. They have shown by example what it is to lead a fulfilling life and contribute back to the
society one belongs to. My second family would be my lab family at the University of Massachusetts, Amherst. The Gierasch lab has been my home away from home for the past several years and I have had the privilege to work with extraordinary people throughout my stay here. Many of the lessons I have learnt in graduate school have come from working in this stimulating environment. My third family would be my closest friends who have stuck with me through life and graduate school. Even though a few of them do not entirely understand what I do or why I am a student at this age, they offer support without reservation and continue to be a source of joy in my life.

Next, I would like to thank my teachers, Nandita Mangalore, Prashant Ratnaparkhi and Samuel Namdas who instilled a love for biochemistry during my undergraduate studies. I am grateful to Derek Wilson for taking me on during my year in Toronto and introducing me to biological mass spectrometry. To my committee members Scott Garman, Peter Chien and Lynmarie Thompson I thank them for being critical and approachable during my stay at graduate school, sharing thoughts and ideas hasn’t been more fun! I am indebted to Evan Powers, who showed me the true power of combining mathematics with biology to understand fundamental processes. I would credit him second only to Lila for the development of my thought and work during graduate school.

Finally, I would like to acknowledge that because of the tremendous amount of luck I have had through life and graduate school, I am duty bound to help others (using my training) and spread some of the good fortune I have received.
ABSTRACT

INVESTIGATING THE IMPACT OF SMALL MOLECULES LIGANDS AND THE PROTEOSTASIS NETOWRK ON PROTEIN FOLDING INSIDE THE CELL

SEPTEMBER 2016

KARAN S. HINGORANI

B.Sc., St. XAVIER’S COLLEGE, UNIVERSITY OF MUMBAI

PhD., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lila M. Gierasch

The folded forms of most proteins are critical to their functions. Despite the complexity of the cellular milieu and the presence of high-risk deleterious interactions, there is a high level of fidelity observed in the folding process for entire proteomes. Two important reasons for this are the presence of the quality control machinery consisting of chaperones and degradation enzymes that work jointly to optimize the population of the folded state and interaction partners that re-enforce the functional state and add to the competitive advantage of an organism. While substantial effort has been directed to understand protein folding and interactions in vitro, comparatively little of these processes are explored inside the cell. This work examines two important aspects of protein folding inside the cell: first, the impact of small molecule ligands on protein folding; and second, the impact of the proteostasis network on the folding of an obligatory chaperone client. We deploy a combination of experiments and mathematical modeling based on the principle of kinetic partitioning to understand how these
phenomena sculpt the protein folding landscape inside the cell. We find that ligands specifically deplete unfolded and aggregation- or degradation - prone protein populations by favoring the folded state and the chaperone and degradation proteins work to minimize off-pathway species thus reducing the population of aggregated protein inside the cell.
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1 Chaperone concentrations upon induction</td>
<td>168</td>
</tr>
<tr>
<td>A.2 Concentrations and $F_{agg}$ for MetK synthesized at varying IPTG concentrations</td>
<td>169</td>
</tr>
<tr>
<td>A.3 Concentrations and $F_{agg}$ for MetK synthesized at varying IPTG concentrations at varying DnaKJE concentrations</td>
<td>170</td>
</tr>
<tr>
<td>A.4 Concentrations and $F_{agg}$ for MetK synthesized at varying IPTG concentrations at varying GroELS concentrations</td>
<td>171</td>
</tr>
<tr>
<td>A.5 Concentrations and $F_{agg}$ for MetK synthesized at varying IPTG concentrations at varying Lon concentrations</td>
<td>172</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>Figure 1.1 Cartoon representation of the complexity found in an E. coli cell where protein folding occurs</td>
<td>6</td>
</tr>
<tr>
<td>Figure 1.2 Major components of the folding and degradation branches of the proteostasis network in E. coli</td>
<td>8</td>
</tr>
<tr>
<td>Figure 1.3 Schematic for a coupled folding and aggregation landscape</td>
<td>11</td>
</tr>
<tr>
<td>Figure 1.4 Evolutionary pressures changes the protein folding landscape inside the cell</td>
<td>22</td>
</tr>
<tr>
<td>Figure 1.5 General schematic for the fates and the kinetic partitioning of a polypeptide chain inside the cell</td>
<td>30</td>
</tr>
<tr>
<td>Figure 1.6 The FoldEco model of proteostasis and kinetic partitioning in E. coli</td>
<td>32</td>
</tr>
<tr>
<td>Figure 2.1 dDHFR is a destabilized test protein the solubility of which is improved upon addition of a high affinity ligand trimethoprim</td>
<td>42</td>
</tr>
<tr>
<td>Figure 2.2 Chloramphenicol shut-off experiments indicate the effective irreversibility of ligand binding and aggregation</td>
<td>46</td>
</tr>
<tr>
<td>Figure 2.3 General schematic of the kinetic partitioning influenced by ligand binding</td>
<td>52</td>
</tr>
<tr>
<td>Figure 2.4 Titration of TMP shows that dDHFR F_r scales with concentration of TMP and plateaus as a function of its intrinsic folding and aggregation rate constants</td>
<td>54</td>
</tr>
<tr>
<td>Figure 2.5 Titration of DGJ shows that α-Gal F_r scales with concentration of DGJ and plateaus as a function of its intrinsic folding and degradation rate constants</td>
<td>57</td>
</tr>
<tr>
<td>Figure 2.6 Other prominent studies showing how natural and artificial ligands bias the folding – aggregation - degradation kinetic competition for other proteins</td>
<td>61</td>
</tr>
<tr>
<td>Figure 3.1 Hypotheses for chaperone functions based on computational modeling</td>
<td>73</td>
</tr>
<tr>
<td>Figure 3.2 Model test protein MetK</td>
<td>77</td>
</tr>
<tr>
<td>Figure 3.3 Equilibrium denaturation properties of MetK</td>
<td>81</td>
</tr>
<tr>
<td>Figure 3.4 Synthesis and aggregation of MetK at basal chaperone conditions</td>
<td>86</td>
</tr>
<tr>
<td>Figure 3.5 Impact of GroEL expression on MetK synthesis and aggregation</td>
<td>92</td>
</tr>
<tr>
<td>Figure 3.6 Impact of DnaKJE on the folding and aggregation of MetK</td>
<td>97</td>
</tr>
</tbody>
</table>
Figure 3.7 Impact of ion on MetK synthesis and aggregation. ....................... 102
Figure 3.8 Concentrations of chaperones produced as a function of arabinose concentrations. ........................................................................................................ 104
Figure 3.9 The kinetic partitioning of MetK upon synthesis in E. coli .......... 109
Figure 3.10 MetK assembly might be independent of monomer folding ...... 112
Figure 4.1 General design for a MEF system ............................................. 121
Figure 4.2 Equilibrium stability of DHFR constructs with modified surface loops ......................................................................................................................... 125
Figure 4.3 Utility of FlAsH binding for stability measurements............... 131
Figure 4.4 Schematic of the MEF system deployed inside cells ............... 137
Figure 4.5 In vitro and in vivo demonstrations of MEF constructs .......... 140
Proteins are essential biomolecules that execute vital cellular functions. For most proteins to function correctly, they must fold into their correct three-dimensional shape and interact appropriately with their partners. While much is known about protein folding in \textit{in vitro} systems, there remains much work to be done to understand how protein folding occurs with fidelity inside the cell. In this chapter I will discuss some current progress on the 1) status of protein folding research \textit{in vitro}, 2) how protein folding inside the cell can be vastly different from folding inside the test tube and the factors that can contribute to protein folding \textit{in vivo}, 3) the concept of kinetic partitioning which can be used to understand how the differences \textit{in vitro} and \textit{in vivo} manifest in the folding reaction and finally 4) the questions that I have tried to address during the course of my thesis which are directed at understanding certain aspects of the cellular environment that influence protein folding. Sections of this chapter were co-authored with my advisor and published in the review “Comparing protein folding \textit{in vitro} and \textit{in vivo}: Foldability meets the fitness challenge” Current Opinion in Structural Biology February 2014.

\textbf{1.1. Protein folding \textit{in vitro}:}

In the sixty-odd years since Anfinsen’s pioneering work showing the ability of RNaseA to re-fold from a reductively denatured state (Anfinsen, 1973), the mechanism of protein folding and how amino acid sequence encodes a folding reaction have been extensively studied (Daggett and Fersht, 2003; Dill et al., 2008). Increasingly powerful experimental
and computational methods have been focused on the intellectually seductive \textit{in vitro} ‘protein folding problem’, defined to be the relationship between amino acid sequence and the adoption of a native fold. As a consequence, we know a great deal about protein folding, but our knowledge is largely confined to how a protein folds at high dilution in conditions that are optimized for folding success.

Small fast-folding domains have been the subject of extensive in-depth study \textit{in vitro} because they are amenable to detailed physico-chemical analysis. For multiple reasons, it might be anticipated that the intrinsic folding behavior of these domains will solely determine their \textit{in-vivo} folding properties. They fold on time scales (Daggett and Fersht, 2003; Hingorani and Gierasch, 2014; Zhang and Ignatova, 2011) that are much faster \textit{(e.g., microseconds to milliseconds)} than co-translational events \textit{(rate of synthesis in eukaryotes 5 amino acids/sec, or 15 amino acids/sec in \textit{E. coli})}. Also, they generally do not populate long-lived intermediates and do not present extensive hydrophobic surfaces—both necessary for binding interactions with chaperones. If such domains retain their intrinsic properties and their properties are not dominated by context, they may be viewed as the “atomic particles” of well-folded proteins in the cell. This view would allow researchers to treat large proteins as composites of smaller domains and, if parsed into even smaller units, ‘foldons’ (Maity et al., 2005). Thus, the insights on folding of small domains provided by ever more powerful experimental methods and impressive new computational capabilities may be applicable to explaining folding \textit{in vivo}. For example, the description of transition path times using single-molecule Förster resonance energy transfer (FRET) sheds light on timescales of fundamental folding.
events, and the promising simulations of folding at realistic timescales have afforded the opportunity to compare experiment and theory directly (Chung et al., 2012; Lindorff-Larsen et al., 2011; Piana et al., 2013). Analysis of the folding trajectories computed by Anton, the supercomputer designed for protein folding simulations, offered a unifying mechanism for a dozen proteins and suggested that native-like contacts are formed in the unfolded state, with successive stabilization of key contacts driving the folding reaction. Progress in the simulation of folding reactions has also been reported by the Pande group, who showed that Markov state modeling and molecular dynamics (MD) simulations, accessible without Anton, generate kinetically relevant folding trajectories (Pande et al., 2010). These researchers found evidence of glass-like kinetics using Markov state modeling of folding (Weber et al., 2013). Computational studies such as these may provide a possible bridge from theory to experiment.

However, small single domain proteins are quite rare; for example, they represent less than 15% of the E. coli proteome (Braselmann et al., 2013). Moreover, recent work from several groups suggests that, even though large proteins can generally be broken down into smaller units by domain dissection, the folding of these component domains may not be independent, and thus what is seen for free-standing small-domain proteins may not be applicable to the universe of larger proteins in the proteomes of all organisms. Specifically, the domains of repeat proteins have been found to display context-dependent folding (Sawyer et al., 2013; Vieux and Barrick, 2011). In addition, the coupling of domains of large proteins is often a key part of the function of the large protein (Ferreiro et al., 2011).
Thus it will be necessary to push the envelope of in vitro approaches and tackle larger proteins to understand the folding mechanisms of a greater population of several proteomes. Some recent research has taken on this challenge, and results show how new complexities in folding landscapes will emerge when larger proteins are examined: Pirchi et al. deployed single molecule FRET coupled to hidden Markov analysis to uncover six metastable states and multiple folding routes along the folding landscape of adenylate kinase, a three-domain 23.5 kDa protein (Pirchi et al., 2011). The Rief laboratory used optical tweezer pulling experiments and hidden Markov analysis to study the folding of calmodulin and observed four on-pathway intermediates along with two off-pathway intermediates that compete with the productive folding reaction (Stigler et al., 2011). Dahiya and Chaudhuri (Dahiya and Chaudhuri, 2013) examined the folding of the 82 kDa multi-domain protein, malate synthase G, and concluded that weak interdomain cooperativity may add complexity to a folding pathway, including the possibility of a functional intermediate.

Another research topic in in vitro folding that has seen impressive progress recently is the nature of the denatured or unfolded state ensemble and under what conditions this ensemble collapses (Haran, 2012; Meng et al., 2013a; Meng et al., 2013b). A contentious point has emerged regarding how collapsed the unfolded state ensemble is under differing denaturant concentrations, and it now seems that the apparent results depend on the method of observation (Yoo et al., 2012). In any case, it remains unclear whether and when a polypeptide freely explores the unfolded state in vivo, apart from intrinsically
disordered proteins. Domains may transiently unfold or populate non-native states as they interact with chaperones (see below), and molecular machines that facilitate either translocation across membranes or degradation likely actively unfold proteins (Baker and Sauer, 2012; Tomkiewicz et al., 2007). Thus, the connections between non-native, unfolded states in vitro and in vivo should continue to be explored.

1.2. Protein folding in vivo – what’s different?

Proteins initially fold in vivo upon their biosynthesis. Hence, the first environment they are subjected to is created by the ribosome and ribosomally associated chaperone proteins and enzymes. In addition, chains may fold co-translationally before the entire chain has been made. In contrast, folding of proteins in vitro initiates from an unfolded ensemble in which a population of full-length chains (or in the case of single molecule experiments, one polypeptide) is subjected to folding conditions. Thus, the possibility of co-translational folding constitutes a major difference between the de novo folding reaction in the test tube and in almost all cells of living organisms.

Second, in vitro proteins sample their unfolded state in a dynamic equilibrium governed by their thermodynamic stability. Whether proteins spend much time unfolded in vivo is unclear. Many factors may disfavor accumulation of any significant population in the unfolded state, including chaperone binding, ongoing degradation, and kinetic barriers. Nonetheless, there may be lessons to be drawn from in vitro studies of unfolded states.
Third, protein-folding experiments in vitro are done at high dilution. In vivo, macromolecule concentrations range from 200 to 400 mg/ml, and highly interactive surfaces are present all around a folding chain are highly interactive. Thus, the impact of crowding and the influence of protein-protein interactions, including weak “Quinary” interactions (McConkey, 1982), must be taken into account (Figure 1.1).

Figure 1.1 Cartoon representation of the complexity found in an E. coli cell where protein folding occurs

The figure depicts computational modeling by McGuffee and Elcock to capture the crowded and heterogeneous environment of the E. coli cytoplasm. The modeling was performed with 50 of the most abundant proteins in E. coli at atomistic detail (Plos comp boil). Crowding in E. coli represents the high density of macromolecules and co-solutes (discussed in the macromolecular crowding section in this chapter) that have co-evolved with each other over the evolutionary time course. The high concentration of macromolecules necessarily forces weak and transient interactions between them giving rise to the term Quinary structure or if occurring in metabolic pathways – metabolons.
Fourth, while proteins fold on their own in vitro, a significant fraction have ‘helpers’ in vivo: molecular chaperones (Kim et al., 2013). It remains unclear to what extent and how chaperones alter the fundamental folding energy landscapes of proteins (Figure 1.2).

Figure 1.2 Major components of the folding and degradation branches of the proteostasis network in E. coli

A – GroEL (orange) and GroES (blue) (PDB 1aon) is homotetradecameric chaperonin of the Hsp 60 family of chaperones. It is posited to act as a conduit for folding by providing an environment for the unfolded polypeptide to fold without aberrant interactions with other molecules inside the cell. The complete activity of GroEL occurs in the presence of its co-chaperone GroES, which belongs to the Hsp 10 family of chaperones.

B- DnaK (purple – PDB 2kho) J domain of DnaJ (yellow – PDB 1xbl) and GrpE (red – PDB 1dkg). DnaK belongs to the Hsp70 family of molecular chaperones. Substrate binding and release by DnaK is controlled by the binding and hydrolysis of the nucleotide ATP to the chaperone which triggers allosteric conformational changes causing substrates to be released (ATP state) or bound (ADP state). DnaK is hypothesized to act on misfolded (of-pathway) or unfolded polypeptides. DnaJ is a co-chaperone of the Hsp40 family; DnaJ interacts with substrate to deliver it to DnaK. GrpE is the nucleotide exchange factor that accelerates the removal of ADP and incorporation of ATP into DnaK.

C- Lon protease (PDB – 4ypl) is thought to be the major proteolytic enzyme that clears misfolded species thus increasing the relative levels of folded proteins inside the cell. Lon is also proposed to degrade off-pathway misfolded intermediates (Gur and Sauer, 2008; Powers et al., 2012)
B.

C.
Fifth, proteins are vulnerable to competing intermolecular aggregation reactions to an extent that depends quite straightforwardly *in vitro* on the concentration of aggregation-prone species. Aggregation also competes with folding *in vivo*, but translating the parameters and mechanistic insights from aggregation studies *in vitro* to the *in vivo* context must be done with caution (Figure 1.3).

**Figure 1.3 Schematic for a coupled folding and aggregation landscape**

The above figure shows the combined folding-aggregation energy landscape inside the cell. Unfolded protein population serves as the progenitor of all states i.e. native/folded, misfolded and aggregates all arise out of the unfolded state. As the protein sequence codes for all states, the folding and aggregation landscapes become necessarily connected and increasing protein concentration is a major driving force behind aggregation. Cellular proteostasis is maintained by chaperones and degradation enzymes that favor the formation of the native state and prevent the accumulation of aggregated or any other non-productive states. Figure adapted from Hartl et al. 2009 (Hartl and Hayer-Hartl, 2009) and Jahn and Radford 2008 (Jahn and Radford, 2008).
Lastly, folding reactions in vivo are spatially organized such that some interactions will be preferred over others. In vitro it is very difficult to mimic a spatially organized, inhomogeneous environment. This point is absolutely central to the folding of membrane proteins, which, despite their importance, will not be a focus of this chapter.

1.3. Dissecting the factors that contribute to protein folding in vivo

1.3.1. Co-translational folding

How co-translational folding modulates the folding landscape of proteins (the landscape of the whole chain, which defines the pathways to reach the native state or off-pathway trajectories towards aggregation from the unfolded state) of proteins has been examined in a number of recent experimental and computational studies. O’Brien et al. introduced a computational approach to explore the impact of factors such as translation rate on folding (Ciryam et al., 2013; Nissley et al., 2016). Their findings suggest that mutations in mRNA that lead to altered translation rates may markedly alter folding outcomes. In a follow up in the same study, this group compared folding of ribosome nascent chain complexes that are arrested with those that are actively translating and concluded that at in-vivo translation rates, one-third of E. coli proteins would fold co-translationally. Krobath et al. also applied computational methods and found major impact of co-translational folding of arrested chains versus chain fragments folding freely (Krobath et al., 2013). They observed that the ribosome enhanced the population of low energy conformations dominated by local interactions. The interrelatedness of translation rate
and folding points to a level of selective pressure acting at the RNA level. Experiments with synonymous codons and ribosome display indeed point to the encoding of RNA-level information, which might be woven together with the sequence code for folding in vivo. The ribosome itself has been shown to affect folding (Ingolia et al., 2013; Shabalina et al., 2013). Using single molecule force experiments, Kaiser et al. (Kaiser et al., 2011) found that electrostatic interactions between the ribosome and their test protein (T4 lysozyme) retarded premature folding and allowed the nascent chain to remain in a folding-competent state. Knight et al. (Knight et al., 2013) examined the dynamics of a model nascent chain (a disordered protein) with varying charge and concluded that the ribosome surface electrostatically influenced the behavior of the chain, causing nascent protein variants carrying more negative charge to be more mobile. Recent studies by the O’Brien group demonstrate the power of a computational model that predicts changes in co-translational folding by the introduction of mutations in protein sequence that are measureable by experiments (Nissley et al., 2016). Additionally, efforts directed at understanding the structural aspects of co-translational folding have also revealed the role of the ribosome in preventing misfolding and aggregation of certain substrates (Cabrita et al., 2016; Deckert et al., 2016).

Viewing co-translational folding in terms of a naked nascent chain exploring conformational space is, however, greatly oversimplified. A whole host of chaperones (discussed in sections below) and quality control mechanisms lie in wait to greet the emerging polypeptide chain and assist its folding. The nature of this ribosomally associated greeting committee in E. coli has been reviewed by Bukau and co-workers.
(Gloge et al., 2014). Their studies and others have revealed the order of events upon ‘birth’ of a nascent polypeptide, beginning with N-terminal processing, and followed by chaperone interactions with trigger factor, the chaperone that has privileged access to nascent chains of cytoplasmic proteins. These authors have provided compelling arguments for an unfolding role of trigger factor, a ribosome associated chaperone, which interacts with nascent chains (Hoffmann et al., 2012). Single-molecule pulling experiments on maltose binding protein, Mashaghi et al. (Mashaghi et al., 2013) make a strong argument that trigger factor promotes productive folding by protecting partially folded states from misinteraction with neighboring molecules. The emerging role of trigger factor in nascent chain folding is supported by computational work from Dobson and colleagues, which posits that trigger factor interacts with emerging chains and retards folding in addition to shielding the polypeptide from unfavorable interactions (O'Brien et al., 2012). In eukaryotes, the Frydman lab has recently examined the co-translational roles of Hsp70 in yeast through a global analysis of ribosome nascent chains (O'Brien et al., 2012; Willmund et al., 2013). They found that Hsp70 interacted preferentially with large multidomain proteins of complex topology that were unlikely to be able to fold cotranslationally, consistent with the function of Hsp70 in maintaining the nascent chain in a folding-competent state.

1.3.2. Chaperones and proteostasis

Once a newly synthesized chain is away from the ribosome, it is further assisted by non-ribosomally associated chaperones to ensure its successful folding and minimize competing aggregation processes. While data have been rapidly accumulating on the
client repertoire of various chaperones *in vivo*, much less is known about how chaperone interactions affect protein folding reactions. For example, recent studies have asked how many and which proteins in *E. coli* are facilitated by the major chaperone systems GroEL/GroES and DnaK/DnaJ/GrpE: The Hsp70 system interacts with 700 cytoplasmic proteins, with particularly strong interaction with a subset of 180 that are aggregation-prone (Calloni et al., 2012). GroEL/ES was found in a proteomic study to support the folding of 250 proteins, with 84 of these obligatorily dependent on the chaperonin for folding; a recent revisiting of this question concluded that there were fewer true GroEL substrates (Kerner et al., 2005; Niwa et al., 2012), but the two studies agreed on the nature of the obligate substrates: small enough to fit in the chaperonin cavity, and enriched in metabolic enzymes and TIM barrels. Interestingly, Taguchi and co-workers (Niwa et al., 2012) found using a cell-free system that the major *E. coli* chaperone systems GroEL/GroES and DnaK/DnaJ/GrpE improved the solubility of 66% of their test group of 800 marginally soluble *E. coli* proteins. These studies have provided insight into the cellular dependence on chaperones for productive folding. How do chaperone-substrate interactions sculpt folding landscapes? Single particle cryo-electron microscopy has provided glimpses of substrates encapsulated in the GroEL chaperonin cavity, suggesting that they are quite collapsed (Chen et al., 2013; Clare et al., 2009). Using *in vivo* experiments monitoring growth as a criterion for fitness when mutant versions of the essential protein dihydrofolate reductase were expressed in the presence of differing amounts of GroEL/ES or the major protease Lon, Bershtein *et al.* (Bershtein et al., 2013) concluded that both the chaperonin and the protease act on the molten globule intermediate. These studies are consistent with current models in which GroEL
smoothens the folding landscape of poor folders, while DnaK largely acts to unfold its substrates, or maintain folding-competent or unfolded states (Kim et al., 2013; Saibil, 2013). There have been numerous efforts to determine the clients of Hsp90 chaperones, and several labs have applied biophysical methods to deduce the nature of the binding interaction and likely impact on substrate folding, but many questions remain for this chaperone as well (Taipale et al., 2012). Data suggest that Hsp90 substrates are folding intermediates with dynamic character. For example, p53 was observed to adopt a molten-globule state upon interaction with Hsp90, and the model substrate staphylococcal nuclease has been proposed to bind Hsp90 in an unfolded state via a local structural element (Nissley et al., 2016). Similarly tau also interacts with multiple hydrophobic contacts on Hsp90 suggestive of a late folding intermediate (Karagoz et al., 2014). The elegant recent single molecule study of trigger factor-substrate interactions described above demonstrated directly an unfolding activity – i.e. trigger factor unfolded the client protein to allow a refolding reaction to occur (Hoffmann et al., 2012). The interactions of small heat shock proteins with their clients have been a subject of constant examination, but here also we lack mechanistic understanding about how these chaperones affect folding. Similarly, the periplasmic chaperone HdeA binds molten globular substrates at low pH but the consequent effects on their folding are as yet unexplored (Basha et al., 2012; Foit et al., 2013; Hong et al., 2012). Strikingly, another periplasmic chaperone, spy has been recently shown to allow proteins to fold while still being associated with the chaperone (Stull et al., 2016). The eukaryotic chaperonin, TRiC, has to deal with larger proteins than encountered in *E. coli*, and a recent study concludes that it binds partially folded intermediates at domain boundaries, which helps explain how it may act on
multidomain substrates but does not reveal details of its impact on their folding (Russmann et al., 2012). Recent studies pioneered by the Kay lab (Sekhar et al., 2015, 2016) have found that DnaK keeps substrates in a more extended conformation yet allowing the substrate to retain some of the local secondary structure it has in the unfolded state. All told, current understanding of the impact of chaperone interactions on the folding landscapes of proteins remains incomplete, and the confluence of data and ideas from both in vitro and in vivo experiments will be needed to shed further light on this key question.

Chaperones work in teams and in partnership with degradation enzymes to facilitate folding in vivo and maintain protein homeostasis. A recent thrust is focused on admitting the complexity of integrated chaperone networks to elucidate the impact on folding of a substrate. A kinetic model centered on the major E. coli proteostasis components (chaperones, degradation enzymes, disaggregase), beginning with a translated nascent chain, has been developed jointly by Powers' and our labs (Powers et al., 2012). This model called FoldEco enables generation of hypotheses about the involvement of the proteostasis machinery and the folding success of a polypeptide when a few in vitro parameters are known. Also, by implementing in vivo FRET on fluorescently labeled chaperones Kumar and Sourjik (Kumar and Sourjik, 2012) were able to capture some of the interplay between the chaperone systems in E. coli, thus showing that the quality control systems are not isolated, but rather synergistic. The authors show how DnaK (or Hsp70) seems to be a central player in the de novo and re-folding branches of the proteostasis system.
1.3.3. Evolutionary pressures

The canonical definition of ‘the protein folding problem’—viz., how is the information for a protein folding landscape encoded in a given sequence—does not acknowledge the many selective pressures that have led to the existence of that sequence in the proteome of the organism from which it came. There is growing awareness of the importance of understanding the impact of evolutionary selection on protein sequences and consequently, their folding. It is not obvious when taken out of evolutionary context why factors such as function, turn-over interactions would change the protein folding landscape. Adding to the complexity is the fact that there are also many pressures acting on base sequences, for example to adjust translation rate and to enable regulatory processes to occur in transcription and translation (Li et al., 2012).

Protein stability naturally appears to be a property that would be selected for during evolution (Serohijos et al., 2012). Using a theory-based and simulations approach, Shakhnovich and colleagues make a strong case of how destabilizing mutations are selected against in highly abundant proteins, thus explaining their slow evolutionary rate. Yet, proteins designed in a laboratory generally display significantly higher stability than naturally occurring proteins (Koga et al., 2012). This observation suggests that stability is not the dominant driving force for sequence selection (Reynolds et al., 2013). A protein must possess a number of other properties to survive a selection for organismal fitness.
Perhaps the most obvious evolutionary pressure that impacts folding properties is the requirement for function. Many have noted that the selection for folding and function leads frequently to a trade-off (Dellus-Gur et al., 2013; Gosavi, 2013). Tawfik recently noted that some folds, like TIM barrels, may possess a property, which he terms polarization, that enables them to adapt to new functions (innovability) while maintaining foldability and stability (Dellus-Gur et al., 2013). Mechanistic impacts of the folding-function tradeoff were postulated for interleukin-1β (IL-1β). Capraro et al. (Capraro et al., 2012) observed that a functionally important structural feature, in IL-1β a β-bulge, acts to shape the functional landscape so that only one folding route is followed, whereas variants in which this bulge was mutated follow multiple routes.

Tawfik’s term ‘innovability’ may also apply to the ability of a protein evolutionary path to lead to new folds. In a very thought provoking study, He et al. (He et al., 2012) experimentally identified ‘mutational tipping points’ that enabled proteins to switch folds and evolve new functions. On the other hand a study of ancestral thioredoxin proteins by the Gavira group points out that although the ancestral protein differed considerably in sequence from the present version and was more thermo-stable, it folds into the same conformation as extant thioredoxins (Tokuriki and Tawfik, 2009). This highlights the robustness of a protein sequence to tolerate destabilizing mutations yet fold to carry out its function. It may well be an evolutionary advantage to retain this sequence nimbleness—the ability to absorb mutations that may cause a change in fold or function, which may improve organismal fitness, and in turn will have impact on the ‘winning’ sequences we see in current proteomes. In a recent study with repeat proteins, Smock et al. (Smock et
al., 2016) arrive at the conclusion that stability is inversely correlated with function, which is not highly surprising given that an optimal balance between stability and functionality decides the outcome of the protein sequence under evolutionary pressure (Dellus-Gur et al., 2013; Koga et al., 2012).

The fine tuning of protein sequences to achieve the fittest fold and optimized function do not operate in isolation in vivo. The idea that chaperones can buffer destabilizing mutations that directly improve their function, or serve as stepping-stones to proteins of improved function, has been experimentally supported (Tokuriki and Tawfik, 2009; Wyganowski et al., 2013). Mapa and colleagues (Mapa et al., 2012) performed experiments on a set of model substrates that populated kinetic intermediates and demonstrated that each selectively bound its cognate chaperone from the whole spectrum of E. coli chaperones present in lysate. They postulated that chaperone preferences co-evolve with foldability of protein sequences. This notion was recently emphasized in a provocative review on the origins of proteostasis (Powers and Balch, 2013). Furthermore, the authors of this review, along with another, have pointed out that protein evolution under the aegis of proteostasis is also environment dependent, and that integration of all factors operating on an organism leads to proteomic diversity (Bogumil and Dagan, 2012; Powers and Balch, 2013) (Figure 1.4).
Figure 1.4 Evolutionary pressures changes the protein folding landscape inside the cell

Protein folding *in vitro* can be thought of as navigating an energy landscape to reach from a high-energy unfolded state to a low energy folded state. Hills and valleys along the landscape represent kinetic barriers to protein folding which determine the speed at which a protein folds/unfolds and the energetic difference between the top and bottom of the landscape represents the folding free energy. *In vivo* the protein folding landscape is subject to modification by a myriad of factors which include folding, function, evolvability, and interactions with other biomolecules – all factors ultimately tying into evolutionary optimization of protein to benefit the organism.
The Protein Folding Problem

*In vitro*

Unfolded

Native

*In vivo*

Foldability
- Stability
- Co-T folding
- Landscape compatible with chaperones

Turnover
- Degradation
- Regulating translation

Protein-Protein Interactions
- Pathways
- Quinary structure
- Proteolysis
- Avoidance of deleterious interactions

Function
- Catalysis
- Signaling
- Transport

Adaptive Potential
- New functions
- New folds
- Chaperone-aided

Integrated Optimization of Organismal Fitness
Another factor that constrains sequence evolution is the requirement that proteins in vivo form productive interactions and avoid non-productive interactions (Pastore and Temussi, 2012). A corollary of this selective pressure is the avoidance of pathological aggregation, which may be viewed as a non-productive interaction. As noted recently by Levy et al. (Levy et al., 2012) the constraints on evolution of proteomes imposed by the need to form productive interactions and to avoid non-productive interactions is enhanced under the crowded conditions of the cell. A computational study by Yang and co-workers postulates (Yang et al., 2012) that avoiding deleterious interactions causes abundantly expressed proteins to evolve more slowly. In addition, evolutionary trends also suggest that there has been a decrease in the fraction of hydrophobic residues and a tendency for increased disorder within the proteome over time (Mannige et al., 2012). Such changes may arise as a function of natural selection; however, they have consequences on folding and protein-partner interactions. Furthermore, organization with favorable interaction partners has been recently hypothesized to add to protein stability (Dixit and Maslov, 2013). This concept is similar to that of chaperones being evolutionary buffers as discussed above, allowing proteins to accrue destabilizing mutations, yet fold and be better at their function (Tokuriki and Tawfik, 2009).

1.3.4 Macromolecular Crowding, high density of co-solutes and quinary interactions

- Macromolecular crowding:

In vivo, proteins must fold and be stable in a heterogeneous environment as concentrated as 400 g/L (Figure 1.1). Early hypotheses about macromolecular crowding placed heavy
emphasis on the contribution of the void volume effects and its impact on the enhancement of protein stability (Zhou et al., 2008). However, recent work by Pielak and colleagues reveals that the influence of the crowded in-vivo environment may be dominated by the prevalence of weak interactions, rather than the effects of excluded volume from macromolecular crowding, as previously believed (Miklos et al., 2011b; Monteith et al., 2015; Smith et al., 2016; Wang et al., 2012). These researchers found that protein crowding agents (bovine serum albumin, lysozyme) destabilized a test protein, CI2, in contrast to the stabilization expected from excluded volume effects (Miklos et al., 2011a; Miklos et al., 2011b). Such effects are expected to be protein- and context-dependent, and indeed Guo et al. used a novel rapid laser temperature stepping method capable of measuring complete thermal melts and kinetic traces in vivo to deduce that phosphoglycerate kinase was more stable in mammalian cells than in vitro (Guo et al., 2012). The seemingly contrasting results may differ because the experiments were performed at different temperatures, and the entropic component of crowding is temperature-dependent (Zhou, 2013). In addition, Dixit and Maslov (Dixit and Maslov, 2013) have argued compellingly that protein-protein interaction networks will stabilize proteins in vivo relative to in vitro. Recent studies performed by the Oliveberg (Danielsson et al., 2015) and Pielak (Smith et al., 2016) groups have provided a much more detailed understanding of how the cellular environment and more importantly the place of the protein in the a particular environment is an important factor determining changes to protein stability when compared to measurements made in vitro. Using elegant NMR experiments, the authors (Danielsson et al., 2015; Smith et al., 2016) determine the impact of macromolecular crowding agents, the E. coli cytosol and cellular
environment and the mammalian cellular environment on the stability of test proteins (the SH3 domain and superoxide dismutase) and arrive at the conclusion that protein stability \textit{in vivo} is a function of multiple parameters – namely, protein surfaces, Quinary interactions and in general the cellular milieu that interacts with the protein of interest.

• Quinary Interactions:
Protein function also involves the formation of higher order protein structures such as quaternary and quinary structures (McConkey, 1982), which involve the proteins to productively interact with each other. These higher levels of “folding” have long been implicated in metabolic functions, where the resulting organized pathways were termed 'metabolons' (Roguev et al., 2013; Velot et al., 1997) and in signaling pathways. It had long been hypothesized that the crowded cellular environment necessitates interactions among biomacromolecules, and although such weakly associated complexes are difficult to study \textit{in situ}, and would be difficult to isolate, recent efforts have led to new methods to interrogate them (Fraser et al., 2013; Lane et al., 2012). A recent study from the Teichmann and Robinson labs (Marsh et al., 2013) utilizes nano-electrospray ionization and gene fusion analysis to determine how several multimeric complexes are assembled and disassembled. Through their analyses the authors find that the formation of quaternary structure and protein assembly pathways also appear to be under evolutionary pressure. Gruebele and colleagues (Wirth and Gruebele, 2013) have weighed in on terming the panoply of weak interactions influencing a protein \textit{in vivo}, both specific and non-specific: 'quinary structure', as originally suggested by McConkey (McConkey, 1982) and re-introduced in an earlier review (Gershenson and Gierasch, 2011).
• Inorganic ions and Metabolites:

The total concentration of cytoplasmic inorganic ions in *E. coli* is ~300 mM according to the CyberCell database created by the Wishart laboratory (Sundararaj et al., 2004). The concentration of K\(^+\), by far the most abundant inorganic ion, varies drastically with osmotic conditions. Two hundred millimolar is reported (Jewett and Swartz, 2004; Record et al., 1998) to be physiologically relevant and CyberCell notes a concentration range of 200-250 mM. Similarly, concentrations of Mg\(^{2+}\) have been reported, with estimates ranging from 20 to 100 mM, although the amount of free Mg\(^{2+}\) is estimated to be much smaller at 1 - 2 mM (Moncany and Kellenberger, 1981; Outten and O'Halloran, 2001; Tyrrell et al., 2013). Estimates for other common inorganic ions include Na\(^+\) at ~5 mM, Ca\(^{2+}\) at ~0.1 mM, and CyberCell reports concentrations of Cl\(^-\) and total phosphate (H\(_2\)PO\(_4\)/HPO\(_4^{2-}\)/PO\(_4^{3-}\)) at 6 and 5 mM, respectively (Outten and O'Halloran, 2001; Shabala et al., 2009). Although variations between tissues exist, average concentrations of inorganic ions in *E. coli*, yeast and mammalian cells are in the same range.

Changes in intracellular ion levels can have pronounced effects on the conformational properties of proteins and can have detrimental effects as seen in the enhanced aggregation of a-synuclein at high salt concentrations (Hoyer et al., 2004). Growing evidence suggests that changes in metal homeostasis and altered intrinsically disordered proteins-metal interactions contribute to the pathogenesis of several neurodegenerative disorders (Botelho et al., 2012; Breydo and Uversky, 2011; Kepp, 2012) suggesting the critical balance that requires maintenance inside the cell. Indeed, many amyloidogenic
IDPs such as α-synuclein, tau or amyloid β (Aβ) peptides directly bind metals, and metal interactions modulate their in vitro aggregation behaviors (Theillet et al., 2014).

Recent advances in metabolomics technologies have allowed the concentrations of large numbers of metabolites to be measured in E. coli (Bennett et al., 2009). In glucose-fed, exponentially growing E. coli cells the combined concentrations of metabolites have been estimated to be ~300 mM, with glutamate (Glu) being the most abundant metabolite by far (96 mM), followed by glutathione, fructose-1,6-bisphosphate and adenosine triphosphate (ATP) at 17 mM, 15 mM and 9.6 mM, respectively (Bennett et al., 2009). However, these concentrations depend on the culture medium. By changing the carbon source from glucose to glycerol or acetate, intracellular Glu levels change from 96 mM to 149 mM, to 45 mM, respectively (Bennett et al., 2009). Similarly, intracellular glutathione concentrations change from 17 mM, to 18 mM, to 8 mM; Fructose-1,6-bisphosphate from 15 mM, to 6 mM, to <0.15 mM; ATP from 9.6 mM, to 9.0 mM, to 4.1 mM (Bennett et al., 2009). Significant variations in intracellular Glu levels due to changes in glucose levels in the growth media or due to changes in osmotic conditions have also been seen in other studies (Cayley et al., 1991; Roe et al., 1998). When E. coli cells were grown in McIlvaine's medium (minimal medium with citrate/phosphate buffer, supplemented with thiamine and glucose) at pH 6, and harvested at mid-exponential phase, the total concentration of all amino acids was determined to be ~90 mM, of which Glu comprises ~60 mM (Roe et al., 1998). In the presence of 200 mM glucose, Glu concentration increase to ~117 mM. At 400 mM glucose, it is ~160 mM. By contrast, CyberCell lists the combined E. coli concentrations of all 'small organic molecules' as 40-
50 mM (undefined growth-medium and -stage), concentrations of free amino acids total ~15 mM, and ATP is indicated between 1.3-7.0 mM, depending on growth conditions and sugar sources (Theillet et al., 2014). These numbers indicate the sheer abundance of co-solutes and metabolites, all of which can influence the stability and interactions of proteins inside cells.

Some metabolites such as glycerol, trehalose and zwitterions such as trimethylamine-N-oxide, proline, betaine and ectoine, stabilize proteins at intracellular concentrations between 100-300 mM (Bandyopadhyay et al., 2012). These compounds may represent a special class of metabolites, termed ‘osmolytes’ because they also function as powerful stabilizing agents in vitro. Nevertheless, these data indicate that metabolite concentrations in the range of ~300 mM are sufficient to modify the properties of individual proteins.

While it is unlikely that metabolites generally induce folding of proteins in cellular environments, they may modify the structural features of some of them. Furthermore, the binding of these metabolites might shift the populations of proteins that exist in the folded/unfolded state to selectively populate one or the other. Given the fact that proteins with binding function are amongst the most populate class in the proteome, there remains little light shed on how small molecule ligands can impact the folding fate of a large fraction of the proteome.
1.4. Kinetic partitioning – folding/aggregation/chaperone binding/ligand binding:

To understand how folding inside the cell differs from folding in vitro one has to consider all the different parameters that can alter the folding of a protein and the fate of the polypeptide chain in vivo. To fully capture the processes going on inside the cell, one has to understand that the polypeptide sequence is under competition from the various factors mentioned above. While the correct and functional folded state may be optimized by evolution, protein levels, avoidance of aggregation and dependence on chaperones and binding ligands can all be described by invoking a model of kinetic competition or kinetic partitioning (Figure 1.5).

**Figure 1.5 General schematic for the fates and the kinetic partitioning of a polypeptide chain inside the cell**

A synthesized protein (S) can fold co-translationally (not shown) or can fold to the native (N) state from the unfolded (U) state. The U state is the progenitor for all states inside the cell. The U state can misfold into an off pathway (M) state which has the propensity to form protein aggregates (A). The N state has the capability to bind ligands (blue star) and the M and U states can be degraded by cellular proteases. The transition from the U state to any other state, or interconversion between states is principally governed by the intrinsic rate constant for that conversion and the concentration of the species which acts as a reactant in that conversion.
A model of kinetic partitioning takes into consideration biophysical parameters and the concentrations of the interacting components and uses a set of coupled ordinary differential equations to establish relationships between folding outcomes and the impact of the cellular environment on them. For example, one of the best-described model for kinetic competition is FoldEco (Cho et al., 2015; Powers et al., 2012) (Figure 1.6).

**Figure 1.6 The FoldEco model of proteostasis and kinetic partitioning in E. coli**

The figure is a schematic of FoldEco which is a system of differential equations that computes the flux of synthesized protein through different states (folded, unfolded, misfolded, aggregated and degraded) in E. coli taking into consideration their intrinsic folding and misfolding rate constants and how these proteins interact with the proteostasis machinery of the cell, i.e. the association rates between different states and a chaperone or degradation enzyme. It is a computational system which overlays the chaperones and degradation machinery of *E. coli* on top of the kinetic competition that naturally occurs for protein sequence. It serves as an important tool for generating hypotheses as to how altering folding/misfolding properties can have an impact on the folding fate of a protein inside the cell.
FoldEco is a computational algorithm written to determine how the biophysical properties of a protein interface with the proteostasis components to produce a given yield of folded protein. The biophysical properties of the protein include the folding and unfolding rate constants, misfolding and aggregation rate constants and equilibrium constants for all those processes. The four states, folded, misfolded, unfolded and aggregated, are a minimal set of states states that a monomeric protein can populate after biosynthesis; one could add complexity by adding intermediates or treating multimeric proteins, which could populate different oligomeric states. The program is specifically written to incorporate the major components of the proteostasis machinery of *Escherichia coli*, namely the two major chaperone systems, the GroELS and DnaKJE systems, a disaggregation system which includes ClpB and DnaK and a mechanism for degradation assumed to be catalyzed by the Lon protease. The proteostasis components have the capacity to interact with three of the four possible states a protein can populate. The interaction between the unfolded and misfolded states and the chaperone systems and Lon are defined by association rate constants and the association of the aggregated state with the disaggregation machinery is similarly defined. The program integrates all these processes as coupled differential equations which when solved yield a theoretical output of what to anticipate under a defined set of folding and proteostasis conditions. I will discuss chaperone binding and the dependence of some proteins on chaperones to fold in Chapter 3. By understanding how the protein partitions between folded and aggregated states and with the knowledge of the chaperone concentrations, one should be able to understand the impact of the proteostasis components on the folding of a test protein.
Important, yet not fully captured by FoldEco is the impact of the metabolites and ligands on the folded state of the protein. As mentioned earlier, ligand binding has the capacity to shift the equilibrium that exists between different protein states \textit{in vivo}. FoldEco also has been focused initially on the fate of a single protein, but the program has the capacity to consider competition among different folding clients for the protein homeostasis network.
1.5 Statement of thesis

Despite the advances made in the field of protein folding, understanding how protein folding properties interface with the complex cellular milieu remains insufficiently studied. Most of our quantitative knowledge on folding comes from studies performed on small well-folded proteins in dilute and purified environments. This approach, while revealing the minutiae of the folding pathway, does not fully capture what could be occurring inside the cell. Specifically, such studies do not capture the essence of kinetic partitioning: protein sequences are have several possible pathways they can follow, folding, aggregation, degradation, and the fate of the polypeptide chain depends on both the intrinsic biophysical properties that govern the states a synthesized protein will visit and how the cellular quality control machinery interacts with a synthesized polypeptide chain so as to sculpt its landscape to favor folding and reduce detrimental outcomes. In this thesis, we use a combined approach of experimental testing and mathematical modeling to examine two aspects of the cellular environment that could affect the successful folding of a synthesized protein. In Chapter 2, we dissect quantitatively the effect of ligands on the folding fate of the protein. Our investigation seeks to capture what could be happening to populations of multiple states that can be occupied by synthesized protein in the presence of a high affinity ligand. We hypothesize that ligand binding occurs after the protein successfully folds and that binding prevents non-productive outcomes such as aggregation by effectively lowering the rate at which the protein unfolds and populates an aggregation-prone state. We also hypothesize that the remediation provided by the binding of the ligand is dependent on the relative
biophysical properties of the protein and how it is the competition between populating the folded state versus non-productive states that determines how effective the ligand, would be. We further test our hypothesis using a system pertinent to human disease and where protein degradation can be alleviated using a high affinity small molecule, an intervention now dubbed pharmacological chaperone therapy. In Chapter 3, we investigate why certain proteins obligatorily require the aid of chaperones to successfully reach the folded state. To understand this, we utilized a test protein that obligatorily relies on one of the chaperone systems to correctly fold and investigate how this protein’s folding outcome changes when we alter the availability of different chaperone systems and degradation machinery to the protein upon synthesis in the cell. Specifically, we work in *E. coli*, and manipulate its two major chaperone systems, the GroELS and DnaKJE systems, and a major degradation system (the Lon protease). We work in *E. coli* because it is a simple organism with approximately 4000 genes and is capable of being grown and manipulated to express varying amounts of protein in a facile manner. Using the above-mentioned FoldEco model of proteostasis kinetic partitioning to explain the observed experimental folding fate of the protein we deduced what the folding, misfolding and aggregation parameters for such a chaperone-dependent protein were; parameters, which would not be easy to measure in *in vitro* environments. Combining modeling results with experiments monitoring the folding fate of our test protein, we were able to understand how ligands and the intrinsic biophysical properties of a protein that obligatorily depends on cellular proteostasis machinery affect the folding outcome of a protein to reach the folded state, thus shedding light on the processes that govern protein folding inside the cell. In chapter four we use our knowledge on ligand binding and enhancement of stability to build a
protein folding sensor to measure protein thermodynamic stabilities inside the cell. We discuss one approach for building such a sensor, the mutually exclusive folding model and show how it can be implemented using a suitable protein – ligand pair as a starting point for our sensor. We then go on to show preliminary experiments that utilize our folding sensor in vitro and in vivo to demonstrate that such a construct can theoretically be used to test protein stability and the limitations associated with the use of a mutually exclusive folding construct as a sensor for protein stability measurements.
CHAPTER 2
INVESTIGATING THE IMPACT OF SMALL MOLECULES ON PROTEIN FOLDING INSIDE THE CELL

This chapter investigates how the presence of a high affinity ligand biases the competition between folding and aggregation to favor the folded state, thereby reducing the effective unfolding rate and preventing the channeling of protein into non-productive pathways such as aggregation or degradation. The work for this chapter was performed in collaboration with Matthew Metcalf, Scott Garman and Evan Powers, and has been submitted for publication.

2.1. Introduction

Cellular proteins are bathed in a sea of diverse ligands ranging from other macromolecules to metabolites. The intracellular concentration of macromolecules is high, up to 400 mg/ml in bacteria and eukaryotic organelles, and small molecules such as metabolites are also present at high concentrations inside cells. For example, in glucose-fed *E. coli*, glutamate, ATP, glutathione, NAD+ and other metabolites are present at millimolar levels (Bennett et al., 2009; Gershenson and Gierasch, 2011). Thus, in a normal physiological milieu proteins participate in binding reactions and equilibria with many other intracellular solutes; these binding interactions range from strong, specific interactions to weak, transient interactions. Surprisingly, little attention has been paid to the impact of these many binding interactions on the folding energy landscapes of proteins in the cell, despite active research directed towards enhanced understanding of protein folding in cells (Bershtein et al., 2015; Bershtein et al., 2013; Cho et al., 2015;
Dixit and Maslov, 2013; Hingorani and Gierasch, 2014; Kim et al., 2013; Monteith et al., 2015). It is in the intracellular environment that is replete with co-solutes that proteins must successfully navigate an intersecting folding – aggregation landscape to reach their functional folded states (Jahn and Radford, 2008). In addition, small molecules added to cells exogenously as potential therapeutics will influence the folding – aggregation landscape. In the present study, we have explored the impact of ligand binding on folding in the cell using two simple test systems, combining experimental and mathematical modeling approaches to gain insight into the magnitude of the effect of ligand binding on the partitioning of the test protein between proper folding and aggregation.

2.2. An aggregation-prone dihydrofolate reductase mutant can be rescued using a high affinity ligand

To determine the effect of a high affinity ligand on the kinetic balance between folding and aggregation, we chose to work with a destabilized version of E. coli dihydrofolate reductase (dDHFR) as our test protein. dDHFR was destabilized by the introduction of two glycine residues in a surface loop between residues 106 and 107 (Figure 2.1A). The mutation did not perturb any residues in the ligand-binding site. We tested the affinity for the ligand trimethoprim using an intrinsic (present in the DHFR molecule) tryptophan fluorescence quenching assay (see methods) and found that the our mutation had only slightly perturbed affinity for this ligand (approximately 40 nM compared to 8 nM as reported in literature by (Watson et al., 2007)) (Figure 2.1C). Upon over expression in E. coli BL21(DE3) cells, dDHFR largely (80%) aggregated in the absence of any added ligand. Strikingly, the presence of a non-hydrolysable analog of the natural DHFR
substrate trimethoprim (TMP) during protein synthesis re-partitioned the protein to the soluble fraction to a large extent (Figure 2.1D).

*In vitro* characterization of dDHFR revealed a $\Delta \Delta G$ of approximately 3 kcal/mol with respect to wild type (Figure 2.1B) (Ionescu et al., 2000) at 37 °C. The equilibrium free energy of unfolding was determined as a function of increasing concentration of trimethylamine N-oxide (TMAO) by performing urea denaturation experiments at desired concentrations of TMAO. Extrapolating the free energy to 0 TMAO allowed us to calculate the $\Delta G$ to be about 0.5 kcal/mol at 37 °C (the equilibrium stability measured by urea denaturation experiments of wildtype DHFR measured at 37 °C was reported at 3.5 kcal/mol). This translates to roughly 70% of the protein population being folded at equilibrium, suggesting that the protein tended to significantly populate the unfolded state even at equilibrium. Together with our *in vitro* and protein expression data, the dDFHR-TMP seemed to be an ideal model system to test the impact of a high affinity ligand on the kinetic partitioning between folding and aggregation.
**Figure 2.1** dDHFR is a destabilized test protein the solubility of which is improved upon addition of a high affinity ligand trimethoprim.

A. Test protein dDHFR (PDB 7dfr) with a gly-gly insertion between residues 106 and 107 (highlighted in red) and the cognate ligand trimethoprim, which mimics substrate and is a competitive inhibitor.

B. Fluorescence quenching of dDHFR (50 nM) upon titration with TMP enables the calculation of the fraction of dDHFR that is bound at each TMP concentration. The error bars for the individual data points represents the standard error of the mean from triplicate measurements. Fitting single site binding model to the data yields an estimated equilibrium dissociation constant ($K_d$) of 37 ± 6 nM. The error in $K_d$ is the standard error of the best fit parameter based on the fit to the averaged data.

C. Stability of dDHFR was measured at varying TMAO concentrations as a function of urea concentration. Extrapolation from stabilities measured at different TMAO concentrations yields $\Delta G$ (folding free energy) at 37 °C to an average of 0.5 kcal/mol which is a $\Delta \Delta G$ of ~ 3 kcal/mol compared to wildtype (Ionescu et al., 2000).

D. dDHFR synthesized in the absence of ligands partitions predominantly to the aggregated fraction, however presence of the high affinity ligand trimethoprim (0.8mM in the above sample) redistributes dDHFR to now be mostly soluble.
A. PDB: 7DFR

B. Fraction dDHFR Bound

\[ \text{[TMP] \text{\(\mu\)M}} \]

Trimethoprim
C.

D.

<table>
<thead>
<tr>
<th></th>
<th>- TMP</th>
<th>+ TMP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

Thermodynamic stability (kcal/mol) vs. [TMAO] (M)
2.3. Testing the reversibility of aggregation and ligand binding inside the cell

We next tested if the effect of TMP had to do with prevention of dDHFR from forming aggregates, or whether the ligand caused a rescue from the aggregated state. Secondly, we also wanted to determine if the dDFHR-TMP complex formed inside cells was essentially irreversible formed on the time scale of our experiments, or transient and readily dissociated. We performed chloramphenicol shut-off experiments to answer both questions (Aakre et al., 2013). We shut off translation using chloramphenicol and replaced the media containing TMP with fresh media containing no TMP. We hypothesized that if the dDHFR-TMP complex dissociated, then we would observe a rise in the aggregated fraction of the protein. Similarly, we shut off translation in the presence of high concentrations of TMP (80 µM). As TMP is added to the outside of cells, it acts as a large reservoir of ligand. Our hypothesis in this case was that if the aggregates of dDHFR were disaggregated and re-dissolved upon addition of TMP, we would observe an increased level in the soluble protein even after the protein had aggregated inside the cell. Both treatments were allowed to proceed for an hour before interrogating the soluble and insoluble fractions (Figure 2.2). From both experiments we could conclude that on the time scale of our experiments, the dDHFR-TMP complex was stable in cells (no increase in aggregates upon depletion of TMP) and that the aggregate that formed could not be re-dissolved by the addition of high TMP concentrations, indicating a lack of rescue from the aggregated state (Figure 2.2). Our model could now treat these two processes (aggregation and ligand binding) as irreversible.
Figure 2.2 Chloramphenicol shut-off experiments indicate the effective irreversibility of ligand binding and aggregation

A. The partitioning of dDHFR in the presence of TMP before and 1 hour after stopping translation with chloramphenicol (CAM). T = total dDHFR in all forms; P = aggregated dDHFR in the pellet after centrifugation at 18,000×g for 30 min; S = soluble dDHFR in the supernatant after centrifugation at 18,000×g for 30 min. Gels were stained with Coomassie G-250. The numbers above each lane are the mean ± standard error in the fraction of protein present in the corresponding states for three trials. In the presence of TMP, the protein found in the aggregate does not redistribute to the soluble fraction over the 1 hour.

B. As A, but showing the partitioning of dDHFR in the presence of TMP, and 1 hour after stopping translation and changing the media to remove TMP. Removal of the TMP does not re-partition the soluble dDHFR into the aggregate fraction over 1 h. The associated bar graphs show the average aggregated and soluble fractions of three measurements and the associated standard error of the mean.
A.

<table>
<thead>
<tr>
<th>T</th>
<th>P</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>-CAM, +TMP</td>
<td>+1 h,</td>
<td>+CAM, +TMP</td>
</tr>
</tbody>
</table>

![Graph showing F_agg values for different conditions]
B.

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>P</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>-CAM, +TMP</td>
<td>+1 h,</td>
<td>+CAM, -TMP</td>
<td></td>
</tr>
</tbody>
</table>

![Bar chart showing F_{agg} for different conditions](chart.png)
2.4. A mathematical model allows us to hypothesize the effect of high affinity ligands on the kinetic partitioning of their cognate proteins (developed by Evan Powers)

In the model in shown in Figure 2.3, the protein is synthesized at a rate $\sigma$ (µM s$^{-1}$) and released from the ribosome in an unfolded state (U), which can fold reversibly to the native state (N) with folding and unfolding rate constants $k_f$ and $k_u$ (both s$^{-1}$), respectively. The unfolded state can also aggregate through visiting some off-pathway misfolded intermediate which rapidly self associates to form aggregated protein (Agg).

Each of these processes is treated as being irreversible (based on our chloramphenicol shut-off assays and the fact that degradation is an irreversible process) and having a first-order rate limiting step with the rate constant $k_{agg}$ (s$^{-1}$). The native state can bind to a ligand to form a complex (N:L). Again based on our controls (described in the previous section), we treat this process as irreversible, with a second-order association rate constant $k_a$ (µM$^{-1}$ s$^{-1}$). The model described here is not just limited to a competition between folding and aggregation but in general competition between folding and other pathways a polypeptide chain can visit inside the cell. Shown in Figure 2.3 is also the competition between folding and degradation (for example, Endoplasmic Reticulum Associated Degradation, ERAD) where again, the unfolded population of the protein is the progenitor to states that are degraded (Deg) with a rate constant $k_{deg}$ (s$^{-1}$). As degradation is intrinsically an irreversible process, we can account for it in the same fashion we used for the cytosolic folding – aggregation competition. Finally, the native state (free or ligand bound) can in some cases (depending on the organelle of interest) be
secreted with a first-order rate constant $k_{sec}$ (s$^{-1}$). We assume that the turn-over rate of secreted protein (Sec) is much longer than the time scale of the experiment so that degradation after secretion is negligible. Also, we assume that the intracellular concentration of any ligand that is present is constant because there is a large reservoir of ligand in the media and the ligands rapidly equilibrate across the cell membrane. Thus, the “terminal states” in this model, that is, the states in which protein will continue to accumulate indefinitely, are the degraded, aggregated, and secreted states. If there is no secretion, then the ligand-bound native state (N:L) will be a terminal state rather than the secreted state.

We thank Evan Powers for generously providing the mathematical rigor to our model and the detailed steps are noted in the appendix.

The fraction of protein that remains soluble (i.e., is not degraded and does not aggregate) at a given time is straightforward to measure experimentally. We refer to this quantity as $F_r$. We replace $F_r$ with either $F_{soluble}$ or $F_{preserved}$ depending on the competition (folding vs aggregation or folding vs degradation) under investigation. When a ligand is present, the system rapidly reaches a “pseudo-steady state” in which the concentrations of $U$ and $N$ are constant and $Deg, A, N:L$, and/or $Sec$ accumulate at a constant rate. $F_r$ can then be written as follows

$$F_r = 1 - \frac{[A]_t + [Deg]_t}{[P_{tot}]_t} = 1 - \frac{1}{1 + B_1 \left( \frac{1}{1 + \frac{1}{B_2 + B_3[L]}} \right)}$$

(1)

Where $[A]_t$ and $[Deg]_t$ are the concentrations of aggregated and degraded protein at time $t$; $[P_{tot}]_t$ is the total concentration of protein that has been synthesized at time $t$;
[L] is the ligand concentration; $B_1 = k_f/(k_{agg} + k_{deg})$; $B_2 = k_{sec}/k_{u}$; and $B_3 = k_a/k_{u}$. In the absence of a ligand, pseudo-steady state is reached more slowly and $F_r$ is time dependent. An approximate expression for $F_r$ under these circumstances is:

$$F_r|_{[L]=0} = 1 - \frac{1}{1 + B_1 \left( \frac{B_2}{1 + B_2} \right)} + \frac{1}{k_u t} \left( \frac{1 + B_1}{1 + B_1 \left( \frac{B_2}{1 + B_2} \right)} - 1 \right)$$  \hspace{1cm} (2)$$

This expression does not hold at early time points, but the approximation improves with time as the system approaches steady state (see Appendix). Equation (2) becomes identical to equation (1) with $[L] = 0$ at very long times ($k_u t >> 1$).

Equation (1) provides a framework for understanding the effect of ligand binding on protein partitioning to folding, degradation and aggregation in vivo. Importantly, it predicts that $F_r$ increases with the ligand concentration until it reaches the following limit:

$$F_r|_{[L] \rightarrow \infty} = F_{r, max} = 1 - \frac{1}{1 + B_1} = 1 - \frac{k_f}{1 + k_{agg} + k_{deg}}$$ \hspace{1cm} (3)$$

Thus, at very high ligand concentration, a protein’s partitioning becomes purely a matter of kinetics and is no longer dependent upon folding thermodynamics.

Specifically, $F_r$ is controlled by the relative values of $k_f$ vs. $k_{agg} + k_{deg}$. 
Figure 2.3 General schematic of the kinetic partitioning influenced by ligand binding

The schematic describes the kinetic partitioning of a protein upon synthesis in the presence of a high affinity ligand. The species in the model are: U, unfolded protein; N, natively folded protein; N:L, ligand bound natively folded protein; Deg, degraded protein; A, aggregated protein; Sec, secreted protein. The total protein synthesized, \( P_{\text{tot}} \), includes all of these states. The rate constants are: \( \sigma \), protein synthesis rate (\( \mu \text{M s}^{-1} \)); \( k_{f} \), folding rate constant (s\(^{-1}\)); \( k_u \), unfolding rate constant (s\(^{-1}\)); \( k_a \), protein–ligand association rate constant (\( \mu \text{M}^{-1} \text{s}^{-1} \)); \( k_{\text{deg}} \), degradation rate constant (s\(^{-1}\)); \( k_{\text{agg}} \), aggregation rate constant (s\(^{-1}\)); \( k_{\text{sec}} \), secretion rate constant (s\(^{-1}\)).
2.5. Testing the hypothesis of an upper limit on aggregation prevention in the DHFR–TMP system.

We tested our model by titrating increasing concentrations of TMP into cells expressing dDHFR. As described by equation 1, we saw a titratable rise in the $F_r$ as a function of our ligand concentration up to a certain limit beyond which any addition of ligand does not cause any increase in the amount of protein that aggregates or remains soluble (Figure 4). Given our observations, it is reasonable to conclude that the parameters from the fit of equation (1) accurately reflect the biophysical properties of dDHFR inside the cell. Of particular interest is the value of $B_1$, which is $2.9 \pm 0.8 = k_f/(k_{agg} + k_{deg}) = k_f/k_{agg}$ (the BL21 strain of *E. coli* lack Lon protease which is mainly responsible for DHFR degradation (Bershtein et al., 2013; Cho et al., 2015), so $k_{deg} = 0$), because it reveals that folding is about three times faster than aggregation. According to equation (3), this ratio controls the maximum extent to which TMP can rescue the folding of dDHFR. No matter how high the concentration of TMP is, a portion of the dDHFR will be diverted to aggregation before TMP binding can rescue it; for dDHFR, that portion is about 25% (Figure 2.4B).

The dDHFR expression experiments were 90 min long ($t = 5400$ s). Given that $B_1 = 2.9 \pm 0.8$ and that $F_{r|L=0} = 0.15 \pm 0.02$, equation (3) can be solved for $k_u$ to yield $k_u = 0.0035 \pm 0.0010$ s$^{-1}$. This value of $k_u$ is consistent with unfolding rate constants measured for other destabilized DHFR variants. However, it should be noted that dDHFR can bind to a
second ligand, the coenzyme NADPH. The effects of this ligand on dDHFR’s behavior are subsumed into the folding energetics, and in particular into the effective unfolding rate constant.

**Figure 2.4 Titration of TMP shows that dDHFR \( F_r \) scales with concentration of TMP and plateaus as a function of its intrinsic folding and aggregation rate constants**

A. SDS PAGE gels showing that the titration of TMP shifts the dDHFR population from mostly aggregated to mostly soluble in a dose dependent manner.

B. A plot of \( F_r \) vs. TMP concentration after expression for 90 min in *E. coli* BL21 (DE3) cells at 37 °C (open circles). The solid curve is the fit of equation (40) to the data with \( B_1(\frac{k_f}{k_{agg} + k_{deg}}), = 2.9 \pm 0.8, B_3 = 0.048 \pm 0.025 \mu M^{-1}, \) and \( R^2 = 0.84. \) The value of \( F_r \) at \([\text{TMP}] = 0\) is shown as a filled red circle. The dashed line represents the maximum value of \( F_r \) according to equation
A. 

\[
[TMP] = 0 \mu\text{M} \quad 2 \mu\text{M} \quad 5 \mu\text{M} \quad 10 \mu\text{M} \quad 20 \mu\text{M} \quad 40 \mu\text{M} \quad 60 \mu\text{M} \quad 80 \mu\text{M} \quad 100 \mu\text{M}
\]

\begin{array}{ccccccccccc}
\text{T} & \text{P} & \text{S} & \text{T} & \text{P} & \text{S} & \text{T} & \text{P} & \text{S} & \text{T} & \text{P} & \text{S} \\
\text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} & \text{dDHFR} \\
\end{array}

B. 

\[d_{\text{DHFR}} F_r \quad [\text{TMP}] \quad \mu\text{M} \]

\[F_{r,\text{max}} \quad \phi \quad \phi \quad \phi \quad \phi \]

\[0 \quad 1 \quad 2 \quad 3 \quad 5 \quad 10 \quad 20 \quad 30 \quad 50 \quad 100\]
2.6. Testing the model on a pharmacological chaperone – The alpha galactosidase and DGJ model

The chemical chaperone 1-deoxygalactonojirimycin (DGJ) is currently under investigation as a pharmacological chaperone therapy for Fabry disease (Benjamin et al., 2009; Khanna et al., 2010; Wu et al., 2011). DGJ has been shown to increase the cellular concentrations and activity of several disease related α-galactosidase (α-GAL) point mutants in cell culture experiments (Benjamin et al., 2009; Wu et al., 2011). We reasoned that the intracellular rise in protein concentration was a manifestation of the impact of ligand binding on the kinetic competition between folding and degradation for the mutant α-GAL proteins. We tested the effect of increasing concentrations of DGJ on the intracellular protein levels of α-GAL R301Q (a mutation associated with late onset Fabry disease) in HEK cells (Fig 2.3A,B). Predictably, the levels of α-GAL increase with increasing concentrations of the pharmacological chaperone until a saturation point, analogous to what was observed in the DHFR experiment (Figure 2.5). To determine the level of α-GAL in the absence of any degradation we used the proteasome inhibitor lactacystin in one of our samples. Predictably, when compared to the sample with the added proteasome inhibitor, the levels of α-GAL increase with increasing 1-DGJ concentrations until a saturation point is reached (Figure 2.5A), analogous to what was observed in the dDHFR experiment and consistent with the predictions of equation (1).
α-GAL R301Q does not aggregate intracellularly ($k_{agg} = 0$), $B_1 = 1.24 \pm 0.08 = k_f/(k_{agg} + k_{deg}) = k_f/k_{deg}$, indicating that degradation and folding are almost evenly poised.

According to equation (1) and using this value of $B_1$, the maximal rescue possible for α-GAL R301Q is approximately 55% of the protein. Unlike dDHFR, the best fit of equation (1) to the α-GAL R301Q expression data is still very good at low ligand concentrations, indicating that the system reaches pseudo-steady state over the 24 h time course of these experiments (Figure 2.5B).

**Figure 2.5** Titration of DGJ shows that α-Gal $F_r$ scales with concentration of DGJ and plateaus as a function of its intrinsic folding and degradation rate constants

A. Western blot showing the increase in alpha galactosidase levels upon titration with DGJ, lac is proteasome inhibitor lactacystin as our 0 degradation sample to which the increase in protein concentration brought about by the ligand is compared to. The different bands observed represent the differentially glycosylated species.

B. A plot of $F_r$ vs. 1-DGJ concentration after expression for 24 h in HEK 293T cells at 37 °C (open circles). The solid curve is the fit of equation (1) to the data with $B_1 = 1.2 \pm 0.2$, $B_2 (k_{sec}/k_u) = 0.33 \pm 0.11$, $B_3 = 0.14 \pm 0.10 \mu M^{-1}$, and $R^2 = 0.88$. The value of $F_r$ at $[1-DGJ] = 0$ is shown as a filled red circle. The dashed line represents the maximum value of $F_r$ according to equation (3).
2.7. Conclusions about ligand binding

This study has broad implications for how pharmacological chaperones function and may provide explanations in cases where they do not ameliorate protein misfolding diseases where the ratio of the folding rate constant to the sum of the aggregation and degradation rate constants is less than one. Migalastat (DGJ) is a well known pharmacological chaperone that improve the folding of the α-galactosidase mutants implicated in Fabry disease, respectively (Benjamin et al., 2009; Ishii et al., 2007). However, there are many variants of these proteins that are unresponsive to pharmacological chaperone therapy (Wu et al., 2011). We posit that a simple mechanism that invokes increased thermodynamic stability upon ligand binding is not sufficient to explain how such molecules work. To completely understand the mode of action of pharmacological chaperones it is imperative to understand the kinetic partitioning between productive (folded) and non-productive (aggregated/degraded) outcomes. We argue that pharmacological chaperones would be most effective on proteins where the rate of folding is not the limiting step, i.e. the protein needs to fold faster than it aggregates or is degraded so that the small molecule can bind it and prevent the partitioning into non-productive states. Thus, to put forward a molecule as an effective pharmacological chaperone requires knowledge of how the molecule alters the kinetic competition of the protein it targets rather than just the affinity with which it binds.

A case in point is the limited success of pharmacologic chaperones as monotherapies for cystic fibrosis, which is caused by mutations in the cystic fibrosis transmembrane
conductance regulator (CFTR) (Clancy et al., 2012). CFTR is a large membrane protein that folds very slowly, and as a result a substantial fraction of even wild type CFTR is degraded before it can be exported (Lukacs et al., 1994). This situation is worse for CFTR mutants, many of which partition almost exclusively to ERAD leading to loss of CFTR function (Lukacs et al., 1994; Ward et al., 1995). Treating cells expressing a common disease-associated mutant (ΔF508 CFTR) with lumacaftor, a small molecule that binds CFTR, modestly increases the trafficking of CFTR to the cell surface (to 14% of wild type levels) (Van Goor et al., 2011). In terms of our model, this “low ceiling” for the efficacy of a pharmacologic chaperone for ΔF508 CFTR is an unavoidable consequence of its slow folding, and it suggests that alternative strategies for improving ΔF508 CFTR trafficking—for example, using proteostasis regulators either to increase $k_f$ (by improving chaperoning) or decrease $k_{\text{deg}}$ (by inhibiting ERAD)—could yield improved results (Powers et al., 2009).

The influence of ligand binding on the energy landscapes for folding and aggregation/degradation is a widespread phenomenon in nature, and ligands often shepherd a protein towards well-folded states. For example both antibody heavy chain and light chain diseases are caused by insufficient amounts of their binding partners. Similarly, β-2 microglobulin aggregation is triggered by its dissociation from the heavy chains of MHC ClassI molecules (Eichner and Radford, 2011). Ligand binding can also cause transitions from an unfolded state to a more ordered state. In an elegant study Daniels et al. (Daniels et al., 2014) demonstrate how the ligand pyrophosphate drives a conformational change of the disordered protein Bacillus RNaseP from unfolded to
folded. Such binding of ligands can provide an additional layer of protection against aggregation as seen in the case of Tau, where the microtubule binding region and the region implicated in aggregation are the same (Mukrasch et al., 2005). Lastly, the impact of binding on the ability of proteins to withstand mutations without succumbing to deleterious consequences places ligands in the category of “evolutionary buffers”; in the same manner as chaperones, ligands may allow proteins to mutate via steps that would otherwise be highly risk prone in an evolutionary pathway (Gershenson et al., 2014). To this effect a small summary of examples where ligand binding alters the kinetic partitioning of proteins between folding and non productive states such as aggregation or degradation is listed in Figure 2.6.

Figure 2.6 Other prominent studies showing how natural and artificial ligands bias the folding – aggregation – degradation kinetic competition for other proteins

The above table shows a few representative examples of how ligands have shown to alter the aggregation or degradation of the proteins they bind to. Essentially, the presence of a ligand is shown to favor the folded form, but our analysis shows that this is only possible because folding is not the rate limiting step for these proteins.
<table>
<thead>
<tr>
<th>Protein</th>
<th>Suppressor</th>
<th>Effect</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-2 microglobulin</td>
<td>Rifamycin SV</td>
<td>Prevention of oligomerization</td>
<td>Woods et al. 2011</td>
</tr>
<tr>
<td>p53</td>
<td>DNA aptamer</td>
<td>Prevention of aggregation</td>
<td>Ishimaru et al. 2009</td>
</tr>
<tr>
<td>Transthyretin</td>
<td>Thyroxine</td>
<td>Inhibition aggregation</td>
<td>Miroy et al. 1996</td>
</tr>
<tr>
<td>CRABP-I</td>
<td>Retinoic acid</td>
<td>Reduction of aggregation propensity</td>
<td>Ferrolino et al. 2013</td>
</tr>
<tr>
<td>Ataxin-3</td>
<td>Ubiquitin</td>
<td>Prevention of aggregation</td>
<td>Pastore et al. 2011</td>
</tr>
<tr>
<td>SOD-1</td>
<td>Multiple</td>
<td>Prevention of aggregation</td>
<td>Ray et al. 2005</td>
</tr>
<tr>
<td>Rhodopsin</td>
<td>Retinal</td>
<td>Reduction of aggregation and increase membrane trafficking</td>
<td>Saliba et al. 2002</td>
</tr>
<tr>
<td>FKBP-1</td>
<td>Shield</td>
<td>Reduction of protein degradation</td>
<td>Banaszynski et al. 2006</td>
</tr>
</tbody>
</table>
2.8. Materials and methods

- Cloning, expression, and purification of dihydrofolate reductase.

The wild type cysteine-free *E. coli* dihydrofolate reductase (DHFR) gene was a gift from the Matthews lab (University of Massachusetts, Medical School, Worcester MA) (Ionescu et al., 2000). It was subcloned into the pET28 expression vector using NdeI and BamHI restriction endonuclease sites. The -glycine-glycine- insertion between residues K106 and A107 was introduced by site-directed mutagenesis using a QuikChange protocol (Stratagene) yielding the “dDHFR” mutant. The vector was transformed into *E. coli* BL21 (DE3) cells. Cultures were grown at 37 °C to optical density at 600 nm (OD$_{600}$) of 0.6, induced with 1 mM final IPTG concentration and expressed at 37 °C for 4 h. The cells were harvested by centrifugation and re-suspended in buffer containing 20 mM HEPES, pH 7.4. The cells were lysed using a microfluidizer, and the protein was purified by refolding from inclusion bodies. Inclusion bodies were dissolved in 8 M urea and subsequently dialyzed to refold the protein, which was further purified using anion exchange chromatography on a DEAE column using a 0 M to 1 M sodium chloride gradient. After elution, the salt was removed by dialysis and the protein was flash frozen and stored at −80 °C until further use.
• *In vitro* trimethoprim binding assay for dDHFR.

The *in vitro* trimethoprim (TMP) binding assay was performed with 50 nm purified dDHFR following the protocol from Watson and co-workers (Watson et al., 2007). Briefly, dDHFR was incubated with increasing concentrations of TMP and the degree of quenching of dDHFR intrinsic tryptophan fluorescence was monitored at 345 nm. Spectra were corrected for any contribution to the fluorescence by TMP by running a parallel titration without protein. Fraction bound was calculated by using the formula below.

\[
F_{\text{bound}} = \frac{\text{fluorescence}_{\text{measured}} - \text{fluorescence}_{\text{max}}}{\text{fluorescence}_{\text{min}} - \text{fluorescence}_{\text{max}}}
\]

• Measurement of dDHFR thermodynamic stability.

dDHFR was titrated with increasing concentrations of urea (0 – 8 M) in the presence of a final TMAO concentrations of 0.7, 0.8, 0.9, 1, 1.2 and 1.5. The titrated samples were incubated at 37 °C over 10 hours. Fluorescence was collected from 320 to 380 nm after specific tryptophan excitation at 295 nm. Intensity at wavelength 344 nm was used to fit the data to the two-state equation (Bolen and Santoro, 1988) to obtain thermodynamic stability. All solutions were prepared with 20 mM HEPES pH 7.4, and the urea concentration was measured using its refractive index.

Cell growth, protein induction and partitioning for dDHFR expression experiments.
For all experiments involving partitioning of dDHFR, *E. coli* BL21(DE3) cells were grown in LB medium until mid-log phase (OD_{600} = 0.6). Cells were induced with 1 mM final IPTG concentration for 90 minutes. All growth and induction was performed at 37 ºC. The cells were harvested after equalization for OD_{600} by centrifugation at 4,000 x g for 2 minutes, the media was discarded, and the cells were lysed using BPER-II reagent at room temperature for 10 minutes with 1µg/ml final DNase I added to the lysis mixture. The sample was then centrifuged at 18,000 x g for 30 minutes to separate insoluble (pellet) and soluble (supernatant) components. The samples were subsequently prepared by boiling in SDS running buffer for running on 12% SDS-PAGE gels.

- **TMP titration in *E. coli* cultures.**

  TMP was dissolved in 100% methanol and then diluted to the desired concentration in media immediately after the addition of IPTG. Final methanol concentration did not exceed 7% v/v at the end of the experiment. Cells were harvested and partitioned as mentioned above.

- **Chloramphenicol shut off experiments.**

  Protein was induced for 90 minutes in the presence of 80 µM TMP. A fraction of the cells was harvested before any chloramphenicol was added and a second sample was collected after an hour of incubation with chloramphenicol at a final concentration of 50 µg/ml. The partitioning measurements were performed as described above. For the change of media experiment, a fraction of cells was harvested after the 90 minute induction of DHFR in the presence of 80 µM TMP. The remaining cells were centrifuged for 10
minutes at 4000 x g, washed with equal volume LB, centrifuged again and resuspended in LB supplemented with chloramphenicol at 50 µg/ml without any TMP. The second sample was collected after an hour of incubation with chloramphenicol and processed as described above.

• Mammalian cell growth, and experiments with α-galactosidase.

We seeded 10 cm culture dishes with HEK 293T cells and allowed them to reach ~80% confluency at 37 °C, 5% CO₂. We then transfected the cells with the R301Q α-GAL pCMV 3xFLAG-14 (R301Q α-GAL) vector using Lipofectamine 2000. At 24 hours post-transfection we trypsinized the cells and used the resulting suspension to seed 6-well plates for the DGJ titration experiment. The cells were allowed to adhere over a 4 hour incubation period, after which desired amounts of 1-deoxygalactonojirimycin (DGJ) or lactacystin (20 µM final concentration) were added. After 24 hours the cells were washed thoroughly with PBS, then lysed with ice cold lysis buffer (1% Triton X-100 in PBS). The lysate was centrifuged at 18,000 x g for 30 minutes and the soluble fraction was used for analysis. The samples were normalized prior to loading using the bicinchoninic acid (BCA) protein quantification method to ensure that equal amounts of protein were loaded in each case. The western blot was performed using as primary a polyclonal rabbit antibody against human α-GAL (purchased from GeneTex).

• Gel band quantification.

SDS-PAGE gels were stained with Coomassie G-250, scanned and analyzed on the LI-COR ODYSSEY CLx and quantified using the associated image studio software (version
4.0) (Luo et al., 2006). Western blots for the α-GAL experiments were similarly analyzed.
CHAPTER 3
INVESTIGATING THE IMPACT OF THE PROTEOSTASIS NETWORK ON THE FOLDING OF AN OBLIGATE CLIENT IN VIVO

This chapter describes my investigation of the folding, misfolding and aggregation properties of an obligatory chaperone client protein inside the cell. Using a combinatorial approach of computational modeling and experiments designed to test the impact of chaperones on the folding of the client protein at varying expression levels we have uncovered how such proteins fold in vivo. The work in this chapter was in collaboration with Todd Morse, Ha Dang and Evan Powers.

3.1. Chaperone assisted protein folding inside the cell

Besides aiding in stress responses, preventing protein aggregation and at times reversing it, molecular chaperones often play a vital role in protein folding (Kim et al., 2013). An important aspect of protein folding inside the cell is the dependence of certain proteins on molecular chaperones to attain the native/folded state. This reliance on chaperones, which involves a physical interaction with the chaperones steers the protein away from deleterious aggregation states and towards a more productive native state. As one can now imagine a kinetic competition for protein sequence (see Chapter 1), chaperones bias the competition to favor the formation of the native state from the unfolded state (which we consider the progenitor of all states inside the cell). In our model organism E. coli there are two major chaperone families that can selectively favor folding over non-
productive outcomes such as degradation and aggregation. The first is the Hsp60 family member GroEL and its co-chaperone GroES (Kerner et al., 2005; Kim et al., 2013). This chaperone assembles in a tetradecameric ring creating a cavity that can be occupied by a protein molecule undergoing the folding reaction. There remain many open-ended questions pertaining to the molecular mechanism of the action of GroEL. For our purposes we assume that GroEL is a channel that funnels proteins from the unfolded state to the native state in any of the two predominant modes of its action (passive folding cage/active substrate interaction leading to folding). The second set of chaperones is the Hsp70 family, with DnaK as the *E. coli* member, which works with its associated co-chaperones DnaJ and GrpE (Calloni et al., 2012; Kim et al., 2013). There is little direct evidence that DnaK directly facilitates the formation of the native state from the unfolded state, i.e. it does not directly affect the folding rate; rather DnaK seems to prevent transitions into off-pathway misfolded states and provide an escape route from such a misfolded state. This role contributes to an effective increase in the concentration of the folded native product (Clerico et al., 2015). Proteomic analysis has yielded a list of ~250 proteins that can interact with GroEL, and of these 80 proteins in the *E. coli* proteome which are unable to fold without the assistance of the GroELS chaperone system (Kerner et al., 2005). Such proteins have been termed as class III or obligate GroEL substrates. Furthermore, a depletion of GroELS from cells results in the accumulation of client proteins on the DnaK chaperone (Kerner et al., 2005). These interesting observations lead us to investigate the how these two major chaperone systems work together to achieve folded protein *in vivo*. An interesting question arises from these observations: what are the biophysical properties of the obligate chaperone-dependent client proteins?
Specifically, what are the biophysical folding/misfolding properties of such client proteins in terms of folding and unfolding rate and equilibrium constants, and how do these constants compare with one another for an obligatory chaperone client? Furthermore, can we estimate these properties from cellular experiments and observed dependence on chaperones? This would be useful information, particularly considering the fact that refolding these proteins in vitro without chaperone assistance is extremely challenging. These questions are all part of understanding how the biophysical properties of proteins interface with the proteostasis network inside the cell such that folding is optimized and deleterious consequences such as aggregation are avoided.

3.2. What can we hypothesize about proteins that depend on chaperones in vivo?

Kinetic competition is a central theme underlying protein folding inside the cell. Once a protein is synthesized, the intrinsic propensities to fold, misfold and aggregate will dictate the trajectory of that protein inside the cell. One class of the major factors that influences these trajectories are molecular chaperones and degradation enzymes (described above). Using a computational model (FoldEco) developed by our lab in collaboration with Evan Powers, we asked what types of proteins would be most aided by different classes of molecular chaperones (Powers et al., 2012)? In this calculation, a test dataset of 4000 proteins of varying biophysical properties (within published limits) was created. In the modeling these proteins were synthesized at various synthesis speeds using the FoldEco program in the absence of any molecular chaperones. FoldEco tracks the time evolution of pools of protein in different states of the protein including native, misfolded and aggregated as well as chaperone-bound, and the program was used to generate the
concentration of protein that aggregated at the end of the experiment. In a second experiment with the same protein set, the molecular chaperones DnaK (the KJE system) and GroEL (the ELS system) were reintroduced individually, and the fate of the aggregated pool of proteins was similarly tracked. Based on these simulation experiments, it was inferred that at slower synthesis rates the presence of DnaK lowered the aggregation of those proteins that had a high propensity to misfold compared to the average misfolding tendency of the set of proteins tested. Specifically the proteins most helped by DnaK had a high $k_m$ (misfolding rate constant). Similarly, the presence of GroEL seemed to have its greatest effect on those proteins that folded much more slowly than the average folding speed of the set of tested proteins. However, at faster synthesis rates, the roles of GroEL and DnaK were seemingly very similar, and slow folding proteins were aided by DnaK as well as GroEL, whereas proteins with a higher propensity to misfold were also helped by both chaperone systems. Another experiment from the same study, asked the question of how the presence of degradation machinery, specifically the Lon protease affected the folding of proteins. Up-regulation of Lon, as the study suggests and predicts had an impact on the total protein concentration, but had a stronger impact on the misfolded species as it accumulated (Figure 3.1). Taken together, these simulations suggest a synergistic action of the proteostasis components to shuttle substrates from the unfolded (U) state to the native state while simultaneously lowering the population of the protein that accumulates in the misfolded (M) state either by prevention of the formation of the M state or by its degradation.
Thus, we hypothesize that a protein that requires GroEL to fold successfully (a class III substrate as classified by Kerner et al. (2005)) (Kerner et al., 2005), is a slow folding protein, i.e. with a low $k_f$ (rate constant of folding). Based on how the protein aggregates, one can then determine what the misfolding and aggregation rate constants would be relative to the folding rate constants. Additionally, if one can introduce probes for the unfolded state (such as GroEL) or for the misfolded state (such as Lon) or perhaps both (such as DnaK) based on the hypotheses generated by the FoldEco simulations, one can start to dissect the folding behavior of an obligate chaperone client inside the cell.

In experimental studies performed by Cho et al. (Cho et al., 2015) on studying the impact of the proteostasis network on protein folding inside *E. coli* the FoldEco model was useful to back calculate some of the biophysical parameters of the proteins in the study. And while that study was informative in terms of how chaperone systems can work together to benefit folding, it did not address the folding behavior of a naturally occurring protein that has co-evolved over millennia with cellular chaperones and to fold under their influence.
Figure 3.1 Hypotheses for chaperone functions based on computational modeling

A. Based on the FoldEco simulation study with 4000 proteins with a distribution of folding, misfolding and aggregation rate constants, at slower synthesis rate proteins with high misfolding were prevented from aggregation by the KJE system, similarly slow folding proteins were helped by GroELS.

B. At increased synthesis rate the effects of the GroELS and DnaK were seen on proteins with similar properties of folding and misfolding. Both GroEL and DnaK reduced the aggregation of those proteins with low folding rate constants and had a higher than average misfolding propensity.

C. Lon has ability to suppress aggregates by reducing protein concentration. The concentration of protein reduced by lon depends on the presence of the other chaperones that can act to sequester the protein away from. Figure adapted from Powers et al (Powers et al., 2012).
A.

**Slower synthesis**

<table>
<thead>
<tr>
<th>Top KJE substrates</th>
<th>Top ELS substrates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_m$ (s$^{-1}$)</td>
<td>$k_m$ (s$^{-1}$)</td>
</tr>
<tr>
<td>$K_m$ (μM)</td>
<td>$K_m$ (μM)</td>
</tr>
<tr>
<td>$k_f$ (s$^{-1}$)</td>
<td>$k_f$ (s$^{-1}$)</td>
</tr>
<tr>
<td>$K_f$ (μM)</td>
<td>$K_f$ (μM)</td>
</tr>
</tbody>
</table>

B.

**Faster synthesis**

<table>
<thead>
<tr>
<th>Top KJE substrates</th>
<th>Top ELS substrates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_m$ (s$^{-1}$)</td>
<td>$k_m$ (s$^{-1}$)</td>
</tr>
<tr>
<td>$K_m$ (μM)</td>
<td>$K_m$ (μM)</td>
</tr>
<tr>
<td>$k_f$ (s$^{-1}$)</td>
<td>$k_f$ (s$^{-1}$)</td>
</tr>
<tr>
<td>$K_f$ (μM)</td>
<td>$K_f$ (μM)</td>
</tr>
</tbody>
</table>
C.

![Graphs showing the relationship between synthesis rate and the minimum [Lem] needed to suppress aggregation with and without chaperones.](image)

- **A. Without chaperones**: The graph illustrates the native state concentration (µM) of proteins, fraction of proteins degraded by Lon, and heat shock [Lem] for different synthesis rates (µM s⁻¹) and percentage of aggregates. The x-axis represents the synthesis rate, while the y-axis represents the minimum [Lem] needed to suppress aggregation.
- **B. With chaperones**: Similar to A, but with the addition of chaperones, showing a lower minimum [Lem] required for aggregation suppression.
3.3. Choosing an experimental system

We chose to work in *E. coli* BL21(DE3) for experimental tractability and the fact that it lacks the Lon protease allowing us to study chaperone dependence in a background without proteolysis (which for FoldEco we assume is performed by Lon). In our lab, we have methods to upregulate individual chaperone components by mean of a plasmid provided in trans to the genetic copy of the chaperone or the degradation enzyme lon (BL21(DE3) lacks a functional Lon protease thus making it easier to determine the effects of varying concentrations of Lon on the substrate in the presence of other chaperones). Much of the power of FoldEco also comes from the biochemical investigations that have been carried out over the past several years on the chaperone systems of this organism. As there are only two major chaperone systems, GroEL and DnaK, we do not have to worry about redundant versions of the chaperone systems playing a major role in the folding of our test protein.

For our test protein we chose a native *E. coli* protein S-adenosylmethionine synthase (gene name metK). MetK is an enzyme with a transferase function (E.C. 2.5.1.6) (UniProt, 2015). Specifically it catalyzes the formation of S-adenosyl methionine using ATP and Methionine as substrates. MetK is a cytosolic, soluble 384 amino acid protein with a monomeric molecular weight of ~42 kDa. MetK assembles into homotetramers as a final quarternary structure (Figure 3.2) (Komoto et al., 2004). MetK has been shown to
be a class III GroEL substrate. In *in vitro* refolding studies by the Hartl lab (Kerner et al., 2005), MetK is unable to fold and function when introduced into refolding conditions in the absence of the GroELS system. Refolding in the presence of GroELS alone restores MetK activity to ~70% of native MetK whereas refolding in the presence of DnaKJE and GroELS restores up to 90% activity of the protein. Refolding in the presence of the KJE system alone, however, does not restore any measurable activity. These *in vitro* tests indicate that MetK obligatorily requires an interaction with GroEL to reach the native state. MetK on a pET 29b plasmid was kindly provided to us by the Hartl laboratory (Max Planck Institute, Martinsried Germany) and the proteostasis components were available on arabinose inducible pBAD plasmids that were provided from previous FoldEco studies conducted jointly with the Kelly/Powers laboratories (Cho et al., 2015; Kerner et al., 2005).

**Figure 3.2 Model test protein MetK**

A. The choice of test protein is MetK (PDB 1RG9), monomer and tetramer shown above. MetK is a homotetramer and GroEL catalyzes the folding of this protein. For our experimental consideration, we treat the species subject to proteostasis treatment to be the monomer.

B. The above cartoon shows the reaction catalyzed by MetK. The enzyme is essential in *E. coli* and is responsible for the biosynthesis of S-Adenosyl methionine.
A.

MetK monomer \rightleftharpoons \text{MetK tetramer}

B.

\text{Methionine} + \text{ATP} \rightarrow \text{S-Adenosyl methionine}
3.4. MetK – Equilibrium folding properties of MetK

To understand the folding properties of MetK, i.e. the folding rate constant, unfolding rate constant, misfolding rate constant and the aggregation rate constant in the cellular context we need to first dissect the contribution of tetramerization to the folding fate of MetK. Tetramerization can be regarded as self-liganding mechanism for MetK as is seen in the case for transthyretin and other multimeric enzymes. To determine the tetramerization dissociation equilibrium constant, we performed equilibrium denaturation experiments with guanidine hydrochloride as a function of MetK concentration. MetK denaturation shows a three-state equilibrium unfolding curve and the midpoint of the first transition shows a concentration dependent sensitivity to guanidine induced denaturation, whereas the midpoint of the second transition is insensitive to MetK concentration at higher guanidine concentrations. This observation is consistent a dissociation of the tetramer at low guanidine concentrations causes while the denaturation of the monomer occurs at higher concentrations of the chaotrope. By performing guanidine denaturation experiments as a function of MetK concentration we were able to determine two equilibrium parameters for MetK. The first was the limiting tetramerization concentration. This was determined as the concentration of MetK at which the first transition was no longer seen in the guanidine denaturation experiment. The second parameter was the apparent unfolding equilibrium constant, which would theoretically be the ratio of the unfolding to folding rate constants. This number could serve as a guide to
help fit some of the MetK rate constants using FoldEco. From our experiments, the
limiting tetramerization concentration was ~0.13 µM as at that concentration the first
transition was not seen for MetK denaturation (Figure 3.3A). This suggests that even at
low concentrations, MetK readily forms the tetramer and once it is in the tetrameric state,
it can be assumed to be resistant to aggregation that would initiate as consequence of
tetramer dissociation and unfolding (see chapter 2 for how ligands prevent aggregation).
The second equilibrium constant we obtained was the apparent unfolding equilibrium
constant which was $9 \times 10^3$, indicating that the monomer of MetK was stable as the
folded species outnumbered the unfolded species by roughly 10000 molecules to 1
(Figure 3.3B). It is important to note here that the unfolding equilibrium constant is an
apparent constant, since the reaction is not reversible. Effectively, folding does not occur
without GroEL assistance.
Figure 3.3 Equilibrium denaturation properties of MetK

A. The family of curves depicts MetK denaturation over a range of concentrations (brown – 0.136, red – 0.272, green – 0.408, blue – 0.572 and pink – 1 µM concentration of tetrameric protein respectively). One can observe two transition points, one under 1M where Cm changes with MetK concentration and one at approximately 2.5 M where Cm does not change. The first transition which represents tetramer dissociation disappears at 0.13 µM thus allowing us to determine the equilibrium constant for tetramer dissociation.

B. Denaturation of monomeric MetK. This set of points is derived from the full denaturation, this transition does not change with MetK concentration. The equilibrium unfolding constant for the monomer (K_u) was calculated to be – 9 x 10^3 by fitting the data to the two-state equation to estimate folding free energy (Bolen and Santoro, 1988).

C. Model for equilibrium MetK unfolding. We propose a model for MetK denaturation where dissociation of the tetramer is the first step and it is limited by the concentration of MetK and the second step is the denaturation of the monomer. Monomer denaturation is shown irreversibly as the protein depends on GroEL to get to the folded state.
B.

![Graph showing the wavelength (max) as a function of [GdnCl] molar concentration. The x-axis represents the concentration of [GdnCl] in molar, ranging from 1 to 4, and the y-axis represents the wavelength (max) ranging from 338 to 355 nm. The graph shows a transition from the native/folded state to the denatured/unfolded state.]

C.

![Diagram illustrating the transition from MetK tetramer to MetK monomer and ultimately to the unfolded state. The diagram includes a native/folded state and a denatured/unfolded state, with arrows indicating the direction of the transition.]
3.5. MetK overexpression and aggregation at “basal” proteostasis levels.

To determine the nature of the competition between folding and aggregation, we overexpressed MetK in *E. coli* BL21(DE3) cells. We examined the folding fates of MetK by quantifying the total amount of MetK synthesized and then interrogating what fraction of the protein produced became insoluble after our synthesis time. It is important to note here that as there is no degradation component yet, MetK has two possible folding fates; either it folds correctly (with the help of GroEL) or it misfolds, self-associates and aggregates, which is to say it escapes any positive interaction with GroEL and has misinteractions with other MetK molecules. We use solubility or insolubility after expression as a readout after protein synthesis to determine where MetK partitioned after production. We assume that soluble MetK is well folded, but the soluble MetK could be a mixture of well folded, chaperone bound or a ligand bound (MetK binds ATP, methionine and the product S-adenosyl methionine) species, a partition that presently evades separation. However, the insoluble MetK is obtained from inclusion bodies, which contain the misfolded MetK that aggregates inside the cell after synthesis. Thus we use the aggregated fraction (*F*<sub>agg</sub> or *Fraction*<sub>aggregated</sub>), defined as the ratio of the intensity of the aggregated fraction to the intensity of the total protein synthesized as our readout for the folding of MetK (see equations below). The intensity of the total and aggregated fraction was calculated after running the samples on a 12% SDS-PAGE gel and staining with Coomassie G-250 stain. Stained gels were imaged using the LI-COR system and were analyzed using the associated gel analysis software.
As mentioned earlier, the MetK gene was subcloned into a pET29b vector, with protein expression driven from an isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible T7 promoter. To investigate the folding-aggregation competition of MetK, we varied its concentrations inside the cell by titrating increasing amounts of IPTG (from 1 µM to 200 µM final concentrations) into cells grown to mid-log phase (OD600 ~ 0.6). By comparing total cell lysates to purified MetK standards, we were able to calculate the amount of MetK synthesized at each concentration of IPTG. We observed that the MetK concentration ranged from 170 µM to approximately 2000 µM over the range of IPTG concentrations we used. We also observed that protein concentration peaked at ~2000 µM and did not rise further with increase in concentration of IPTG (Figure 3.4A). The concentrations reported are MetK monomer concentrations and not the tetramer. For our experiments we assume that the monomer is the species acted upon by the cellular proteostasis network. We then calculated the concentrations of the chaperones DnaK and GroEL present at our growth conditions, again by comparison with standard purified amounts loaded on SDS-PAGE gels. The aggregated fraction of MetK was measured after harvesting cells at the end of the protein induction time, followed by cell lysis by sonication and separation of soluble and insoluble components by high-speed centrifugation. The aggregated fraction of MetK increased with an increase in the
concentration of MetK synthesized (not surprisingly, as aggregation is concentration dependent). The fraction aggregated (measured from 0 to 1) rose from an average of 0.2 at an intracellular concentration of MetK of 170 μM MetK to an average of 0.45 at approximate [MetK] of 2000 μM (Figures 3.4 A, B). For these experiments performed at what we term “basal” levels (no chaperone overexpression), the concentrations of GroEL (as an active tetradecamer) and DnaK were calculated to be 130 and 70 μM on average Figures 3.8A, B see pages 120-123).

**Figure 3.4 Synthesis and aggregation of MetK at basal chaperone conditions**

A. The above figure shows the concentration of MetK synthesized at varying IPTG levels under “basal” levels of chaperones. MetK levels rise with IPTG concentration until 200uM, then plateau. The concentration in cell measured by comparison with purified standards.

B. Fagg of MetK as a function of concentration. MetK Fagg increases with increase in concentration from 0.25 to 0.47 under “basal” chaperone conditions. Due to a limitation of MetK synthesized the aggregated fraction does not rise further.
### Table A

<table>
<thead>
<tr>
<th>[IPTG] µM</th>
<th>[MetK] µM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>6.1</td>
</tr>
<tr>
<td>10</td>
<td>8.2</td>
</tr>
<tr>
<td>50</td>
<td>10.2</td>
</tr>
<tr>
<td>200</td>
<td>12.3</td>
</tr>
<tr>
<td>500</td>
<td>14.4</td>
</tr>
<tr>
<td>1000</td>
<td>16.4</td>
</tr>
</tbody>
</table>

### Diagram A

- Bar graph showing [MetK] µM intensity at different [IPTG] µM concentrations.
- Line graph showing intensity (AU) vs. [MetK] µM concentration with an R² value of 0.9967.
B.
Our observations with MetK folding/aggregation at these basal chaperone conditions allow us to draw some important conclusions. First, even at nearly stoichiometric amounts of MetK to GroEL (at low concentrations), MetK aggregates inside the cell. This suggests that the intrinsic propensity of MetK to aggregate is higher than the association rate with GroEL (and perhaps even the equilibrium association with GroEL). Thus if we were to now begin to compare the folding and misfolding rate constants for MetK (which we suggest are the major drivers for the distribution of MetK), at a first approximation the ratio of the misfolding/aggregation rate constants to the folding rate constant is larger than 1 (As $k_m, k_{agg} > k_f$). The next set of experiments involves the overexpression of either the GroELS system or the DnaKJE system or the Lon protease to determine how the fluxes between folding and aggregation change when you have mechanisms to selectively target one or more states. Knowledge of its folding outcomes in the presence of lon and the DnaKJE system would refine our understanding how the other biophysical parameters – misfolding and aggregation would be compared to the folding parameters – we plan to bolster our experimental findings with computational modeling to achieve a more quantitative understanding of the system. Furthermore, these experimental readouts can then be used as inputs for FoldEco simulations to predict what the biophysical parameters for MetK might be inside the cell.
3.6. Effect of GroELS titration on MetK aggregation

(These experiments were performed in collaboration with Ha Van Dang – Mount Holyoke College)

Since GroEL serves as a conduit that funnels MetK from the unfolded state to the folded state, we hypothesized that increasing the concentration of GroELS would enhance the efficiency of MetK folding and lead to a diminution of the aggregated MetK population. We co-transformed *E. coli* BL21(DE3) cells with the MetK plasmid and a GroELS plasmid, with the expression of GroELS driven by an arabinose inducible promoter and that of MetK by an IPTG inducible T7 promoter. GroELS was titrated to varying concentrations using concentrations of arabinose from 0.002 to 0.2% (w/v final concentration) (Figure 3.8A). The range of [arabinose] used allowed us to test [GroEL] from the basal level of ~130 µM to as high as approximately 1600 µM. For these experiments, we first grew the co-transformed cells to an OD$_{600}$ of 0.3 at which point we induced GroELS with the desired amount of arabinose. After 1 generation of growth at OD$_{600}$ of 0.6 we induced MetK with varying IPTG concentrations. Determination of total MetK and fraction MetK aggregated was performed similarly.

We made two major observations regarding MetK synthesis and aggregation with the GroEL titration experiment. First, as GroELS intracellular concentrations increased, the aggregation of MetK decreased. This result is not surprising as MetK is an obligate GroEL substrate. It is important to note here that at no point is the aggregation of MetK completely abolished. There appears to be a limit of about 0.15 fraction MetK aggregated regardless of the concentration of GroELS (Figure 3.5A). At basal concentrations of
GroEL (~130 µM), MetK synthesized at about 2000 µM has an average fraction aggregated of ~0.45. However, at the same concentration of MetK, the higher level of GroEL (~1600 µM) reduces the fraction aggregated to ~0.2. The observation that MetK aggregation is not entirely abolished even at high GroEL levels (MetK ~ 170 µM, GroEL ~ 1600 µM) allows us to suggest that the intrinsic misfolding rate of MetK ($k_m$) is larger than the association rate constant between MetK and GroEL (and thus the rate constant of GroEL mediated folding) (Figure 3.5C). This lower limit on aggregation is similar to the one observed for dDHFR observed in Chapter 2 and also suggests a chaperoning limit akin to the limit of the small molecule ligand which depends on the intrinsic ratio of the folding to aggregation rate constants ($k_f/k_{agg}$).

The second observation was the fluctuation of MetK concentration at high levels of expression of GroEL. While the concentrations of MetK remained relatively unchanged at higher levels of MetK induction (in the presence of all concentrations of GroEL induced), we observed a reduction of the concentration of MetK synthesized at lower levels of induction (at lower IPTG concentrations). Generally we observed a reduction in MetK synthesis at low levels of IPTG (particularly 5 and 10 µM, which under basal conditions yields three-fold more MetK) and high levels of arabinose (at 0.02 and 0.2% w/v where we achieve a GroEL concentration roughly 9 to 12 times of our basal level) (Figure 3.5B). This observation indicates that there is a competition for protein synthesis machinery, and that though we used orthogonal promoters to induce MetK and GroEL, the protein synthesis outcome depends on the partitioning of the protein production resources inside the cell. Interestingly, although not pursued, is the debunking of the
notion that an IPTG inducible T\textsubscript{7} promoter essentially overrides the protein synthesis system preventing the translation of other proteins. This example of MetK – GroEL (recollect that GroEL is induced with arabinose before MetK induction with IPTG) is a good demonstration that the T\textsubscript{7} driven expression can be affected by the production of protein driven by an alternative promoter.

**Figure 3.5 Impact of GroEL expression on MetK synthesis and aggregation.**

A. The SDS PAGE gels show that higher GroEL concentrations lower the aggregation of MetK at equivalent MetK concentrations (synthesized with 50µM IPTG). The F\textsubscript{agg} never reaches zero even at sub-stoichiometric amounts of MetK compared to GroEL indicating a propensity to aggregate faster than association with GroEL.

B. GroEL induction at high levels causes diminution of MetK levels when MetK is induced with low concentrations of IPTG likely due to competition for protein synthesis machinery. C – trends observed for MetK aggregation as a function of GroEL and MetK concentrations at the GroEL concentration Data and figures courtesy Ha Van Dang.

C. The chart shows the concentration of MetK and the fraction aggregated at different average GroEL levels (top left corner – indicated by colours from red to blue). MetK aggregation generally reduces with increasing GroEL concentrations, however there is no concentration of GroEL at which MetK aggregation is entirely abolished. The error bars represent an SEM of three experiments.
A. 

[Arabinose] % w/v

<table>
<thead>
<tr>
<th></th>
<th>0.005</th>
<th>0.02</th>
<th>0.05</th>
<th>0.2</th>
<th>0</th>
</tr>
</thead>
</table>

[ITPG] 50 μM - 1424 μM [MetK] average

B. 

[Arabinose] % w/v

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0.005</th>
<th>0.02</th>
<th>0.05</th>
<th>0.2</th>
</tr>
</thead>
</table>

[ITPG] 10 μM
C.

![Graph showing fraction of aggregated MetK vs. average MetK monomer concentration with different GroEL concentrations: 130 μM, 950 μM, 1260 μM, and 1540 μM.](image)
3.7. Effect of DnaKJE on MetK aggregation

From the GroELS titration and the MetK aggregation at stoichiometric and higher amounts of GroELS, it is clear that MetK has an intrinsic tendency to populate an off-pathway aggregation-prone misfolded state faster than it can encounter GroEL. This indicates that MetK populates the unfolded state (the precursor of all states), which partitions into the misfolded state (the same one that arises from the unfolded progenitor pool) and the state that binds GroEL. Based on the hypothesis presented earlier, the presence of DnaK should aid in the correct folding (and thus lower the aggregation) of proteins with a higher propensity to misfold. Experimental data from the Hartl lab (Calloni et al., 2012) also suggests that GroEL substrates accumulate on DnaK upon depletion of GroELS inside the cell. This too indicates a synergistic tendency of DnaK to bind unfolded or misfolded proteins and prevent aggregation. Similar to the above experiments with GroELS, we co-transformed *E. coli* BL21(DE3) cells with the MetK plasmid and the plasmid containing genes for DnaK, DnaJ and GrpE. Also similar to GroELS, the expression of the components of the KJE system was driven by arabinose. The KJE components and MetK were induced by IPTG, similar to the protocol used for the GroELS and MetK experiments.

Mimicking the observation in the GroEL induction experiments, the biosynthesis of MetK showed a diminution at high concentrations of arabinose and low concentrations of
IPTG, although the reduction in MetK synthesis was not as drastic as the one observed with the GroEL experiment. By varying the concentration of arabinose (from 0.002 to 0.2% w/v final concentration), we could increase the concentration of DnaK from ~70 µM at basal (no addition of arabinose) level to ~700 µM (approximately 10-fold higher) at the highest concentration of arabinose (0.2% w/v) used (Figure 3.8B). Over the range of DnaK concentrations we tested, we observed a general decrease in the fraction of MetK aggregated across nearly all MetK concentrations (Figure 3.6C). The decrease in MetK aggregation also scaled with the increase in the concentration of DnaK, i.e. as we synthesized more DnaK inside the cell, we observed a reduction in the aggregated fraction at almost all concentrations of MetK produced, and this aggregated fraction further diminished as DnaK concentrations rose (compare Figure 3.6A and B, compare across concentrations of MetK and DnaK where DnaK concentrations increase from Blue to Red and the MetK concentrations are represented on the X axis). The aggregated fraction at lower levels of expression of MetK (170 µM) reduced from ~0.25 to ~0.1 at the highest concentration of DnaK produced inside the cell. Similarly, the fraction MetK aggregated when approximately 1500 µM was synthesized also reduced from an average of 0.35 to 0.15. It was only at the highest levels of MetK synthesis that the aggregated fraction of MetK was not reduced to a large extent (last concentration point shown in Figure 3.6B, all colors). What does this tell us about the impact of DnaK on the folding of MetK? First, as we know that MetK folds in a GroEL-dependent manner, it is not the overexpression of DnaK that aids the folding reaction directly. However, given enough DnaK, the cellular pools of GroELS (which did not appear to increase during these experiments) become sufficient to achieve a high proportion of folded MetK. This
suggests co-operation between the DnaK and GroELS chaperones. It appears that DnaK can form a complex with MetK to protect it from aggregation, and then that DnaK releases MetK in a form that is ushered by GroEL to the folded state. It should also be noted that at no concentration of DnaK was the aggregation of MetK (at any concentration) entirely abolished. All these observations are consistent with the hypothesis that proteins with a high propensity to misfold are aided by the KJE system to avoid aggregated states, and importantly this role of DnaK can work in concert with the action of the GroELS system for proteins that are obligate GroEL substrates.

**Figure 3.6 Impact of DnaKJE on the folding and aggregation of MetK.**

A. DnaK at high levels ~ 650 uM strongly inhibits MetK aggregation as seen by the low intensity aggregation bands at nearly all MetK concentrations on the SDS PAGE gel.

B. Similar to GroEL, as DnaK concentration increases (top right corner shows the average DnaK concentration at the levels of MetK synthesized) the MetK aggregate fraction reduces, yet never going to 0. DnaK helps prevent aggregation and allows what little GroEL (estimated at 130 µM average) is present inside the cell to successfully fold a large concentration of MetK.
A.
B.

The graph shows the fraction of aggregated MetK as a function of the average MetK monomer concentration, with different symbols and colors representing various concentrations of DnaK: 70 μM (blue circles), 180 μM (green squares), 460 μM (orange triangles), and 660 μM (red diamonds). The error bars indicate the variability in the measurements.
3.8. Effect of Lon titration on MetK aggregation in basal levels of KJE and GroELS

In addition to the folding limb of the proteostasis network, the other component of protein quality control that keeps aggregation under check is the degradation machinery. For proteostasis in *E. coli* the degradation machinery proposed to do this function is the Lon protease. Experimental and modeling results suggest that while Lon decreases the total concentration of a protein, this also causes a reduction in the amount of protein that aggregates. There are two species that are believed to be degraded by the Lon protease are the unfolded and misfolded state (Gur and Sauer, 2008; Gur et al., 2012; Powers et al., 2012). Modeling of the fraction degraded and fraction aggregated data of the model proteins used by Cho et al. (Cho et al., 2015) suggests that Lon targets either the unfolded state or the misfolded state depending on how much the protein substrate intrinsically populates the off-pathway misfolded state. For the DHFR mutant used in that study, the effect of the Lon protease was on the unfolded state whereas the effect for the mutant Cellular Retinoic Acid Binding Protein I (CRABPI) Y133S was on the misfolded state. BL21(DE3) has no functional Lon protease. Up to now, all our investigations were in the absence of any degradation caused by the Lon protease. Thus, re-introduction of Lon at varying concentrations allowed us to directly test the effect of the protease in the background of basal GroEL and DnaK conditions. Similar to the protocol used with the GroELS and DnaKJE systems, the *E. coli* Lon gene was cloned into a plasmid where its expression is controlled by arabinose. The experiments with Lon induction and protein partition were performed similarly to the GroELS and DnaKJE experiments.
Lon concentrations increased with increasing arabinose concentrations added to the cells. The concentrations rose from ~3 µM hexameric (the active state is a hexamer) Lon at an arabinose concentration of 0.002% final w/v to an average of 23 µM Lon at the highest concentration (0.2% w/v final) used (Figure 3.8C). Over the range of Lon concentrations tested, MetK concentrations at all levels of IPTG were depleted relative to that in the absence of Lon (Figures 3.7A, B). For example, the final concentration of MetK for the same addition of IPTG was reduced by half when Lon was present at 23 µM (compare blue and red points in Figure 3.7C). Furthermore, the aggregated fraction was also diminished at higher Lon concentrations. This is not a novel observation as a reduction in protein concentration diminishes the concentration of the aggregation prone species. However, we clearly observed that at nearly equal concentrations of MetK (see data points between 600 and 1100 µM MetK), an increase in Lon concentration reduces a larger amount of the aggregated protein (Figure 3.7C). This is indicative of a mechanism where Lon facilitates the clearance of a species en route to aggregation. For MetK, it appears that the presence of Lon has an impact on the misfolded species as it selectively reduces the aggregated fraction within the total protein population. This observation is consistent with the hypothesis that Lon degrades off-pathway misfolded proteins (Powers et al., 2012). Also consistent with the impact on Lon in BL21(DE3) cells, cells expressing a high concentration of Lon also grew more slowly than cells with little or no Lon present. This is likely due to the degradation of the yoeB antitoxin which is the cognate antitoxin in the yefM – yoeB toxin – antitoxin pair, which has been implicated in biofilm
formation and regulation of its own gene products via transcriptional attenuation (Christensen et al., 2004).

**Figure 3.7 Impact of lon on MetK synthesis and aggregation.**

A. The SDS PAGE gel shows the impact lon at medium levels ~ 17 uM. This concentration of lon reduces concentration of MetK synthesized at the same concentration of IPTG added under the basal conditions (compare with Figure 3.4). It also reduces the aggregated fraction, not surprising as total levels of MetK are lowered.

B. Increase in lon concentrations (from Blue to Red, top right corner shows average values) lowers MetK synthesized proportionally with the lon synthesized. Similar concentrations of MetK between (700 – 1100 uM MetK) show altered aggregated fraction and higher the lon levels, lower the Fagg at these concentrations. This indicates that lon degrades an aggregation prone, off-pathway misfolded state.
Figure 3.8 Concentrations of chaperones produced as a function of arabinose concentrations.

A. GroEL synthesized during the GroELS titration experiments is measured by comparison against purified GroEL.

B. Similarly DnaK synthesized at the various DnaKJE titration experiments is measured by comparison with purified proteins.

C. Lon concentrations are also measured by comparison with purified standards. Purified GroEL, DnaK and lon were gifts from Ivan Budyak, Joseph Tilitsky and Rilee Zeinert respectively.
A.

![Image of gel electrophoresis](image)

<table>
<thead>
<tr>
<th>[GroEL] μM</th>
<th>% [Arabinose]</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>0.005</td>
</tr>
<tr>
<td>15</td>
<td>0.02</td>
</tr>
<tr>
<td>7.5</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

MWt. kDa

GroEL

MetK

Intensity (AU)

[GroEL] M

% [Arabinose]
C.

\[
\begin{array}{cccccccc}
2 & 1.5 & 1 & 0.5 & 0.25 & 0.125 & 0.002 & 0.02 & 0.2 \\
\text{[Lon] } \mu\text{M} & & & & & & & & \\
\% \text{ [Arabinose]} & & & & & & & &
\end{array}
\]

![Image](image_url)

![Graph](graph_url)
3.9. Modeling the data with FoldEco

(The mathematical modeling was performed by Evan Powers)

To understand the folding behavior of MetK inside the cell, we used FoldEco to fit our experimental data. The major parameters in FoldEco that can be manipulated are the equilibrium constants and kinetic rate constants for the folding, misfolding and aggregation reactions. The FoldEco simulation was performed iteratively to match the experimental dataset. The best fit between the FoldEco simulation and the experimental data showed a root mean squared deviation of 0.13 and allowed us to infer three important conclusions. The first, is that the ratio misfolding rate constant to the folding rate constant ($k_m/k_f$) is 2 which is in good agreement with our hypothesis from section 3.5 where we state $k_m$ is larger than $k_f$. Specifically the FoldEco fitting allows us to approximate the values of $k_m$ and $k_f$ to 0.02/s$^{-1}$ and 0.01/s$^{-1}$ respectively. These values are also in good accord with the predictions made by FoldEco about proteins that rely on the GroELS system (and MetK is an obligate substrate). The second conclusion is that aggregation is almost instantaneous after accumulation of the misfolded state suggesting a low life time for this off-pathway state. Third, the fit also does not converge well on the equilibrium constants again suggesting that the partitioning of MetK inside the cell is under kinetic control rather than thermodynamic control. In Figure 3.9 we see a general schematic for the kinetic partitioning for MetK and using this model (that is essentially how FoldEco describes the kinetic partitioning of proteins) we can now being to see how the folding and misfolding reactions compete with each other in the presence of chaperones and how chaperones change the folding outcome even when the misfolding rate constant is larger than the folding rate constant.
Figure 3.9 The kinetic partitioning of MetK upon synthesis in *E. coli*

The figure below shows the general kinetic partitioning scheme for MetK. S, U, N, D, A and M represent the synthesized, unfolded, native, degraded, aggregated and misfolded populations of MetK. The rate constants for the individual reactions is denoted by $k_{\text{reaction}}$. Folded MetK is represented as a monomer ($N_{\text{mono}}$) or a tetramer ($N_{\text{tetramer}}$). GroEL catalyzes the transition from U to N. DnaK is shown to interact with either the U or the M state whereas lon is shown to degrade either the U or M state. The aggregates are shown as inclusion bodies which is how proteins aggregate in *E. coli*. 
$\text{U} = \text{Synthesized chain}$

$\text{U} = \text{Unfolded}$

$\text{N}_{\text{mono}} = \text{Folded monomer/tetramer}$

$\text{M} = \text{Misfolded}$

$\text{D} = \text{Degraded}$

$\text{A} = \text{Aggregated}$

$k_d = \text{rate constant for the reaction}$

$N_{\text{tet}}$
3.10. Conclusions

Our experiments provide insight into the species populated by MetK upon synthesis in *E. coli*. Consistent with hypotheses generated by FoldEco, MetK is a slow folding protein that tends to populate the off-pathway misfolded state with a larger than average (with respect to the average properties of the 4000 test proteins used in the computational study) (Powers et al., 2012) propensity. This misfolded population is hypothetically a target for the Lon protease, and our results support this hypothesis in that we observed a greater diminution in the fraction aggregated than the total protein at similar concentrations of MetK as Lon concentration increased. The propensity of MetK to aggregate even at larger than stoichiometric concentrations of both GroEL and DnaK indicates that MetK has an intrinsic aggregation rate faster than a GroEL facilitated (as without GroEL the rate is not readily observable in reasonable time scales) folding rate. Furthermore, we have inferred a chaperone cooperation wherein DnaK is able to prevent the aggregation of MetK and subsequently co-ordinate with GroEL to ensure the successful folding of a large population of the protein.

While the folding of MetK is stringently dependent on GroEL in *E. coli*, it is interesting to note that organisms lacking GroEL such as *Mycoplasma synoviae* have MetK homologs (Georgescauld et al., 2014). The *M. synoviae* MetK shares approximately 47% percent sequence identity with the *E. coli* MetK. On imposing the similar residues on to the structure of the *E. coli* MetK, we observed that the similar residues were concentrated
at the interface of the MetK dimer (Figure 3.10). This provides a hypothesis that the assembly of MetK might be conserved evolutionarily, whereas the folding of the monomeric protein has co-evolved with the organismal proteostasis network (or lack thereof).

**Figure 3.10 MetK assembly might be independent of monomer folding.**

A. The image shows the dimer of *E. coli* MetK (PDB 1RG9) with one chain shown in spheres and the other as a cartoon. Comparing the sequences of homologous MetK, one from *E. coli* and another from an organism that lacks GroEL (*M. synoviae*), it is apparent that identical residues (coloured red) are clustered in the dimer interface whereas non-identical residues (coloured green) are distal to the dimer interface. This figure looks at the dimer interface through one of the monomers (blue cartoon).

B. This cartoon shows the MetK dimer interface from the side where the residues identical between *E. coli* and *M. synoviae* are easier to visualize. This conservation pattern can be used to hypothesize that MetK assembly might be evolutionary conserved but the folding of the monomer has co-evolved with the chaperone in the case of *E. coli*. 
Lastly, as seen in the section of how DnaKJE help out MetK under limiting concentrations of GroEL, it has not escaped our attention that in this scenario, DnaK is able to compensate for the action of GroEL at most concentrations of expressed MetK. There could be two ways DnaK could exert its effect, the first being serving as a holding point for freshly synthesized MetK and keeping it from aggregation. In this scenario, the synthesized MetK would partition simultaneously to both DnaK and GroEL but as the concentration of MetK exceeds the concentration of GroEL the remainder of the protein would partition to the excess DnaK and finally the protein population that escapes both chaperone systems will aggregate. In this scenario, DnaK would hold on to the MetK polypeptide and release it in a manner capable of being acted upon by GroELS. The release of MetK and the processing by GroEL (which is necessary for MetK folding) appears to have some stoichiometric balance at lower levels of MetK as the limiting GroEL is able to handle that protein load in the presence of high concentrations of DnaK. The second model invokes the possibility that DnaK sequesters other GroEL client proteins and thus increases the capacity of GroEL to accommodate larger concentrations of the synthesized MetK. However, this model may not hold as most of the cellular protein synthesis machinery is occupied for the synthesis of the chaperone components (induced by arabinose) or for the synthesis of MetK (under IPTG control). Thus most of the protein synthesized will likely be MetK and thus reduces the chances (essentially outcompetes) of any other protein substrate to be bound by DnaK.

3.11. Materials and methods

• Cloning, expression and cell culture:
The MetK gene on a subcloned into a pET29b+ vector (Kanamycin\textsuperscript{r}, IPTG inducible) was a gift from the Hartl laboratory (Max Planck Institute, Martinsried Germany). DNAKJE and lon genes were subcloned into the pBAD33 vector (Chloramphenicol\textsuperscript{r}, arabinose inducible) and were the same as the constructs used in Cho et al. (FoldEco 2015). Similarly the pGro7 was also the same construct used in Cho et al. with GroELS synthesis driven by arabinose. The chaperone plasmids were a gift from the Kelly laboratory (The Scripps Research Institute, La Jolla USA).

BL21(DE3) E. coli cells (Invitrogen) were transformed with either the MetK or the MetK and one of any three of the chaperone plasmids mentioned above. The cells were plated on Luria Bertani (LB) agar containing Kanamycin and Chloramphenicol at a final concentration of 50 µg/ml and 34 µg/ml respectively. Cultures for the induction and partition experiments were started from overnight cultures initiated from a single colony of the transformed E. coli. For induction of proteins, the transformed cells were grown in LB medium at 37 °C until they reached an OD\textsubscript{600} of 0.3 at which point the chaperone components were induced with the desired concentration of arabinose (range between 0.002% and 0.2% w/v final concentration). Cells were then allowed to grow till an OD\textsubscript{600} of 0.6 (after ~30 minutes) at which point MetK was induced using a desired concentration of IPTG (range between 1 and 1500 µM final concentration). Cells were harvested by centrifugation at 4000 x g for 2 minutes after 1.5 hours of MetK induction after normalizing OD\textsubscript{600} to 1 across samples.

- Partition experiments and concentration measurement
Harvested cells were resuspended in 20mM Tris at pH 7.4 and were lysed using a sonicator. Sonicated cells were then centrifuged at 18,000 x g for 30 minutes to separate the soluble and insoluble components. The total (unlysed), the supernatant and the insoluble component (after resuspension in equal volume buffer prior to cell lysis) were then run on 12% SDS PAGE gels. The gels were stained with coomassie brilliant blue G-250 dye and scanned on the LICOR gel documentation system after destaining. Quantification of the gel band intensity was carried out using the Odessy 4.0 software associated with the LICOR. To measure concentration of protein, known concentrations of purified protein (MetK, GroEL, DnaK or lon) were loaded on the same gels as the unknown sample. Comparison of the unknown to the standard curve generated from the purified protein yielded the concentration of protein (µG) present in the sample loaded on the gel. By accounting for the number of cells used for the experiment we estimated the cytosolic volume (Volkmer and Heinemann, 2011) and used that to measure the concentration (µM) of the protein inside the cell. Purified chaperone components GroEL, DnaK and Lon were kindly provided by Ivan Budyak, Joseph Tilitsky and Rilee Zeinert respectively.

• Purification of MetK

Cell growth and protein induction.

MetK (pET 29b) was transformed into *E. coli* BL21(DE3) cells and a single colony was used to start an overnight culture. 1.5% of the overnight culture was used to start a fresh 1 L LB + Kanamycin (final 50µg/ml) culture. The 1 L culture was grown to OD 600 of 0.6 at which point it was induced with a final IPTG concentration of 0.2mM for 4-5 hours.
All growth and induction was done at 37 °C. Cells were harvested by centrifugation at 4000 x g and resuspended in 20mM HEPES pH7.4 and flash frozen with liquid N2 until further use.

• Cell lysis.
Frozen cells were thawed under running cold water until completely thawed and then were lysed using the microfluidizer. After being microfluidized, the lysed cells were centrifuged at 20,000 rpm for 30 minutes and the insoluble fraction was discarded. The soluble fraction was processed further.

• Ammonium Sulphate precipitation.
The first step was to perform an ammonium sulphate precipitation of the protein and or contaminants. Solid ammonium sulphate was gradually added to the supernatant after cell lysis in 10% increments with slow stirring in 10 minute intervals. At 60% ammonium sulphate, the protein (soluble MetK) was mostly in the precipitate.

The 60% ammonium sulphate precipitate was resuspended in 20mM HEPES at pH 7.4 and dialyzed against 20mM HEPES at pH 7.4 to remove as much ammonium Sulphate as possible before further treatment.

• Hydroxyapatite column chromatography.
The dialysate from the ammonium Sulphate precipitation was loaded a pre-equilibrated hydroxyapatite column (pre equilibrated with 10mM Sodium Phosphate monobasic-dibasic buffer at pH 6.8. MetK binds both ATP and L-met and hydroxyapatite is a chemical mimic of ATP. The protein was loaded on the column and washed with 20mM sodium phosphate (monobasic - dibasic component buffer at pH 6.8) and eluted with a gradient ranging from 20mM to 300mM of the same buffer.

MetK eluted in the wash, which was at low concentration of sodium phosphate and the impurities eluted at higher concentrations of the elution buffer. The sample was then dialyzed against 20mM Tris pH 7.4 and flash frozen and stored at -80 °C till further use. The concentration was determined by sending out pure protein for amino acid analysis at the Texas A and M university protein chemistry laboratory.

- Equilibrium denaturation experiments

MetK at fixed concentration was diluted into a range of guanidine hydrochloride concentration and incubated over 12 hours at 37 °C to ensure unfolding. The unfolding of MetK as a function of guanidine hydrochloride was followed by collecting tryptophan emission spectra between 320 and 380 nm after excitation at 295 nm. Plotting the maximum wavelength as a function of guanidine chloride calculation yielded a three state equilibrium denaturation curve. Repeating the experiments over a range of MetK concentrations revealed the first transition sensitive to guanidine concentration and the second transition somewhat insensitive to MetK concentration. Fitting the second
transition to a two state equation yielded an apparent thermodynamic stability for the MetK monomer (Pace 1986).

Data for MetK concentration, fraction aggregated and concentration of chaperones at various IPTG and arabinose concentrations is listed in the appendix section of the document.
CHAPTER 4
MUTUALLY EXCLUSIVE FOLDING – SENSORS FOR PROTEIN STABILITY

This chapter investigates how we can utilize a fusion protein system that relies on ligand binding to test the thermodynamic stability of a protein of choice inside the cell. This study attempts to bring together the concept of how ligand binding (explored in Chapter 2) can be potentially harnessed to alter the stability of a protein connected distally to the ligand binding protein (the mutually exclusive folding model – discussed below). The work in this chapter was performed in collaboration with Beena Krishnan and Alexandra Walls.

4.1. The mutually exclusive folding principle and the suitable choice of host protein

In order to determine the thermodynamic stability of a protein inside the cell we have attempted to construct a mutually exclusive folding (MEF) system consisting of a host and test protein of our choice (see Figure 4.1) designed originally by Stewart Loh (Cutler and Loh, 2007; Cutler et al., 2009; Radley et al., 2003). The MEF system works on the basis that the coupling of the two proteins is architecturally incompatible. The differences in distance between the site of insertion on the host and the N to C distance of the test protein result in folding tug-of-war where both domains attempt to reach their individual native states. However a native state can only be achieved by one domain at the energetic cost of unfolding the other domain. Thus the ‘winner’ of this folding tug-of-war would be the protein with larger equilibrium stability (Figure 4.1A) (Cutler and Loh, 2007).
Furthermore such MEF systems can act as molecular switches where a change in conditions can alter the folding equilibrium, thus selectively unfolding one of the two proteins (Radley et al., 2003). The choice of change in condition we desire to exploit is the stability imparted by the binding of a high-affinity ligand. Thus in our MEF scenario, the stability of the test protein is poised between the stability of the apo-host protein and the holo-host protein. Hypothetically, titration of the ligand would shift the population of the host protein from apo to the holo state, thus increasing the stability of the host protein relative to the test protein causing the test protein to unfold as a function of the concentration of the ligand added (Figure 4.4 pages 157-158).

Figure 4.1 General design for a MEF system

A. The cartoon shows the simple design for a mutually exclusive folding construct with two proteins, a host and a test. N-C termini of test (red arrows) are inserted into the host at a surface loop causing architectural mismatch and setting up a folding tug-of-war.

B. One of our MEF constructs has DHFR as the host protein and Ubiquitin as the test protein. Ubiquitin with its N-C distance of 37 Å is inserted into the DHFR surface loop (between 106 – 107) with a distance of 6.7 Å. With this distance incompatibility, DHFR and ubiquitin are now competing with each other to fold and the winner in this case will be ubiquitin as it has the larger thermodynamic stability.
4.2 Modified DHFR as a host protein

For the choice of host protein we have selected DHFR modified at the same surface loop (between residues 106-107). This loop has also been manipulated in murine DHFR in a fragment based recombination assay (Remy et al., 2007). It has been previously shown that the DHFR fragmented at this loop can complement and can bind ligand (the two fragments re-constitute the structure) (Remy et al., 2007). DHFR was a feasible choice for the following reasons. First, the protein, while aggregation prone in vivo can be easily refolded from inclusion bodies suggesting that if we can control the folding/aggregation competition, we can obtain higher amounts of soluble protein inside the cell. Second, we have characterized the folding stability of this protein in vitro at 37 °C (Chapter 2), and the low stability value of 0.5 kcal/mol at 37 °C is desirable as a stability value to be at the lower extreme of our MEF set up (the apo-test protein). Third, we know that the modified DHFR is capable of binding the high affinity ligand trimethoprim (TMP) with nanomolar affinity. A dissociation equilibrium constant in the range of $10^{-9}$ M imparts a stability enhancement of approximately 12 kcal/mol (we can calculate this by using the formula $\Delta G = -RT\ln K_d$, where R is the gas constant, T the absolute temperature and $K_d$ is the equilibrium dissociation constant). Lastly, the ligand trimethoprim is highly specific for DHFR and thus unlikely to interact with other proteins inside the cell, thus making the system specifically tunable inside the complex environment of the cell. These four properties allow us to consider DHFR as suitable host protein. Figure 4.1b shows an example of a MEF construct where DHFR is the host protein and Ubiquitin is the test protein inserted into the surface loop by its N and C termini. The distance incompatibility
between the two proteins at the point of insertion sets up a folding tug-of-war, which is the principle of the MEF design.

Furthermore, we tested the thermodynamic stability of the modified DHFR to determine how much impact modifying the surface loop would have on the stability of this protein. While we know that the protein is destabilized by approximately 3 kcal/mol at 37 °C compared to wild type, this is only for one type of modification. To test the impact of altering the loop length, we tested the stability of wildtype DHFR and compared to it the stability of the gly-gly inserted variant and a variant with a (gly-gly-gly-ser)x2 (GGGS2) variant inserted in the same position. The idea behind expanding the loop was to ask whether insertion of larger fragments would increasingly impact the already perturbed stability of DHFR (Figure 4.2A, B, C). The increase in loop length was hypothesized to minimally perturb stability, as it had no inherent structure, yet had an entropic component that could affect stability. This would not be the case when a folded domain would be inserted into the surface loop of DHFR (which would then be offset by ligand binding DHFR – as designed in the MEF system). We tested the thermodynamic stability of wildtype, GG and GGGS2 at 25 °C using urea denaturation experiments. The stability of wildtype DHFR was calculated to be about 6.4 kcal/mol which is in good agreement with the published value of 6.15 kcal/mol (Ionescu et al., 2000). The stability of the GG and the GGGS2 variants was estimated to be 3.4 and 3.1 kcal/mol respectively suggesting an appreciable reduction in stability compared to wildtype, but not when compared to each other. These values of stabilities agree with our hypothesis that while expansion of the loop with gly-gly perturbs stability significantly, further increasing the
loop length does not further perturb stability drastically. These experiments suggest a certain baseline to which DHFR would be destabilized as the host protein in the MEF constructs.

**Figure 4.2 Equilibrium stability of DHFR constructs with modified surface loops**

A. Thermodynamic stability of wildtype DHFR at 25 °C is estimated to 6.4 kcal/mol.
B. The insertion of gly-gly residues at the surface loop reduces stability to approximately 3.4 kcal/mol.

C. Increasing the loop length with a (gly-gly-gly-ser)x2 causes no drastic alteration of stability compared to the original loss of stability. In all panels the family of curves shows the fluorescence spectra recorded between 320 and 380 nm after excitation of tryptophan residues at 295 nm. The fluorescence at 344 nm is shown next to the family of curves showing an equilibrium two state denaturation.
C.

![Graphs showing fluorescence intensity at 344 nm and intensity at 344 nm x 10^4 as functions of wavelength and urea concentration.](image)
4.2. The choice of reporter on the folded/unfolded state of the host or test protein.

In collaboration with Beena Krishnan and Alexandra Walls

In order to determine the status of the host or the test protein in our MEF system, we need a signal which reports on the folded or the unfolded status of the either the host or the test protein. We seek to build upon the existing fluorescence based methods to detect the folded state of a protein. Previous attempts to measure the folded populations inside cells have involved the use of the fluorophore Fluorescein Arsenical Hairpin binder (FlAsH) binding to either a contiguous or a split tetra-cysteine motif present within the protein of interest (Ignatova and Gierasch, 2004; Ignatova et al., 2007; Krishnan and Gierasch, 2008). Here we design a third generation of tetra-cysteine motif, where the cysteine residues are inaccessible to the FlAsH fluorophore when the test protein is folded, but upon being unfolded, are solvent exposed, and thus capable of binding the FlAsH fluorophore. The quantum yield of FlAsH dramatically increases upon binding the tetra-cysteine motif and the signal can easily be differentiated from unbound FlAsH.

We chose to work with *E. coli* DHFR (our host protein), so that the ligand binding and the reporter modules could be housed within one protein eventually. The other protein we chose was human ubiquitin. We chose ubiquitin as one of our test proteins as well (see below). In order to not perturb secondary structural elements of the protein, we engineered the tetra-cysteine motif into loops of the proteins. In, DHFR loop 87-92 was modified from DVPEIM to CCPECC. Similarly, in ubiquitin loop EVEPSD was
modified to CCEPCC (Figure 4.3A, C). The proteins modified to contain the tetra-cysteine (tetra-cys) motif are now dubbed TC-Ub (for ubiquitin) and TC-DHFR (for DHFR) (Figure 4.3A, B). We hypothesize that the position of the tetra-cysteine motif precludes the binding of FlAsH (low fluorescence intensity), and that the fluorophore only binds upon the unfolding of the protein (high fluorescence intensity), thus reporting on the unfolded populations of the proteins.

To test the stability of TC-Ub and TC-DHFR, 5µM protein was incubated with increasing concentration of Guanidine hydrochloride (for TC-Ub) and Urea (for TC-DHFR). The protein was allowed to incubate for 12 hours at 25 °C. FlAsH was added to a final concentration of 0.5µM in the presence of the reductant ethane dithiol (EDT) to 50µM. The high concentration of the reductant allowed us to assume equilibrium between free FlAsH (conjugated with EDT) and bound FlAsH (bound to protein). The EDT also prevents non-specific binding of FlAsH to other proteins – a scenario likely to occur inside the cell. The protein was allowed to incubate with the dye for 30 minutes prior to fluorescence measurements. Fluorescence spectra were collected from 522 to 560 nm after excitation at 508 nm. We found that the stability of TC-DHFR was approximately 2.5 kcal/mol which is a ΔΔG of 3.5 kcal/mol compared to wild type at the same temperature (Figure 4.2). Ubiquitin stability was measured to be 4.7 kcal/mol, a value which is within the stability range of apo and holo DHFR (Figure 4.3C, D).

Our next objective was to test whether the thermodynamic stability measured by change in FlAsH fluorescence is comparable to the stability measured by other methods. This is
particularly important for the test protein as any modifications can further perturb its stability. We suspected based on previous studies that FlAsH binding does not heavily perturb the folding equilibrium (Krishnan and Gierasch, 2008). Also, as the dye is added in sub-stoichiometric amounts in the presence of a binding competitor (1,2-ethanedithiol (EDT)), we assume that FlAsH acts merely as a reporter of the unfolded state. We tested this assumption by determining the stability of Ubiquitin in a FlAsH independent manner. As ubiquitin has no native tryptophan residues we decided to use circular dichroism to determine the stability of our tetra-cysteine ubiquitin construct. These experiments were conducted in conjunction with Alexandra Walls, a former undergraduate in the laboratory. To test the stability of TC-Ubiquitin, 1.5µM protein was incubated with increasing concentration of Urea. The protein was allowed to incubate for 12 hours at 25 °C. CD signal was collected from 250 to 210 nm and ellipticity at 222 nm was used to calculate the stability of TC-Ub.

From our measurements monitoring the change in ellipticity as an output we determined the stability of the ubiquitin construct to be 5.3 Kcal/mol. By comparing the stability from FlAsH measurements and CD, the two values are comparable and support our hypothesis that FlAsH binding does not significantly perturb stability (Figure 4.3E).
4.3 Utility of FlAsH binding for stability measurements

A. Tetra cys DHFR modified at a surface loop (between residues 88-92, see text).

B. Thermodynamic stability of TC-DHFR calculated by measuring increase in FlAsH fluorescence as a function of urea concentration. The fluorescence intensity rises with increasing urea but drops at 2.5M. Stability calculated to about 2.5 kcal/mol by fitting to the two state equation.

C. TC-Ubiquitin modified at a surface loop (between residues 17-22, see text).

D. Thermodynamic stability of TC-Ubiquitin calculated by measuring increase in FlAsH fluorescence as a function of GdnCl concentration. Similar to TC-DHFR fluorescence rises with increase in GdnCl, but drops after 4.5 M similar to TC-DHFR. Stability estimated to 4.7 kcal/mol.

E. Estimation of TC-Ubiquitin by circular dichroism yields a value of 5.3 kcal/mol indicating a mild perturbation of the stability of TC-Ubiquitin when measured with FlAsH fluorescence.
A. TC-DHFR

B. 

C. 

Wavelength (nm)  

Intensity (cps x 10^4)  

Intensity at 535 nm x 10^4 vs Urea (M)
TC-Ubiquitin

D.

![Graph showing wavelength (nm) vs. intensity (cps x 10^4)]

![Graph showing GdnCl (M) vs. intensity at 535 nm x 10^4]
E.

![Graph showing MRE (deg.cm²/dmole) against Urea (M). The graph has a range from 0 to 3000 on the y-axis and 0 to 8 on the x-axis. The data points are plotted and connected by a smooth curve.]
4.3. The choice of test proteins

One impediment of the MEF construct is the choice of test proteins that can be used. There are three major limitations when choosing the test protein for the MEF construct, the first being the stability of the test protein. If the stability of the test protein exceeds the stability of the ligand bound host protein, there will not be a sufficient energetic differential to unfold the test protein. Second is the size of the test protein. A multimeric or a large sized protein which physically prohibits the re-union of the split host protein will keep the test protein from refolding and binding the ligand, this will not allow the MEF system to function as originally designed. The third limitation is the relative kinetics of folding and unfolding between the host and test proteins. If the kinetics are offset by a large difference, the unfolding of the test protein at the expense of folding the host protein may not occur on a measurable timescale regardless of the differential between the thermodynamic stabilities (Peng and Li, 2009).

All three of these limitations can be theoretically overcome by either, changing the host protein to accommodate a larger stability difference between holo and apo forms of the host protein or by varying the size of the host protein to better accommodate the test protein. The choice of DHFR manages to address the first limitation somewhat successfully as the stability of most proteins (with measured stability) lies between 4 and 12 kcal/mol which is within the stability range of apo (~0.5 kcal/mol) and holo (TMP
bound ~12 kcal/mol by calculation of stability imparted by ligand binding) DHFR. The kinetics of folding/unfolding will be intrinsic to the protein of choice and will not be dissociable from the protein(s) used for the host and test. For our preliminary work, we have chosen to work with two test proteins which are single domain and whose thermodynamic stability lies between 0.5 and 12 kcal/mol. Furthermore, the N terminus to C terminus distance for these proteins is significantly larger than the point of insertion into DHFR, in agreement with the rationale for the MEF construct.

The first choice of test protein was TC-ubiquitin. With the stability of this construct known (see section 4.2) and the N to C terminus distance of ~ 37 Å (compared to 6.5 at the surface loop of DHFR, see Figure 4.1B), it was a viable candidate for a test protein (Vijay-Kumar et al., 1987). Furthermore, with the tetra-cys motif reporting on the unfolded state of ubiquitin we now have a readout on how the folding of DHFR impacts the folding of ubiquitin by following FlAsH fluorescence. The second test protein we used was superfolder GFP (sfGFP), with a reported thermodynamic stability of 9.1 kcal/mol and an N to C terminus distance of approximately 17 Å (Pedelacq et al., 2006). Additionally, this test protein has the intrinsic advantage of being a self reporter as only folded superfolder GFP produces fluorescence. Both proteins were cloned into the surface loop of DHFR without any flanking linkers into pET16b vectors where the expression of these proteins was driven from an IPTG inducible T7 promoter. As a schematic we have depicted how the thermodynamic tug-of-war would play out inside the cell using the DHFR-TC-Ubiquitin MEF construct. In the first scenario without any ligand, a significantly larger fraction of the ubiquitin would be folded compared to
DHFR, however, upon titration of the ligand, we would see a shift in these population to now favor the folded form of DHFR, thus obligatorily unfolding ubiquitin exposing the tetra-cys motif which will be detected by FlAsH fluorescence (Figure 4.4).

**Figure 4.4 Schematic of the MEF system deployed inside cells**

The cartoon shows a schematic of how the DHFR-TC-Ubiquitin MEF system can work *in vivo*. In the first scenario the more stable test protein (TC-Ubiquitin) remains folded in the absence of any DHFR stabilizing ligand. This sequesters the FlAsH binding site and would show a low level of fluorescence. Upon introduction of the ligand trimethoprim (pink), the MEF tug-of-war would now shift such that larger populations of DHFR are folded, thus forcing the unfolding of the coupled TC-Ubiquitin exposing the FlAsH binding site which would increase the fluorescence signal.
4.4. Preliminary results

In our *in vitro* experiments with purified DHFR-TC-Ubiquitin (see methods), when the DHFR-TC-Ubiquitin construct was titrated with increasing concentrations of trimethoprim, 5µM protein was incubated with increasing concentration of trimethoprim. The protein was allowed to incubate for 12 hours at 25 °C. FlAsH was added to a final concentration of 0.5µM and EDT to 50µM. The protein was allowed to incubate with the dye for 30 minutes prior to fluorescence measurements. Fluorescence spectra were collected from 522 to 560 nm after excitation at 508 nm. According to our hypothesis, the increase in the concentration of the ligand should partition increasing concentrations of DHFR to the folded state thereby causing increasing the concentration of unfolded TC-ubiquitin that will be reported by an increase in FlAsH fluorescence. Indeed, we observed an increase in FlAsH fluorescence with increasing TMP concentrations. This supports our model that ligand binding drives the equilibrium towards DHFR folding and ubiquitin unfolding (see Figure 4.5 A).

For our preliminary *in vivo* experiment we transformed DHFR-sfGFP into *E. coli* and streaked them onto minimal media plates containing increasing concentrations of Trimethoprim. An increase in the TMP concentration should drive the folding of DHFR and thus necessarily the unfolding of the coupled GFP. The unfolding of GFP should cause a decrease in GFP fluorescence. In good agreement with our MEF model, we
observed a loss in fluorescence upon increasing the concentration of TMP. This result showed promise to continue our work with the MEF system *in vivo* where the protein folding equilibrium can be tuned by altering the concentration of the ligand (Figure 4.5 B).

**Figure 4.5 In vitro and in vivo demonstrations of MEF constructs**

A. MEF system works *in vitro*. As hypothesized, increasing TMP concentrations causes a rise in FlAsH fluorescence indicating the unfolding of TC-Ubiquitin as a consequence of DHFR folding. The molar ratio of DHFR-TMP indicates a 1 to 1 binding of the ligand to DHFR. As a control experiment, the DHFR-Ubiquitin construct (without any tetra-cys motif) does not show any change in fluorescence upon increase in TMP levels thus showing the usefulness of the tetra-cys reporter.

B. *E. coli* cells transformed with DHFR-sfGFP construct were grown overnight on minimal media containing increasing concentrations of TMP. Two main observations are that the cells are able to grow on large amounts of TMP suggesting that the DHFR provides some resistance to the high levels of TMP. Second, the GFP fluorescence is reduced at higher TMP concentrations indicating the presence of folded DHFR and thus unfolded sfGFP, which is unable to fold and fluoresce.
A.

Test protein - TC-Ubiquitin

Relative Fluorescence

Molar ratio [DHFR]/[TMP]

0 1 2 3 4 5 6

0 1 2 3 4 5 6 7 8

TC-Ubiquitin

Ubiquitin
B.

*In vivo*

Test protein - sf-GFP

<table>
<thead>
<tr>
<th>[TMP]</th>
<th>DHFR</th>
<th>DHFR-GFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5 μg/ml</td>
<td>![DHFR Image]</td>
<td>![DHFR-GFP Image]</td>
</tr>
<tr>
<td>1 μg/ml</td>
<td>![DHFR Image]</td>
<td>![DHFR-GFP Image]</td>
</tr>
<tr>
<td>10 μg/ml</td>
<td>![DHFR Image]</td>
<td>![DHFR-GFP Image]</td>
</tr>
</tbody>
</table>
4.5. Pitfalls in the MEF design

While the MEF system appears to have sound theoretical basis and has been shown to work in preliminary experiments, there are certain challenges one has to consider when working with this system. First, is the question of what it is that we are measuring? While unfolding the test protein as a consequence of folding the host protein can be measured as a function of the ligand, it is not truly the test protein that is being tested. The protein being tested is now part of the MEF system that is a single chain system and not two separate proteins. This introduces a penalty on the thermodynamic stability of the test protein inserted into the system. This thermodynamic penalty is further examined by Cutler and Lon (Cutler and Loh, 2007) and it is essentially the penalty for physically coupling the two proteins exerted on the system. This changes the measured stability of the test protein ($\Delta G_{\text{test}}$) to a relative stability ($\Delta \Delta G_{\text{test}}$) where $\Delta \Delta G_{\text{test}}$ is $\Delta G_{\text{test}} - \Delta G_{x}$. In this scenario, $\Delta G_{\text{test}}$ is the stability of the free test protein when not associated with the MEF system and $\Delta G_{x}$ is the thermodynamic penalty for insertion of the test protein into the MEF system. The major disadvantage of this is that the $\Delta G_{x}$ will have to be measured for each construct made with the MEF system (due to the inherent differences in stability amongst proteins and the varying levels of architectural incompatibilities introduced by the varying N to C terminus lengths of the test proteins). Thus, only when the measurements are made to obtain $\Delta G_{x}$ can we deduce the stability of the test protein, otherwise it will only be a relative measurement. This $\Delta G_{x}$ can also vary in vivo where
environmental factors (see Chapter 1) can influence the folding of the MEF construct. However even without the measurement of $\Delta G_x$, the strategy would be useful to determine the impact of mutations on the stability the test protein relative to wildtype as the answer will be a ratio and not true numbers of stability.

Second, mutually exclusive folding constructs are highly aggregation prone in vivo. This is not surprising as the design of the construct lends to a large fraction of either the test or the host protein being unfolded. In the absence of any degradation, we have found that MEF constructs expressed in E. coli are highly aggregation prone. We have seen that the presence a TMP solubilizes a large portion of the dDHFR construct from chapter 2. However, as we noted in the mathematical derivation (see below), the presence of any amount of aggregated protein and the prevention of aggregation is intimately tied to the intrinsic ratio of the folding and aggregation rate constants of the protein. Thus any aggregation of the construct will not allow us to effectively separate the $K_f$ (folding equilibrium constant, a ratio of $k_f$ to $k_u$ – the folding and unfolding rate constants) and $k_{agg}$, the aggregation rate constant.

Recall from chapter two that the prevention of aggregation by the ligand is dependent on the concentration of the ligand to a certain extent and the intrinsic folding vs. aggregation propensities of the protein (Equation 1). Furthermore, the impact of the ligand at very high concentrations is limited by the ratio of the folding equilibrium constant to the aggregation rate constant (Equation 2).
At infinite ligand concentration equation 1 transforms such that

\[ f[A] = \frac{k_{\text{agg}} (k_u + k_a[L])}{k_f k_a[L] + k_{\text{agg}} (k_u + k_a[L])} = \frac{1}{1 + \frac{k_f k_a[L]}{k_{\text{agg}} (k_u + k_a[L])}} \]  

(1)

One potential solution to this is to purify the protein and exogenously introduce it into the cell via a method like electroporation. Although this too does not ensure that the cellular environment will not change the status quo of the MEF construct and cause it to aggregate inside the cell. A similar issue exists with using the MEF construct in cell lysates.

A potential solution to such problems is to design a muturally “inclusive” folding system. Such a system has been reported in the joint efforts of the Bardwell – Radford groups (Foit et al., 2009). The method they deploy is composed of a β-lactamase fused to the protein of interest with sufficiently long linkers such that there is not architectural incompatibility and the folding of the two proteins is independent. The system depends on periplasmic degradation machinery of _E. coli_ to rapidly degrade marginally stable test proteins, and thus the associated β-lactamase. This decrease in β-lactamase levels is then reflected in the ability of _E. coli_ (transformed with the β-lactamase construct) to grow on β-lactam antibiotics. If mutation in the test protein causes destabilization, that construct
will be more susceptible to degradation and thus will not offer any protection against the antibiotic. This system allows for the testing of impact of mutations on the stability of the test protein *in vivo* but does not allow comparison between two entirely different test proteins. Another way to approach the problem is by using NMR. Only in the last year advances in measuring protein stability inside the cell using NMR have gained traction. By either using fluorinated amino acids or by electroporating isotopically labeled protein into cells have the effects of the cellular environment been tested on the stability of proteins (Danielsson et al., 2015; Smith et al., 2016). Not unsurprisingly, protein stability is affected although not drastically depending on the protein and the cellular environment surrounding the protein, i.e. stability is context dependent. This is in good agreement with the few reports on protein thermodynamic stability inside cells measure in the past decade (Ebbinghaus et al., 2010; Guo et al., 2012; Guzman and Gruebele, 2014) (Ghaemmaghami and Oas, 2001; Ignatova and Gierasch, 2004).

From Chapter 3 we learned the effect of various chaperones and the degradation enzyme lon on the fate of one test protein. Our group has also reported similar effects for several other test proteins. These observations highlight the importance of thinking about protein folding inside the cell as a kinetic competition for the polypeptide chain. Thermodynamic analyses and methods such as the MEF system will allow us to examine the impact of mutations on proteins, but this impact will also be manifested in how the protein partitions into folded or aggregated states or to the degree it is degraded. Thus an approach that considers all potential fates of a protein inside the cell will ultimately yield more information about how this process works inside the cell.
4.6. Materials and Methods

• Cloning, protein expression and purification.
Cysteine free AS-DHFR was a gift from the Matthews lab (UMass Medical School, Worcester MA). The gene was subcloned into the pET16b vector thus getting the protein expression under the control of the T7-IPTG system. Insertions into the DHFR sequence were performed by the introduction of NdeI and Xho I restriction enzyme sites into DHFR by the standard stratagene quickchange protocol allowing us to ligate desired fragments between residues 106 and 107 of the translated protein.

Wildtype DHFR was purified using the following protocol

• Cell growth and protein induction
DHFR plasmid (pET 16b) was transformed into E. coli BL21(DE3) cells and a single colony was used to start an overnight culture. 1.5% of the overnight culture was used to start a fresh 1 L LB + Ampicillin (final 100ug/ml) culture. The 1 L culture was grown to OD 600 of 0.6 at which point it was induced with a final IPTG concentration of 1mM for 4-5 hours. All growth and induction was done at 37 °C. Cells were harvested by centrifugation at 4000 x g and resuspended in 20mM HEPES pH7.4 and flash frozen with liquid N2 until further use.
• Cell lysis:

Frozen cells were thawed under running cold water until completely thawed and then were lysed using the microfluidizer. After being microfluidized, the lysed cells were centrifuged at 20,000 rpm for 30 minutes and the insoluble fraction was discarded. The soluble fraction was processed further.

• Ammonium Sulphate precipitation

The first step was to perform an ammonium sulphate precipitation of the protein and or contaminants. Solid ammonium sulphate was gradually added to the supernatant after cell lysis in 10% increments with slow stirring in 10 minute intervals. At 60% ammonium sulphate, the protein (soluble DHFR) was mostly in the supernatant. The 60% ammonium sulphate supernatant was resuspended in 20mM HEPES at pH 7.4 and dialyzed against 20mM HEPES at pH 7.4 to remove as much ammonium Sulphate as possible before further treatment.

• DEAE column chromatography

The dialysate from the ammonium Sulphate precipitation was loaded a pre-equilibrated DEAE anion exchange column (pre equilibrated with 20mM HEPES at pH 7.4. The protein was loaded on the column and washed with 20mM HEPES at pH 7.4 and eluted with a gradient ranging from 20mM to 1000 mM NaCl made in the HEPES. The eluted sample was then dialyzed against 20mM Tris pH 7.4 and flash frozen and stored at -80 °C till further use.
All other DHFR related constructs were purified from inclusion bodies followed by the same DEAE column chromatography used for wildtype DHFR.

- Extraction from inclusion bodies

After cell lysis and centrifugation, the supernatant was discarded and the precipitate was washed with 0.1% Triton X-100 containing HEPES buffer at pH 7.4 to remove membrane fractions. A second wash step was subsequently performed to remove any triton from the first wash. The precipitate was then dissolved in 8 M urea made in 20 mM HEPES at pH 7.4 for two hours until most of it dissolved. The unfolded protein from the precipitate was then dialyzed against 20 mM HEPES pH 7.4 to remove the urea and initiate protein refolding. The dialysate was further purified using the column chromatography mentioned above.

- Equilibrium denaturation experiments:

For the DHFR variants, 3-5 µM protein was added to increasing concentrations of urea and allowed to incubate overnight at 25 °C for a minimum of 12 hours. Unfolding of DHFR was measured by tryptophan fluorescence after excitation at 295 nm followed by acquiring spectra between 320 and 380 nm. Fluorescence intensity at 344 nm was used to fit to the two state equation to determine equilibrium thermodynamic stability. Similarly, for the tetra cysteine variants, the denaturation was performed using the same procedure mentioned above with either increasing concentrations of urea or guanidine hydrochloride. FLAsH fluorescence was measured after excitation at 508 followed by acquiring spectra between 520 and 560 nm. Intensity at 535 nm was used to fit to the two
state equation to determine stability. Denaturation of TC-Ubiquitin was also followed using circular dichroism by measuring the intensity at 222 nm which reports on the alpha helical secondary structure.

• MEF experiments *in vitro* and *in vivo*

• *In vitro*
Purified DHFR-TCUb (5 µM) was incubated with increasing concentrations of TMP and allowed to incubate overnight at 25 °C. After a minimum of 10 hours of incubation, FlAsH and EDT were added at a final concentration of 0.5 and 50 µM respectively. Incubation with the fluorophore was allowed to occur for 30 minutes and FlAsH fluorescence was subsequently measured. The same experiment was conducted with DHFR-Ub (no tetra cys motif) to ensure that FlAsH fluorescence reports on the unfolding of TC-Ubiquitin with high fidelity.

• *In vivo*

DHFR-sfGFP constructs were transformed into BL21(DE3) cells and the cells were plated on minimal medium containing increasing concentrations of TMP. The plates were incubated at 37 °C overnight and were imaged using eh syngen gel documentation device.
CHAPTER 5
QUO VADIS

5.1. Summary and conclusions

The questions which have motivated my research have a shared thread of understanding how the in cell environment can alter the protein folding landscape. I have asked two specific questions regarding the \textit{in vivo} environment; the first was the impact of high affinity ligands on the folding fate of proteins and the second was the impact of the proteostasis network components on the folding fate of an obligatory chaperone dependent client protein. To understand protein folding inside the cell and the impact of the cellular environment on the protein, we have invoked a generalizable model of kinetic partitioning which states that the synthesized polypeptide chain is partitioned between correct folding, off-pathway intermediates which lead to aggregation, or degradation. All three fates – folded, misfolded (aggregated) and degraded are a result of intrinsic protein biophysical properties such as the rate constants governing each reaction. Upon this generalizable model of kinetic partitioning we have overlaid either the effect of a high affinity ligand – which then allows us to hypothesize the altered fate of the protein in the presence of such a ligand or we have overlaid the proteostasis network which similarly allows us hypothesize how chaperones and degradation machinery alter the folding fate of the protein under consideration.

Using a combination of experiments and mathematical modeling, we were able to arrive at several key conclusions regarding the impact of ligands and of the proteostasis components on the folding of proteins inside the cell. First, we were able to
mathematically deduce that the impact of a ligand was limited by the ratio of the intrinsic folding to aggregation and/or degradation rate constants. Our tests using Trimethoprim and 1-deoxygalactonojirimycin were able to experimentally determine this limit for a test protein (DHFR) and for a protein linked to late onset Fabry disease (α-galactosidase) respectively. Furthermore the results of this study also highlighted the importance of understanding the kinetic partitioning when it comes to pharmacological chaperone therapy along with the thermodynamic effects of the ligand. Finally this study now provides a feasible model as to why certain protein mutations – mutations that hamper the intrinsic folding of the protein are non-responsive to pharmacological chaperone therapy.

In the case of the impact of the proteostasis network on the folding of an obligatory chaperone client, this is to our knowledge the first study, which seeks to address why certain proteins have evolved to rely on chaperones to fold, while many others fold spontaneously. Initial models from simulation experiments regarding the impact of chaperone and degradation components were tested by experimentally varying the levels of protein (MetK) and proteostasis components (GroEL, DnaKJE and lon). By varying these components we shed light into the intrinsic propensity of our chaperone client to aggregate, furthermore, we were also able to show how DnaK can work in collaboration with GroEL to increase the pool of folded protein and finally how Lon specifically targets the misfolded state of MetK and acts to lower the concentration of aggregated protein.
5.2. Significance of the findings

While our experiments and modeling have primarily been in *E. coli* cells, our results especially about ligand binding have far reaching consequences. Importantly, the central concept of kinetic partitioning can be used to describe the distribution and equilibria governing protein folding and function inside the cell. One notable example of how kinetic partitioning can manifest in disease is the case of cystic fibrosis. The CFTR channel can be mutated in ways such that either the trafficking through the endoplasmic reticulum is altered, or there is impairment of the function of the channel upon reaching the membrane or the protein that reaches the membrane is rapidly degraded. All three scenarios cause a loss of function for the cell, which is detrimental to health. By suitably designing ligands such that either or all of the three pathways is/are targeted, we can now hypothesize how the kinetic partitioning of CFTR can be altered to improve cellular health. Secondly, several housekeeping and oncogenic proteins interact with cellular quality control machinery. These include but are not limited to clathrin, tau, p53, and STAT (Kasembeli et al., 2014; Trinidad et al., 2013; Yu et al., 2014). Mutations in these proteins often result in gain of toxicity or loss of function diseases. Thus our knowledge of how the proteostasis network deals with obligatory clients allows us to ask questions about eukaryotic systems as well. For example, how does a p53 mutation affect its interaction with the TRiC/CCT (the eukaryotic Hsp60) and how does that lead to disease? Similarly such questions can also be asked of organellar proteostasis.
5.3. Future directions:

The pharmacological chaperone work described in chapter 2 for the α-galactosidase enzyme was initially pioneered by Amicus therapeutics, the company currently taking the drug through phase 3 clinical trials as a small molecule therapy for Fabry disease. It was their work with the small molecule that initially showed a rise in enzyme concentration and activity upon treating cells expressing several mutant versions of α-galactosidase responsible for Fabry disease. This would be a great platform for a collaboration where the modeling approach we used in chapter 2 can be used to generate biophysical information regarding several Fabry disease causing mutants. The input for these models would theoretically be the enzyme activity – a surrogate for folded protein (or as in our case, the fraction remaining) and the concentration of DGJ used to achieve those enzyme activity levels. This would be a good database describing the folding/degradation competition for several versions of a protein inside eukaryotic cells. Much of this information can also be found in some of the work published by Amicus (Benjamin et al., 2009; Wu et al., 2011). For continuing the work to understand how folding/aggregation or degradation rate constants affect the impact of the ligand, future studies with known DHFR mutants can be pursued using similar experimental approaches. We have not characterized the folding or unfolding rate constants of the dDHFR construct used in our study, however we can test the ligand approach on other DHFR mutants (such as the one described by Cho et al. 2014) to further validate our system. Conversely, an attempt can be made to purify sufficient amounts of the mutant α-GAL (R301Q) to characterize its folding rate constant. Essentially, the ligand binding system can be further validated by either testing the *in vitro* parameters of different mutants of either of our model proteins.
and by performing similar *in vivo* ligand titration experiments on other mutants of these proteins.

Our experiments regarding the folding of MetK were performed in *E. coli* cells using plasmids to overexpress the components under investigation. This approach allows limited control of the levels of protein expressed and thus limits the amount of collectable experimental data that can be fitted by FoldEco. To improve our understanding of the impact of the proteostasis system we need a more precise way to control the concentration of the components without drastically altering the protein folding reaction going on inside the cell. Refolding experiments, where denatured protein is diluted into conditions favoring folding will not be appropriate as they will not recapitulate the involvement of the ribosome in synthesis and folding. To address this problem, we can use an *in vitro* transcription and translation system, such as the PURE system to synthesize the protein of choice while externally supplementing chaperones at the desired concentration to determine their impact on folding. One good test protein to follow up on would be MetK, as we can further probe how having GroEL influences the folding of this protein. Similarly, given the abundance of information we have on CRABP I mutants and their folding, we can use those as test proteins to better understand how varying the biophysical properties changes chaperone dependence to fold successfully. The *in vitro* nature of this experiment now allows us to examine multiple proteins in a more rigorous fashion albeit in *in vitro* conditions which lack certain cellular complexities – particularly molecular crowding and co-solutes.
One aspect where FoldEco is limited is in the number of proteins that can be tackled by the program. Currently, FoldEco is designed to handle how the flux of one protein varies during expression and how that protein relies on proteostasis network components based on its intrinsic folding/misfolding parameters. Experimentally, we can start with the introduction of two proteins and investigate how the proteostasis network would respond to two entities folding simultaneously. Once we learn more about this, we can start thinking about how we could express multiple proteins simultaneously and track their folding fate inside the cell. These experiments would allow us to expand FoldEco to incorporate the folding of multiple proteins inside the cell – which is what routinely occurs under physiological conditions.

A further extrapolation of the folding of MetK investigated in chapter 3 is how this work can be applied to proteins that heavily depend on the proteostasis machinery. This scenario is likely to occur when proteins accrue destabilizing mutations leaving them susceptible to aggregation. Previous work by the Hartl lab has shown how aggregation prone proteins sequester metastable proteins thus causing a secondary gain of toxicity phenotype. Furthermore these aggregates also tend to sequester chaperones that depletes the proteostasis machinery leaving the entire proteome vulnerable to misfolding and aggregation – or premature degradation, all scenarios detrimental for cellular health (Walther et al., 2015) (Olzscha et al., 2011). Using our understanding of how a GroEL substrate interacts with the chaperoning machinery inside the cell, we can now design mutant versions to rely less or more on the chaperone components and thus alter the burden on the proteostasis network. Alternatively, we can use proteins with diverse
biophysical parameters with different chaperone dependencies as stressors for the chaperone machinery. What this would allow us to do is ask the question, that if the chaperones are engaged with the stressor protein, which proteins are left unprotected in the proteome. Using quantitative mass spectrometry approaches such as isotopic labeling in SILAC experiments or using multiple tandem mass tags, we can determine which proteins are aggregation sensitive and to what extent when a stressor protein is introduced into the cell. We can then extend this analysis to introduce a disease variant and ask similar questions.

Lastly, the concept of kinetic partitioning can be applied to study protein folding in eukaryotic systems. One avenue is to pursue the folding of lysozomal proteins and systems where pharmacological chaperones are being investigated. However, another area of interest would be the endoplasmic reticulum (ER) (Braakman and Hebert, 2013). About a third of all proteins synthesized in eukaryotic cells are through the secretory pathway (the lysosome also belongs to this general pathway). Furthermore, several chaperones in the secretory pathway have been investigated extensively. Analogous to the cytosolic version, the chief chaperone in the ER is BiP, which belongs to the Hsp70 family of chaperones. Additionally, there are other modifications such as the addition of carbohydrates – which is also monitored by specific lectin binding chaperones. Finally, some proteins are secreted from the cell to the outside of the cell. Thus a simplified model, which involves kinetic partitioning, emerges from these factors. This models encompasses synthesis, translocation into the ER, binding BiP and the carbohydrate chaperones and secretion and the flux of the protein of interest through these points. Thus
using particular readouts such as concentrations of chaperones, the concentrations of proteins inside and outside the cell one can begin to build an understanding of how protein folding works in the ER.

5.4. Experimental Caveats and potential ways to address them

For chapters two and three, where we study the impact of ligands or chaperones; we use fraction aggregated as a readout for how the protein partitions between the folded and aggregated state. Though we call our readouts as aggregated or soluble, they are in fact placeholders for those properties. Our experimental methods allow for the separation of insoluble (which we call fraction aggregate) and soluble (which we call fraction folded) species after cell lysis. However, the insoluble and soluble are difficult to obtain cleanly and can be comprised of several different species. For example, the soluble fraction can be comprised of well-folded proteins, ligand bound proteins, protein molecules bound to chaperones, soluble higher order species, and unfolded species. It is difficult to separate the contribution of each of these species to the total population of the protein present in the soluble state, however two main follow up experiments can be performed to determine how the population of soluble protein is distributed. The first would be to perform enzyme activity assays. We would assume that only well folded proteins (that are enzymes like MetK) would show enzyme activity. This measurement would help determine how much of the soluble pool is folded to the native state. The second experiment would be to determine how much of the protein is distributed amongst the cellular chaperone pool. In order to do this, one would require immunoprecipitation of the chaperones and analysis of the contents bound to the chaperone. One has to be
cautious during this experiment because the first step would be to deplete the ATP in the soluble fraction so that the chaperone bound protein molecules remain trapped in the chaperone and second, one has to try to precipitate almost all the chaperone molecules to get a clear estimate of the population of proteins that are soluble and chaperone trapped.

Fraction aggregate measurements require the resuspension of proteins left insoluble after cell lysis. Several times, the insoluble pellet will not be entirely resuspended by the buffer, and the contents will not be homogeneously mixed. This is a potential source of error for the measurement of the aggregated species. One way to overcome this is to resuspend the insoluble fraction in a solution containing mild detergent or urea. This treatment will help dissolve the insoluble species better and thus get a better handle on the fraction aggregated. Additionally one can use a proteomic protein perpetually found in the insoluble fraction as an internal control to compare across samples. This will serve as a reference point for all samples and one can determine if the resuspension procedure drastically varied between samples.
APPENDIX
DERIVATION OF LIGAND BINDING AND DATA TABLES FOR CHAPTER THREE

DERIVATION OF LIGAND BINDING—provided by Evan Powers

The derivation below shows the detailed steps of the model describing how the effect of ligand binding (described in Chapter 2) can be explained by the ratio of the intrinsic folding to aggregation or degradation rate constants. This derivation was provided to us by Evan Powers.

The rate equations for this model are:

\[
\frac{d[U]}{dt} = \sigma - k_f [U] + k_u [N] - k_{agg} [U] - k_{deg} [U] \tag{1}
\]

\[
\frac{d[N]}{dt} = k_f [U] - k_u [N] - k_a [N][L] - k_{sec}[N] \tag{2}
\]

\[
\frac{d[A]}{dt} = k_{agg} [U] \tag{3}
\]

\[
\frac{d[N:L]}{dt} = k_a [N][L] - k_{sec}[N:L] \tag{4}
\]

\[
\frac{d[Deg]}{dt} = k_{deg} [U] \tag{5}
\]

\[
\frac{d[Sec]}{dt} = k_{sec} [N] + k_{sec}[N:L] \tag{6}
\]

where the bracketed symbols represent the concentrations of the corresponding states.
Case 1: No ligand – and monitoring the fate of the aggregated protein

When no ligand is present but there is aggregation, degradation, and/or secretion of protein, the terminal states are the aggregated, degraded, and secreted states. The amount of unfolded and native protein within the compartment in which protein is produced will eventually become negligible compared to the amount of protein in terminal states at long times. However, the unbound native state will be a substantial fraction of the total protein at earlier times. To understand the time-dependence of this system it is useful to break up the protein expression time course into two time periods. In the first time period, the amount of native protein is relatively small and $k_u[N]$ is small as well. As a consequence, $[U]$ fairly quickly reaches a “weak” pseudo-steady state in which the unfolding of native protein is negligible (the pseudo-steady state is weak in the sense that it evolves to another, more robust steady state on the time scale of the experiment). The value of $[U]$ at this pseudo-steady state can be determined from equation (1) by setting $k_u[N] = 0$:

$$[U]_{ss,1} = \frac{\sigma}{k_{agg} + k_{deg} + k_f}$$  \hspace{1cm} (7)

This value for $[U]$ can be inserted into equations (3) and (5), which can then be solved to yield expressions for the rate of accumulation of aggregated and degraded protein during the first period:

$$[A]_{t,1} = \frac{k_{agg} \sigma t}{k_{agg} + k_{deg} + k_f}$$  \hspace{1cm} (8)

$$[Deg]_{t,1} = \frac{k_{deg} \sigma t}{k_{agg} + k_{deg} + k_f}$$  \hspace{1cm} (9)
The rate of accumulation of native protein inside the compartment depends on the rates of folding and secretion and can be obtained by substituting the expression for \([U]_{\text{ss,1}}\) into equation (2) and solving the resulting differential equation to yield:

\[
[N]_{t,1} = \frac{k_f \sigma}{k_{agg} + k_{deg} + k_f} \left( \frac{1 - e^{-k_{sec}t}}{k_{sec}} \right)
\]

(10)

Inserting this expression for \([N]_{t,1}\) into equation (6) and solving the differential equation yields the time dependent concentration of secreted protein:

\[
[Sec]_{t,1} = \frac{k_f \sigma}{k_{agg} + k_{deg} + k_f} \left( t - \frac{1 - e^{-k_{sec}t}}{k_{sec}} \right)
\]

(11)

Note that when there is no secretion \((k_{sec} \to 0)\), equation (10) reduces to

\[
[N]_{t,1} = \frac{k_f \sigma t}{k_{agg} + k_{deg} + k_f}
\]

(12)

In the second time period, \([N]\) will have gotten large enough that unfolding can balance folding and a more robust pseudo-steady state is achieved. The concentrations of unfolded and native states in this second time period can be determined by setting equations (1) and (2) equal to 0 and solving:

\[
[U]_{\text{ss,2}} = \frac{\sigma (k_u + k_{sec})}{(k_{agg} + k_{deg})(k_{sec} + k_u) + k_f k_{sec}}
\]

(13)

\[
[N]_{\text{ss,2}} = \frac{\sigma k_f}{(k_{agg} + k_{deg})(k_{sec} + k_u) + k_f k_{sec}}
\]

(14)

The rates of accumulation in the terminal states can be obtained by inserting these values for \([U]_{\text{ss,2}}\) and \([N]_{\text{ss,2}}\) into equations (3), (5), and (6) and solving to give:

\[
[A]_{t,2} = \frac{\sigma (k_u + k_{sec}) k_{agg} t}{(k_{agg} + k_{deg})(k_{sec} + k_u) + k_f k_{sec}}
\]

(15)
\[ [\text{Deg}]_{t,2} = \frac{\sigma (k_u + k_{sec}) k_{deg} t}{(k_{agg} + k_{deg}) (k_{sec} + k_u) + k_f k_{sec}} \]  

\[ [\text{Sec}]_{t,2} = \frac{\sigma k_f k_{sec} t}{(k_{agg} + k_{deg}) (k_{sec} + k_u) + k_f k_{sec}} \]  

The crossing over (temporally) from the first to the second time period happens gradually and continuously. But one way to estimate the cross-over time is to set the right-hand sides of equations (10) and (14) equal to each other and solve for \( t \) (time), which yields

\[ t_c = \frac{\ln \left( 1 + \frac{k_{sec} k_f}{k_u (k_{agg} + k_{deg})} \right)}{k_{sec}} = \frac{\ln \left( 1 + \frac{k_{sec} k_f}{k_u (k_{agg} + k_{deg})} \right)}{k_{sec}} = \frac{\ln \left( 1 + \frac{k_{sec} k_f}{k_u (k_{agg} + k_{deg})} \right)}{k_{sec}} \]  

where \( K_f = k_f / k_u \) is the equilibrium constant for folding and \( t_c \) is the cross-over time. So the time it takes to reach the second period depends on either \( 1/k_u \) or on \( K_f (k_{agg} + k_{deg}) \), whichever is larger. It also depends on \( k_{sec} \), but in a more subtle way because \( k_{sec} \) appears in both the numerator and the denominator. If \( k_{sec} \) is large (secretion is fast), then \( t_c \) will be small, but if \( k_{sec} \) is small or approaches 0—reflecting situations in which secretion is slow or doesn’t happen at all—then \( t_c \) will approach

\[ t_c = \frac{1}{k_u} + \frac{k_f}{k_u (k_{agg} + k_{deg})} = \frac{1}{k_u} + \frac{K_f}{k_{agg} + k_{deg}} \]  

Case 2: Irreversible ligand binding with ligand present at a constant concentration

When the volume of cells is small relative to the volume of the media, the media will serve as a large reservoir of ligand for the cells. The free (unbound) ligand concentration will then be effectively constant inside the cells (note that the concentration of ligand in the cells and in the media may not be the same, but they should be proportional; that is,
an n-fold increase in the ligand concentration in the media should result in an n-fold increase inside the cells). Under these conditions, and if \([L] \gg K_d\) of the complex, natively folded protein should bind effectively and irreversibly to the ligand so that \([N:L]\) accumulates while \([N]\) quickly reaches steady state. Of course, this “steady state” is again not a system-wide steady state because the concentrations of some species are changing, so as before we will refer to it as a pseudo-steady state.

The pseudo-steady state concentrations of \(U\) and \(N\) are:

\[
[U]_{ss} = \frac{\sigma (k_u + k_a[L]) + k_{sec}}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f (k_{sec} + k_a[L])} 
\]  

(20)

\[
[N]_{ss} = \frac{\sigma k_f}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f(k_{sec} + k_a[L])} 
\]  

(21)

Inserting these expressions for \([U]\) and \([N]\) into equations (3) and (5) gives the following for the pseudo-steady state rate equations for \([A]\) and \([P_{deg}]\):

\[
\left(\frac{d[A]}{dt}\right)_{ss} = k_{agg} \frac{\sigma (k_u + k_a[L])}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f (k_{sec} + k_a[L])} 
\]  

(22)

\[
\left(\frac{d[Deg]}{dt}\right)_{ss} = k_{deg} \frac{\sigma (k_u + k_a[L])}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f (k_{sec} + k_a[L])} 
\]  

(23)

Solving these differential equations gives the following for the time-dependent concentrations of \(A\) and \(P_{deg}\):

\[
[A]_t = \frac{\sigma k_{agg} (k_{sec} + k_u + k_a[L]) t}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f (k_{sec} + k_a[L])} 
\]  

(24)

\[
[Deg]_t = \frac{\sigma k_{deg} (k_{sec} + k_u + k_a[L]) t}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f (k_{sec} + k_a[L])} 
\]  

(25)

Inserting the expression for \([N]_{ss}\) into equation (S4) and solving gives the time-dependent concentration of \(N:L\):
\[ [N:L]_t = k_a[L][N_{ss}] \left(1-e^{k_{sec}t}\right) = \frac{\sigma k_a[L] k_f}{(k_{agg}+k_{deg})(k_{sec}+k_{a[L]}+k_{f}(k_{sec}+k_{a[L]})} \left(1-e^{-k_{sec}t}\right) \] (26)

When there is no secretion \((k_{sec} = 0)\), this becomes

\[ [N:L]_t = \frac{\sigma k_a[L] k_f}{(k_{agg}+k_{deg})(k_{a[L]}+k_{f}k_{a[L]})} \] (27)

Finally, inserting the expression for \([N:L]_t\) from equation (26) into equation (6) and solving gives the time-dependent concentration of \(P_{sec}\):

\[ [\text{Sec}]_t = \frac{\sigma k_f}{(k_{agg}+k_{deg})(k_{sec}+k_{a[L]}+k_{f}k_{a[L]})} \left[(k_{sec} + k_{a[L]}+k_{f}k_{sec})t - k_{a[L]} \left(1-e^{-k_{sec}t}\right)\right] \] (28)

When \(k_{sec}t \gg 1\), \([\text{Sec}]_t\) approaches

\[ [\text{Sec}]_t = \frac{\sigma k_f}{(k_{agg}+k_{deg})(k_{sec}+k_{a[L]}+k_{f}k_{a[L]})} \left[(k_{sec} + k_{a[L]}+k_{f}k_{sec})t - k_{a[L]} \right] \] (29)

An observable for \textit{in vivo} protein expression experiments.

Perhaps the most general and easily measured observable for \textit{in vivo} protein folding in the presence of a fast- and tight-binding ligand is the fraction of the total synthesized protein that is neither aggregated nor degraded at a given time point. A general expression for this quantity, which we shall denote \(F_r\) (for “fraction remaining”) is:

\[ F_r = 1 - \left(\frac{[A]_t+[Deg]_t}{[Tot]_t}\right) \] (30)

\([Tot]_t\) is simply equal to the protein expression rate multiplied by the time: \([Tot]_t = \sigma t\). Inserting this and equations (24) and (25) into this equation yields
\[ F_r = 1 - \frac{\sigma k_{agg}(k_{sec} + k_u + k_a[L]) t}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f(k_{sec} + k_a[L])} = 1 - \frac{\sigma k_{deg}(k_{sec} + k_u + k_a[L]) t}{(k_{agg} + k_{deg})(k_{sec} + k_u + k_a[L]) + k_f(k_{sec} + k_a[L])} \] (31)

This can be simplified to

\[ F_r = 1 - \frac{1}{1 + \left( \frac{k_f}{k_{agg} + k_{deg}} \right) \left( \frac{1}{k_{sec} + k_{agg}} \right) \left( \frac{k_{agg} + k_{deg}}{k_u} \right)} = 1 - \frac{1}{1 + B_1 \left( \frac{1}{1 + B_2 + B_3[L]} \right)} \] (32)

where \( B_1 = k_f / (k_{agg} + k_{deg}) \), \( B_2 = k_{sec} / k_u \), and \( B_3 = k_u / k_u \). At very high ligand concentrations, \( F_r \) approaches

\[ F_r, [L] \to \infty = 1 - \frac{1}{1 + B_1} = 1 - \frac{1}{1 + \left( \frac{k_f}{k_{agg} + k_{deg}} \right)} \] (33)

In the absence of ligand and at times such that \( t > t_c \) (where \( t_c \) is defined in equations (S18) and (S19)), \( F_r \) is given by

\[ F_r, |t > t_c| [L] = 0 = 1 - \frac{[\text{agg}]_t |_{t > t_c} + [\text{deg}]_t |_{t > t_c} + \frac{[\text{agg}]_t |_{t > t_c} + [\text{deg}]_t |_{t > t_c}}{[\text{agg}]_t |_{t > t_c}}}{[\text{agg}]_t |_{t > t_c}} \] (34)

Inserting equations (8), (9), (15), and (16) into equation (34) and then collecting terms gives

\[ F_r, |t > t_c| [L] = 0 = 1 - \frac{\sigma (k_{agg} + k_{deg})^t |_{t > t_c} - \sigma (k_{agg} + k_{deg})^t |_{t < t_c}}{\sigma t} \] (35)

Equation (34) can be simplified and rearranged to give:

\[ F_r, |t > t_c| [L] = 0 = 1 - \frac{1}{1 + \left( \frac{k_f}{k_{agg} + k_{deg}} \right) \left( \frac{k_{sec}}{k_u + k_{sec}} \right)} + \frac{t_c}{t} \left( \frac{1}{1 + \left( \frac{k_f}{k_{agg} + k_{deg}} \right) \left( \frac{k_{sec}}{k_u + k_{sec}} \right)} - \frac{1}{1 + \left( \frac{k_f}{k_{agg} + k_{deg}} \right)} \right) \] (36)
The quantity in parenthesis has to be positive (because \(k_{sec}/(k_a+k_{sec}) < 1\)), but the third term has to be less than the second term (because it has been stipulated that \(t_c/t < 1\)), so \(F_r < 1\). Furthermore, \(F_r\) diminishes monotonically with time until, at very long times where \(t_c/t\) is very small, the third term can be dropped and \(F_r\) approaches the time independent value:

\[
F_r|_{t_c/t=0} = 1 - \frac{k_f}{1 + \frac{k_{sec}}{(k_{ag}+k_{deg})(k_{iu}+k_{sec})}}
\]

(37)

In terms of the constants \(B_1, B_2\) and \(B_3\), equations (35) and (36) can be written as follows:

\[
F_r|_{t>t_c} = 1 - \frac{1}{1+B_1\left(\frac{B_2}{1+B_2}\right)} + \frac{t_c}{t} \left(\frac{1}{1+B_1\left(\frac{B_2}{1+B_2}\right)} - \frac{1}{1+B_1}\right)
\]

(38)

Substituting \(B_1\) into the expression for \(t_c\) in equation (19) and then using the result as a rough approximation for \(t_c\) finally gives:

\[
F_r|_{t>t_c} = 1 - \frac{1}{1+B_1\left(\frac{B_2}{1+B_2}\right)} + \frac{1}{k_{iu}t} \left(\frac{1+B_1}{1+B_1\left(\frac{B_2}{1+B_2}\right)} - 1\right)
\]

(39)
Table A.1 Chaperone concentrations upon induction

The above table shows the concentrations of chaperones achieved during each run of the titration experiments in chapter 3 with the designated concentration of arabinose. The concentration listed shows the concentration of the active form of the chaperone, tetradecameric for GroEL, hexameric for lon and monomeric for DnaK.
<table>
<thead>
<tr>
<th>[IPTG] μM</th>
<th>[MetK] μM</th>
<th>F\text{agg}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trial 1</td>
<td>Trial 2</td>
</tr>
<tr>
<td>1</td>
<td>174.8</td>
<td>157.3</td>
</tr>
<tr>
<td>5</td>
<td>229</td>
<td>277.9</td>
</tr>
<tr>
<td>10</td>
<td>592.7</td>
<td>715</td>
</tr>
<tr>
<td>50</td>
<td>1139.9</td>
<td>1354.9</td>
</tr>
<tr>
<td>200</td>
<td>1695.8</td>
<td>1875.9</td>
</tr>
</tbody>
</table>

Table A.2 Concentrations and F\text{agg} for MetK synthesized at varying IPTG concentrations

The above table shows the concentration of MetK (monomeric) achieved under basal conditions of proteostasis at the designated concentration of IPTG and the fraction of MetK which aggregates at each concentration of MetK synthesized.
Table A.3 Concentrations and $F_{agg}$ for MetK synthesized at varying IPTG concentrations at varying DnaKJE concentrations

The above table shows the concentration of MetK synthesized in the DnaKJE titration experiments. The average DnaK concentration is the left most column and the concentration of MetK synthesized (at the designated concentration of IPTG) and the fraction aggregated for each experiment is listed.
Table A.4 Concentrations and $F_{agg}$ for MetK synthesized at varying IPTG concentrations at varying GroELS concentrations

The above table shows the concentration of MetK synthesized in the GroELS titration experiments. The average GroEL concentration is the left most column and the concentration of MetK synthesized (at the designated concentration of IPTG) and the fraction aggregated for each experiment is listed.
Table A.5 Concentrations and $F_{\text{agg}}$ for MetK synthesized at varying IPTG concentrations at varying Lon concentrations

The above table shows the concentration of MetK synthesized in the Lon titration experiments. The average Lon concentration is the left most column and the concentration of MetK synthesized (at the designated concentration of IPTG) and the fraction aggregated for each experiment is listed.


Pande, V.S., Beauchamp, K., and Bowman, G.R. (2010). Everything you wanted to know about Markov State Models but were afraid to ask. Methods 52, 99-105.


