Evolutionary Capacitance and Control of Protein Stability in Protein-Protein Interaction Networks

Purushottam D. Dixit[1], Sergei Maslov[1,2,3]

1 Biology, Brookhaven National Laboratory, Upton, New York, United States of America, 2 Physics and Astronomy, Stony Brook University, Stony Brook, United States of America, 3 Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York, United States of America

Abstract

In addition to their biological function, protein complexes reduce the exposure of the constituent proteins to the risk of undesired oligomerization by reducing the concentration of the free monomeric state. We interpret this reduced risk as a stabilization of the functional state of the protein. We estimate that protein-protein interactions can account for $\sim 2-4\ k_B T$ of additional stabilization; a substantial contribution to intrinsic stability. We hypothesize that proteins in the interaction network act as *evolutionary capacitors* which allows their binding partners to explore regions of the sequence space which correspond to less stable proteins. In the interaction network of baker's yeast, we find that statistically proteins that receive higher energetic benefits from the interaction network are more likely to misfold. A simplified fitness landscape wherein the fitness of an organism is inversely proportional to the total concentration of unfolded proteins provides an evolutionary justification for the proposed trends. We conclude by outlining clear biophysical experiments to test our predictions.

Introduction

The toxicity due to protein misfolding and aggregation has a considerable effect on the viability of living organisms [1–5]. Consequently, cells are under strong selection pressure to evolve thermodynamically stable [6] and aggregation-free protein sequences [7]. The internal region of stable proteins has a tightly packed core of hydrophobic residues. A mutation in the core may disrupt the entire protein structure. Consequently, the core residues are strongly conserved [8,9]. In contrast, mutations on the surface contribute weakly to the thermodynamic stability of proteins [10] yet surfaces show significant level of conservation [11] owing to protein-protein interactions.

Recent high throughput experiments have established that proteins interact with each other on a genome-wide scale [12]. Such 'small world' networks are thought to facilitate biological signaling and ensure that cells remain robust even after a random failure of some of its components [13]. It is thought that evolutionarily, multi-protein complexes are favored over larger size of individual proteins [14] since large proteins are difficult to fold and expensive to synthesize while small interacting proteins can fold independently and then efficiently assemble into large complexes. Individual interaction between proteins can give rise to cooperativity and allostery which results in a finer control over the functional task the protein complex performs. Protein-protein interactions (PPI) are also thought to prevent protein aggregation [15,16]. Lastly, many proteins can perform promiscuous function in that they can partake in multiple protein complexes. Interestingly, proteins in higher organisms are involved in more interactions and form larger protein complexes compared to more primitive life forms [17].

Here, we hypothesize an additional biophysical advantage for protein-protein interactions. Proteins bound to their interaction partners *effectively* present a lower monomer concentration inside the cell. Since free monomers are susceptible to misfolding/unfolding and toxic oligomerization, interacting proteins may face a reduced risk towards the same. This reduced risk can be interpreted as interaction-induced stabilization $\Delta\Delta G_{\mathrm{ppi}}$ — stabilization due to the protein-protein interaction network — of an otherwise monomeric protein (see Fig. 1 for a cartoon). We propose that by giving proteins an additional stability, each protein in the interaction network acts as an *evolutionary capacitor* [18,19] in the evolution of its binding partners: proteins are allowed to explore the less stable regions (regions of low intrinsic stability) of the sequence space as long as they are stabilized by their interaction partners. Inversely, unstable proteins are expected to receive significant additional stability from the interaction network.

Below we outline the empirical evidence for our hypothesis and suggest clear biophysical and evolutionary experiments to test it further.

Results

We present our estimates of the interaction-induced stability $\Delta\Delta G_{\mathrm{ppi}}$ (see Methods) and explore the evolutionary interplay between $\Delta\Delta G_{\mathrm{ppi}}$ and protein stability $\Delta G_{\mathrm{folding}}$ using a simplified fitness model for a toy proteome. We test the predictions of the toy

Author Summary

The folded form of proteins is only marginally stable *in vivo* and constantly faces the risk of aggregation, unfolding/misfolding, and other aberrant interactions. For most proteins, the folded form is also the functionally relevant one and forces of natural selection strongly modulate its stability. *In vivo*, proteins interact with each other on a genome-wide scale. Usually, the interaction of a protein and its binding partners requires both the proteins to be in the folded form and as a result, the interactions tend to shift the population of a protein towards the folded form. Consequently, protein-protein interactions interfere with the evolution of protein stability. Here, we present empirical evidence and theoretical justification for proteins' ability to stabilize the folded form of their interaction partners and allow them to explore the region of the sequence space that corresponds to proteins with less stable structure. We argue that the 'evolutionary capacitance' – previously thought to be a property of the chaperone HSP90, a special class of proteins – is a property of all proteins, albeit to a different degree.

model on the proteome of baker's yeast. The fitness model also sheds light on the interplay between protein stability and protein abundance.

Interaction-induced stability $\Delta\Delta G_{ppi}$ is comparable to inherent stability $\Delta G_{folding}$

Fig. 2 shows the histogram of the estimated interaction-induced stability $\Delta\Delta G_{ppi}$ for ~ 1600 cytoplasmic yeast proteins for whom abundance, interaction, and localization data is available (see Methods for the details of the calculations). Note that the average PPI induced stability is $\sim 2k_B T$ and can be as high as $5-6\ k_B T$. This stabilization is dependent not only on the number of interaction partners of a given protein or the strengths of those interactions but also on the relative abundances of the interaction partners. In fact, the interaction-induced stability of a protein correlates strongly with the relative concentration of its binding partners

$$\sigma_A = \frac{1}{C_A} \sum_{B\ nn\ A} C_B$$

(Spearman $r = 0.64$, $p < 10^{-5}$). This suggests a plausible mechanism of stabilization of a protein without changing its sequence viz. via adjusting the expression levels of its interaction partners (see Discussion below).

The estimated $\Delta\Delta G_{ppi}$ values are of the same order of magnitude as the inherent stabilities of proteins, $\Delta G_{folding}$ ($\sim 5-15k_B T$) [9]. Given that random mutations are more likely to destabilize proteins [6], we expect protein-protein interactions to act as secondary mechanisms to stabilize proteins and to interfere with the evolution of protein stability.

Simplified fitness model explores the interplay between $\Delta\Delta G_{ppi}$ and $\Delta G_{folding}$

To explore the evolutionary consequences of the interaction-induced stability, we investigate a simplified fitness model of a toy proteome consisting of 15 proteins (see Methods, Text S1, and Table S1). Briefly, the fitness of the cell depends only on the total concentration of unfolded proteins in it [20]. During the course of evolution, each protein acquires random mutations that change either a) its inherent stability $\Delta G_{folding}$ or b) the dissociation constant of its interaction with a randomly selected interaction partner. Even though protein abundance and protein-protein interactions evolve at the same time scale as protein stability, the former are dictated largely by the biological function of the involved proteins. Incorporating the fitness effects of changes in expression levels and interaction partners in our simple model is non-trivial. Thus, in order to specifically probe the relation between stability and interactions, we do not allow proteins to change their abundance and interaction partners.

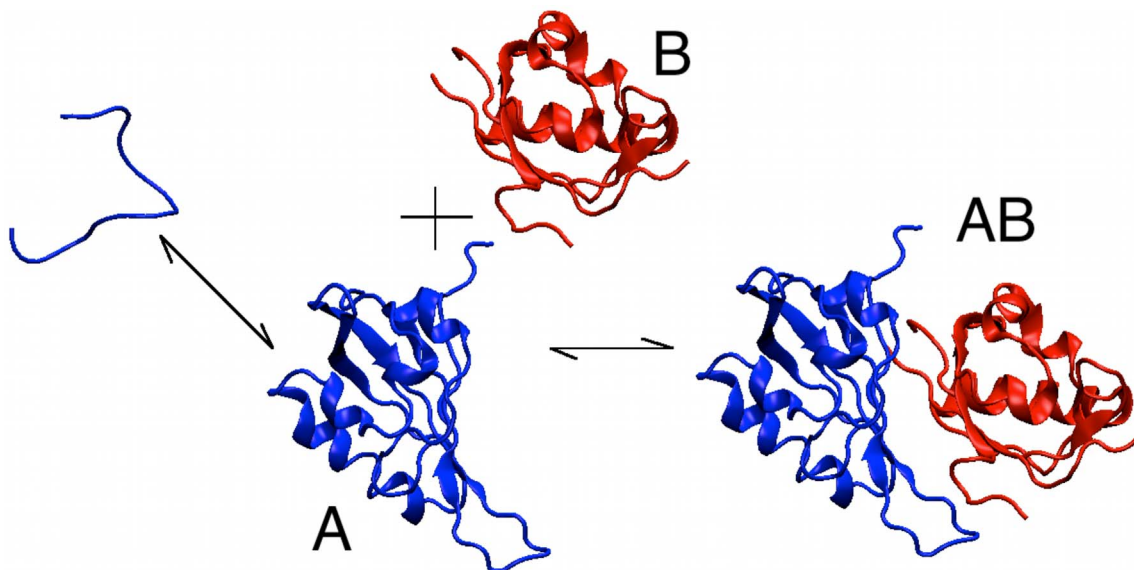**Figure 1. The equilibrium between the folded state of protein A (blue protein) and its unfolded/insoluble state (blue coil) is affected by the interactions of the folded state with its interaction partner B (red).** The formation of the AB dimer lowers the population of the unfolded/insoluble state of protein A and effectively stabilizes the folded state.
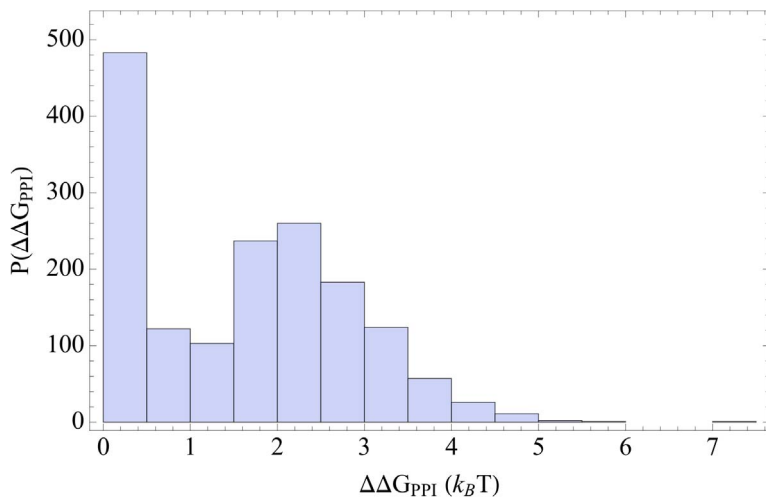doi:10.1371/journal.pcbi.1003023.g001

Figure 2. The histogram of estimated PPI-induced stabilities for the yeast cytoplasmic proteome (See main text). While the average stability is $\sim 2\ k_B T$, some proteins can receive as much as $5-6\ k_B T$ of stability from their binding partners. Note that the peak near $\Delta\Delta G_{\mathrm{ppi}} \approx 0$ is due to proteins which have no interaction partners and are by definition not stabilized by the PPI network.
doi:10.1371/journal.pcbi.1003023.g002

In the model, the concentration of unfolded proteins and thus the fitness of the proteome depends on the total stability $\Delta G_{\mathrm{total}} = \Delta G_{\mathrm{folding}} + \Delta\Delta G_{\mathrm{ppi}}$ of individual proteins. While random mutations are more likely to make proteins unstable, protein-protein interactions increase the total stability. In the canonical ensemble description of the evolution of fitness [21], the inverse effective population size $(1/f_0 Ne)$, the *evolutionary temperature* quantifies the importance of genetic drift. The *effective* population size modulates the competition between destabilizing random mutations and stabilizing protein-protein interactions.

We find that at higher effective populations, proteins are inherently stable and only the least stable proteins (small $\Delta G_{\mathrm{folding}}$) receive high stabilization from the interaction network (high $\Delta\Delta G_{\mathrm{ppi}}$). At low effective population, due to genetic drift, proteins are inherently destabilized and protein-protein interactions serve as the primary determinant of the *effective* stability of proteins. Fig. 3 shows the dependence of average inherent stability ($\Delta G_{\mathrm{folding}}$), average interaction-induced stability ($\Delta\Delta G_{\mathrm{ppi}}$), and average total stability ($\Delta G_{\mathrm{total}}$) with effective population size. Interestingly, the total stability ($\Delta G_{\mathrm{folding}} + \Delta\Delta G_{\mathrm{ppi}}$) of proteins remains relatively insensitive to changes in population size.

We observe that the correlation coefficient between the inherent stability $\Delta G_{\mathrm{folding}}$ and the interaction-induced stability $\Delta\Delta G_{\mathrm{ppi}}$ itself varies with the effective population size. Even though its magnitude decreases, interaction-induced stability becomes more and more correlated with inherent stability as population size increases (See Fig. 4). In real life organisms, interaction-induced stability acts on a need basis for proteins and serve as a secondary stabilization mechanism. In the drift-dominated regime, which is unlikely to be realized in real life organisms (except probably in parasitic microbes with low population sizes), interaction-induced stability becomes the dominant player in the evolution of total stability of proteins [17]. We next examine if this prediction from the toy model holds for real organisms.

Induced stability correlates with aggregation propensity

Proteome-wide information about the inherent stability of proteins $\Delta G_{\mathrm{folding}}$ is currently unavailable. Previously, *in silico* estimates of protein aggregation propensity have been used as proxy for protein stability [22,23]. We use the TANGO [24] algorithm to estimate protein aggregation propensity. It is known

that TANGO aggregation propensity correlates strongly and negatively with protein stability [24]. TANGO has been verified extensively with experiments on peptide aggregation [24] and has been previously used to study the evolutionary aspects of protein-protein interactions [22,25]. Similar analysis for Aggrescan [26] can be found in Text S1 and Table S3. We find that the aggregation propensity Z is correlated positively with the interaction-induced stability $\Delta\Delta G_{\mathrm{ppi}}$ (Spearman $r = 0.11$, $p < 10^{-5}$). As expected [2], the aggregation propensity Z is negatively correlated with protein abundance C (Spearman $r = -0.11$, $p < 10^{-5}$). The correlation between Z and $\Delta\Delta G_{\mathrm{ppi}}$ does not depend on this underlying dependence and persists even after controlling for total abundance C (partial Spearman $r = 0.11$, $p < 10^{-5}$) (See Table S2). This result suggests in the proteome of baker's yeast, protein stability correlates negatively with interaction-induced stability.

Aggregation propensity correlates *principally* with free monomer abundance

The fitness cost of protein aggregation is directly proportional to the amount of aggregate [20]. Thus, the selection forces that make protein sequences aggregation-free act more strongly on highly expressed proteins [1,2,22]. Our hypothesis suggests that the proteins that are bound to their interaction partners present a lower concentration of the *free monomeric* state *in vivo* (low F) and automatically lower the misfolding/aggregation induced fitness cost, even if highly abundant (high C). The selection forces to evolve an aggregation-free sequence may be weaker for such proteins. Consequently, the aggregation propensity Z should be principally correlated with the free monomer concentration F rather than the total abundance C.

Indeed, we observe that the estimated monomer concentration F and the aggregation propensity Z are correlated negatively (Spearman $r = -0.15$, $p < 10^{-5}$). Importantly, this correlation is not an artifact of the underlying correlation between the aggregation propensity and total abundance C (partial Spearman $r = -0.10$, $p < 10^{-5}$). At the same time, the partial correlation coefficient between the aggregation propensity Z and the total protein abundance C controlling for the estimated monomer concentration F is minimal (partial Spearman $r = 0.03$, $p > 0.15$). In short, the total free monomer concentration F of a protein
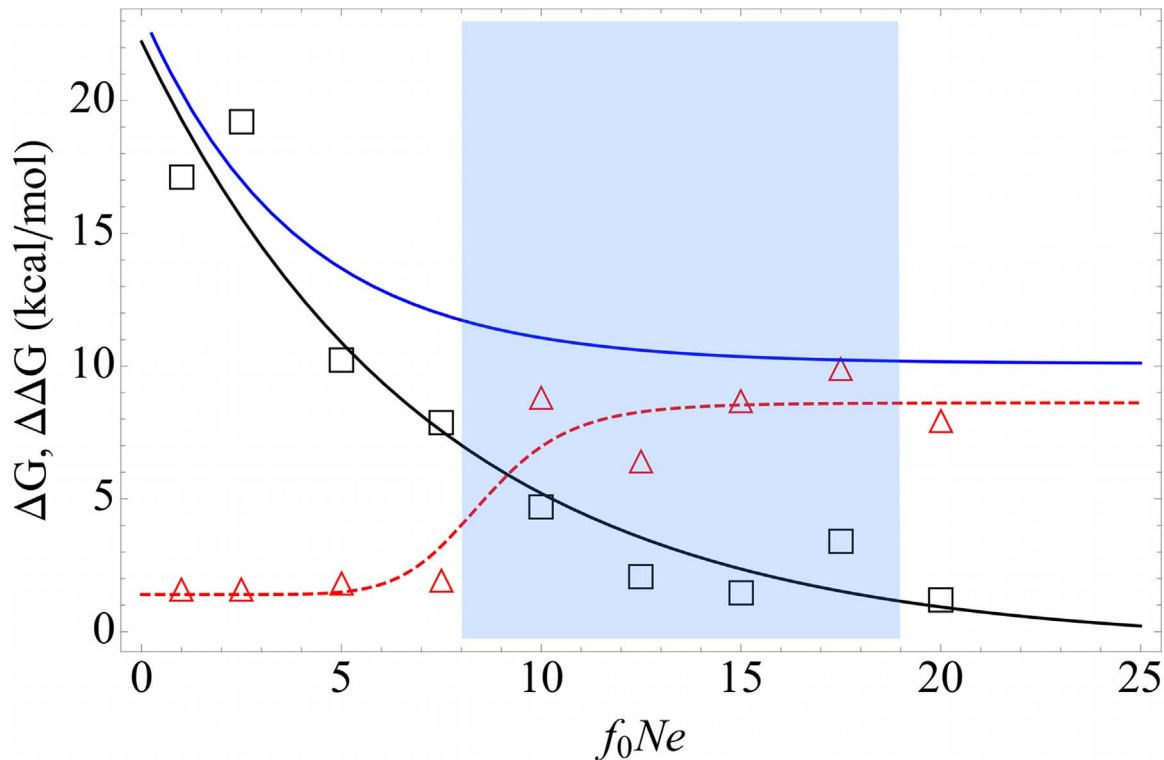
Figure 3. The average of inherent stability $\Delta G_{\text{folding}}$ (triangles) and the interaction-induced stability $\Delta\Delta G_{\text{ppi}}$ (squares) as a function of effective population size $f_0 Ne$ for the toy proteome. The curves are fitted to the data only to highlight trends, blue curve represents the total stability $\Delta G_{\text{folding}} + \Delta\Delta G_{\text{ppi}}$. Population size $f_0 Ne$ is in arbitrary units. The shaded area roughly represents the region of the red and the black curve that correspond to the empirically observed folding free energies $\Delta G_{\text{folding}}$ ($3.5-10$ kcal/mol) [9] and the estimated interaction-induced free energy $\Delta\Delta G_{\text{ppi}}$ ($1.5-3$ kcal/mol).
doi:10.1371/journal.pcbi.1003023.g003

(rather than C, its total abundance) might be a better variable to relate to evolutionary and biophysical constraints on the protein.

Interacting proteins as evolutionary capacitors

We have thus far shown that a protein's interaction partners can significantly stabilize its folded state and this stabilization interferes with the evolution of the inherent stability of the protein. We now explore the reverse viz. the evolutionary consequences of the ability of each protein to impart stability to its interaction partners.

The concept of evolutionary capacitor has been previously introduced for the heat shock protein HSP90 [18,19], which is also a molecular chaperone and a highly connected hub in the PPI network (70 interaction partners in the current analysis). An elevated concentration of HSP90 buffers the potentially unstable variation in proteins, which may allow proteins to sample a wider region of the sequence space, which may often lead to functional diversification [27]. Similar to HSP90, each protein in the interaction network has some ability to stabilize its interaction partners to a certain extent. Consequently, we study the evolutionary *capacitance* C_i of individual proteins in the context of the interaction network by estimating the effect of protein knockout on ppi-induced stability *in silico*. Proteins with higher evolutionary capacitance are defined as those with the higher cumulative destabilizing effect on the proteome. We write,

$$C_i = \sum_{\substack{j \in \text{destabilized} \\ \text{by } i-\text{knockout}}} \left(\Delta\Delta G_{\text{ppi }(j)}^{i,\text{knockout}} - \Delta\Delta G_{\text{ppi }(j)}^{\text{wt}} \right). \quad (1)$$

For each protein i, the sum in Eq. 1 is carried out over all proteins j that are destabilized due to its knockout. Here, we assume that the potential of a given protein knockout to generate multiple phenotypes depends on the loss of stability of its interaction partners caused by its knockout. We hypothesize that, similar to unstable proteins requiring HSP90 to fold, the interaction partners of proteins with high capacitance should be unstable. In fact, the capacitance C_i of a protein and the mean aggregation propensity $\langle Z_{\text{neighbors}} \rangle$ of its interaction partners are strongly correlated (Spearman $r = 0.53$, $p < 10^{-5}$). The capacitance C_i is significantly correlated with $\langle Z_{\text{neighbors}} \rangle$ even after controlling for the abundance of the protein (partial spearman $r = 0.55$, $p < 10^{-5}$) and the number of its interaction partners (partial spearman $r = 0.14$, $p < 10^{-5}$). This suggests that a protein needs to be present in sufficient quantity and should interact with a large number of proteins in order to effectively act as a capacitor.

We have presented evidence that all proteins can act as an *evolutionary capacitor*, albeit with variable effectiveness, for their interaction partners. Traditionally, evolutionary capacitors are understood to be chaperones that buffer phenotypic variations by helping misolding-prone proteins fold in a proper structure [19]. Not surprisingly, when we carried out functional term enrichment analysis using gene ontology [28], we found that approximately half of the top 20 capacitors have 'chaperone' in their name. The top 20 are also over represented in the chaperone-like molecular function of *protein binding* and *unfolded protein binding* ($p \le 0.0025$) and the biological process of *protein folding* ($p \le 0.01$). These findings validate our definition of capacitors that were previously identified as chaperones. Interestingly, some of the predicted capacitors do
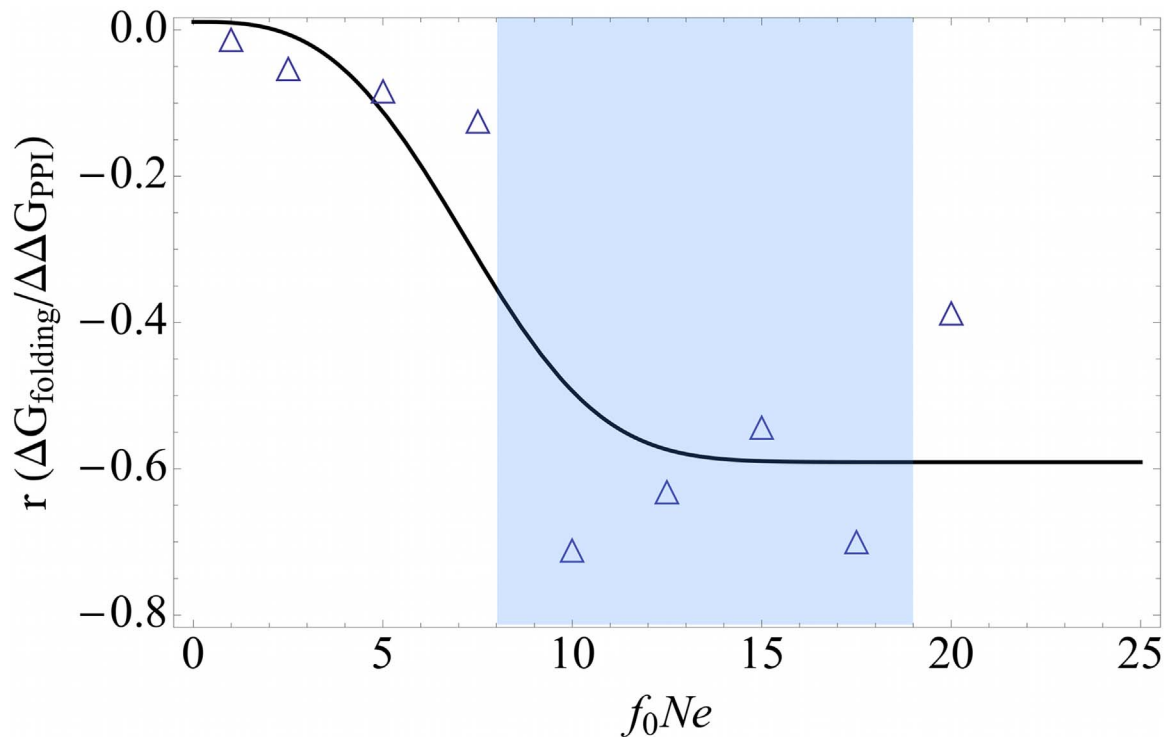
Figure 4. The spearman correlation coefficient between interaction-induced stability $\Delta\Delta G_{ppi}$ and inherent stability $\Delta G_{folding}$ as a function of effective population size $f_0 Ne$ (See supplementary Text S1). Population size is in arbitrary units. The blue region identifies the location of real life proteomes (See Fig. 3).
doi:10.1371/journal.pcbi.1003023.g004

not currently have a protein folding-related functional annotation. These need more experimental investigation (see supplementary File S1 for the list). This suggests that previously identified *evolutionary capacitor* HSP90 may in fact only be one among the broader set of *evolutionary capacitors*. Every protein in the interaction network is an *evolutionary capacitor* for its interaction partners and *evolutionary capacitor* is a quantitative distinction rather than a qualitative one.

Discussion

Recently, Fernández and Lynch [17] showed that random genetic drift is the chief driving force behind thermodynamically less stable yet densely interacting proteins in higher organisms [17]. Additionally, protein complexes in higher organisms have more members than in lower organisms [14]. Recently, it was observed that a destabilizing mutation in the enzyme DHFR in *E. coli* leads to functional tetramerization of the otherwise monomeric enzyme [29] suggesting that protein-protein interactions can at least partially compensate the effect of protein destabilization. β lactoglobulin is an aggregation-prone protein generally found as a dimer. It was shown that the specific interactions responsible for the formation of the dimer considerably reduce the risk of protein aggregation [16]. Ataxin-3 is a protein implicated in polygluta-mine expansion diseases wherein the functional interactions of the protein reduce the exposure of its aggregation prone interface and thereby decrease its aggregation propensity [15].

Here, we have quantified the interaction-induced stability on a proteome wide scale and hypothesized that the PPI-induced stabilization is a secondary evolutionary advantage of the PPI network; alleviating the selection pressure on proteins in functional multi-protein complexes to evolve a stable folded. A simple model

for the fitness of the proteome provided a fundamental justification for the co-evolution of protein stability and protein-protein interactions and made predictions that were tested on the proteome of baker's yeast. In the model, when the effects of natural selection are weak, proteins acquire stability mainly via protein-protein interactions. At a higher population size — in the absence of genetic drift — proteins are intrinsically stable and protein-protein interactions stabilize only those proteins that fail to evolve inherent stability.

We have also presented evidence that *all* interacting proteins stabilize their binding partners to a certain extent and act as the evolutionary capacitance [19] for their evolution. Interestingly, though some of the top 20 capacitors predicted in this study are known chaperones and are over-represented in GO ontology terms such as *protein binding*, *unfolded protein binding*, and *protein folding*; others do not have any protein folding-related functional annotation and need experimental investigation.

The importance of disordered proteins, especially in the proteomes of higher organisms, cannot be neglected. The proteome of baker's yeast does not have many completely disordered proteins but ∼17% of the amino acids in the proteins of yeast are predicted to be in a disordered state [30] (∼10% for the proteins considered in this study, see supplementary Text S1 and Fig. S4). Even though the development presented above applied only to an equilibrium between folded and unfolded/misfolded/aggregated protein, it can be easily generalized to disordered proteins. This is because even though the folded ⇌ unfolded equilibrium is not well defined, similar to well structured proteins, disordered proteins also exist either in a soluble monomeric (instead of the folded state), a misfolded/aggregated, and a complexed state. Many disordered proteins acquire a definite structure when bound to their interaction partners and

seldom dissociate to the soluble monomeric [31]. These serve as even stronger candidates for the beneficiaries of interaction-induced stability compared to folded proteins. Consequently, we include both partially disordered proteins and structured proteins in the current analysis of the ~ 1600 cytoplasmic proteins.

Suggested experimental tests

Modulation of protein stability by overexpression of its partners. We predict that the measured free energy of protein folding in vivo [32,33] will be lower than the in vitro measurement. Moreover, this free energy can be modulated by overexpressing the interaction partners of the protein that increases the equilibrium constant K_2 between the folded monomer and the generic complexed state. Recently, it was observed that the measured stability of phosphoglycerate kinase was higher by $\sim 3k_BT$ *in vivo* compared to *in vitro* [33].

Overexpression-instability epistasis. Does the PPI-induced stabilization have evolutionary advantages? We propose the following experimental test. Consider two mutated phenotypes for an isolated interacting pair of proteins A and B in an organism 1) A^*, a destabilized mutant of protein A and 2) $B\uparrow$ where B is overexpressed. We predict that lowering of the organismal fitness due to destabilization of protein A $(A \rightarrow A^*)$ can be at least partially rescued by the overexpression of the protein B $(B \rightarrow B\uparrow)$ i.e. the combination of two penalizing mutations may perhaps be advantageous to the organism.

Methods

Law of mass action and $\Delta\Delta G_{\text{ppi}}$

In cellular homeostasis, the total concentration C_A of any protein A can be written as the sum of its free folded monomer concentration F_A, a fraction comprising of insoluble oligomers and unfolded peptide U_A, and as part of all protein complexes D_A containing A (See Fig. 5). In our computational model, for simplicity and owing to the nature of the large scale data [34], we restrict protein complexes to dimers [35], thus for all proteins B that interact with A,

$$D_A = \sum_{B\,nn\,A} D_{AB} \tag{2}$$

Conservation of mass implies,

$$C_A = U_A + F_A + \sum_{B\,nn\,A} D_{AB}. \tag{3}$$

The concentration D_{AB} of each dimer AB satisfies the law of mass action,

$$D_{AB} = \frac{F_A F_B}{K_{AB}}. \tag{4}$$

We can write the balance between the three states of the protein, $U_A \rightleftharpoons F_A \rightleftharpoons D_A$ (See Fig. 1), as two *equilibrium* equations

$$F_A = K_{1A} \cdot U_A, \text{ and} \tag{5}$$

$$D_A = K_{2A} \cdot F_A. \tag{6}$$

Note that U_A comprises of a collection of biologically unusable states of the protein viz. the misfolded/unfolded and the

oligomerized state any of which may convert to/interact with the folded monomeric state F_A. Consequently, the first equilibrium $U_A \rightleftharpoons F_A$ is a collection of *thermodynamic equilibriums*. The equilibrium constant K_1 will thus depend not only on the temperature T but also on U_A and F_A. If among the unfolded, misfolded, and the oligomerized states the former dominates the population comprising U_A then, $K_{1A} = e^{\Delta G/k_BT}$ where $\Delta G > 0$ is the thermodynamic stability of the free monomeric state. Similarly, K_{2A} is given by,

$$K_{2A} = \sum_{B\,nn\,A} \frac{F_B}{K_{AB}} = \sum_{B\,nn\,A} \frac{U_B K_{1B}}{K_{AB}} \tag{7}$$

and depends not only on the dissociation constants K_{AB} but also the free concentrations F_B of the interacting partners of protein A and on the topology of the interaction network in the organism. Here too, we assume that a) only the folded monomeric forms of proteins interact with each other and b) there is no appreciable interaction between the collective unfolded state U_A of protein A and any state of any other protein B. We have also neglected the role of chaperones in actively reducing the concentration of the unfolded/misfolded/aggregated state by turning it over to the folded state. In fact, some of the chaperones are included in of our mass action equilibrium model and prevent unfolding by sequestering the folded state (see below and the discussion section).

By combining mass conservation (Eq. 3) with Eq. 5 and Eq. 6,

$$U_A = \frac{C_A}{1 + K_{1A}(1 + K_{2A})}. \tag{8}$$

In the above development, we have made a crucial assumption that only.

Note that in the absence of interactions, $U_A = C_A/(1 + K_{1A})$. We identify $1 + K_{2A}$ as the additional decrease in the insoluble fraction due to protein-protein interactions. We define the interaction-induced stability $\Delta\Delta G_{\text{ppi}}(>0)$ as,

$$\Delta\Delta G_{\text{ppi}} = k_BT \log(1 + K_{2A}) \tag{9}$$

Identification of proteins and the mass action model

We downloaded the latest set of interacting proteins in baker's yeast from the BIOGRID database [36]. To filter for non-reproducible interactions and experimental artifacts, we retained only those interactions that were confirmed in two or more separate experiments. For the sake of simplicity, we only considered cytoplasmic proteins [37] with known concentrations [38]. This lead to ~ 1600 proteins connected by ~ 5600 interactions.

The *in vivo* stability of a protein is a combination of its thermodynamic stability, resistance to aggregation or oligomerization, and resistance to degradation [39]. Note that the interaction-induced stability of a protein depends on the stability of its interaction partners (see Eq. 6, Eq. 7, and Eq. 9). Unfortunately, the exact dependence of the *in vivo* protein stability on its sequence is unclear and there exist no reliable data or sequence dependent computational estimates for the thermodynamic stability of proteins. Moreover, K_{2A}, and thus $\Delta\Delta G_{\text{ppi}}$ (Eq. 6, Eq. 7, and Eq. 9), can be estimated even in the absence of the knowledge of K_{1A}. In our estimates of $\Delta\Delta G_{\text{ppi}}$, we assume that K_{2A} is given simply by
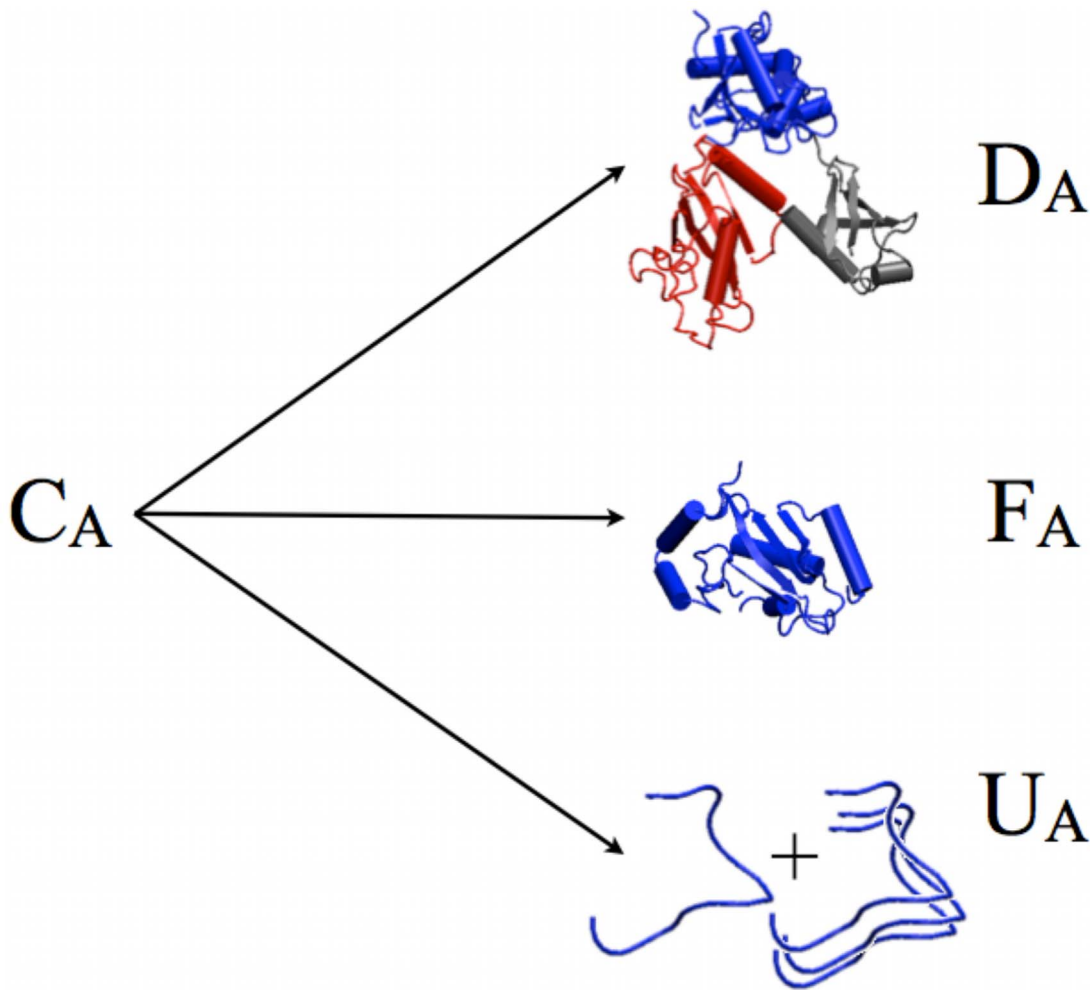
Figure 5. At steady state, protein A can be present either as a mixture of misfolded monomers and insoluble oligomers (U$_A$), a folded monomer F$_A$, or in a complex with its interaction partners (D$_A$).
doi:10.1371/journal.pcbi.1003023.g005

$$K_{2A} = \sum_{B \, nn \, A} \frac{F_B}{K_{AB}}.$$

Here, F_B is obtained by solving the mass action equations [35] iteratively (see below). This is equivalent to assuming that all the proteins are equally and highly stable ($K_{1A} \gg 1$ for all proteins A). The $\Delta\Delta G_{\text{ppi}}$ thus calculated serves as the upper limit of interaction-induced stability. In the supplementary materials (Text S1, Fig. S1, Fig. S2, and Tables S4 and S5), we show that different assignments of the equilibrium constants including a simple model of protein stability [40–42] do not change the qualitative nature of our observations.

The dissociation constants K_{AB} for protein-protein interactions follow a lognormal distribution with a mean $\langle 1/K_{AB} \rangle \approx 5$ nM [35]. The majority of interactions between proteins are neither too weak nor unnecessarily strong. Common sense dictates that it does not make sense to decrease the dissociation constant between two proteins beyond the point where the abundance limiting protein spends all of its time in the bound state. Motivated by these evolutionary arguments to minimize unnecessary protein production and to avoid unnecessarily strong interactions, Maslov and Ispolatov [35] devised a recipe to assign dissociation constants to individual protein-protein interactions. viz. for interacting proteins A and B, the dissociation constant $K_{AB} = \max(C_A, C_B)/20$. We also explore a few other assignment rules for dissociation constants (see supplementary Text S1, Fig. S3, and Table S6).

We solve for free concentrations F_A iteratively [35]. We start by setting $F_A = C_A$ for all proteins and iteratively calculate F_A from

$$F_A = \frac{C_A}{1 + \sum F_B / K_{AB}} \qquad (10)$$

till two consecutive estimates of F_A fall within 2.5% of each other for all proteins.

Simplified fitness model for cellular proteomes

As noted above, the toxic effects of misfolding and aggregation may be the chief determinant of protein sequence evolution [2,4,5]. The dosage dependent fitness effect of misfolded proteins [20] motivates us to introduce a simple biophysical model for fitness f of the proteome (See Eq. 11),

$$f = -f_0 \log \sum_A U_A. \qquad (11)$$

f_0 is the scaling factor. Potentially, f_0 can be estimated from fitness experiments by introducing measured quantities of unfolded protein in the cell [20]. We explore the evolution of a hypothetical proteome to investigate the interplay between protein stability and protein-protein interactions.

We believe that protein abundances and the topology of the interaction network are largely dictated by biological function. It is non-trivial to incorporate the fitness effect of changes in gene expression level and the network topology in our simplified model. Thus, to specifically probe the relation between stability and interactions, we concentrate on the effect of toxic gain of function due to misfolding and aggregation on cellular fitness and not include changes in gene expression levels and network topology. In this aspect, our model is in the same spirit as previously proposed models [6,41–48]. The effect of random mutations on average destabilizes proteins and the dynamics of the evolution of thermodynamic stability of proteins can be modeled as a random walk with negative average velocity [6]. We consider the thermodynamic stability as a proxy for the *in vivo* stability of proteins. We construct the cytoplasm of a hypothetical organism with 15 proteins. The number of proteins is low due to computational restrictions. The proteome is evolved by sampling the dissociation constants from the lognormal distribution while introducing random mutations in proteins that change their stability. At each generation, the fitness is evaluated and the progeny is accepted at a certain *evolutionary temperature* (defined as the inverse of the effective population size, $1/f_0 Ne$) [21]. We run a total of 200000 generations for each *evolutionary temperature* and analyze the organism in the latter half of the evolutionary run (details of the model and a brief description of the population genetics terminology is in supplementary Text S1).

Aggregation propensity

The notion of protein stability relevant to this study is the propensity of a protein to avoid structural transformations that may render it unemployable for biological function. For example, for a small and highly soluble protein, this stability corresponds to the thermodynamic stability of the native state while for a large multi domain protein, it may correspond to the thermodynamic stability of one of its domains against the partially unfolded state. In short, thermodynamic stability of the folded state with respect to the unfolded, partially folded state, and the misfolded state all contribute to the *in vivo* stability of proteins [39].

Though there is a lack of proteome-wide estimates of thermodynamic stability of proteins, the aggregation propensity can be estimated from the sequence [24,26] and is known to be correlated with protein stability [24]. In our correlation analysis, we use the estimated aggregation propensity as a proxy for *in vivo* protein stability and explore the relationship between interaction-induced stability $\Delta\Delta G_{\mathrm{ppi}}$ and protein stability. The aggregation propensity was estimated for the same ~ 1600 proteins used in the mass action calculation to estimate $\Delta\Delta G_{\mathrm{ppi}}$. We tested the TANGO [24] and Aggrescan [26] to estimate the aggregation propensity of proteins. Previously, TANGO has been used [22,23,49] to understand the relation between protein abundance

and instability. We show results for TANGO in the main text. Aggrescan results (supplementary Text S1 and Table S3) are quite similar.

Supporting Information

Figure S1 The histogram of interaction-induced stabilities $\Delta\Delta G_{\mathrm{ppi}}$ when protein stabilities depend on their chain length.
(TIF)

Figure S2 The histogram of interaction-induced stabilities $\Delta\Delta G_{\mathrm{ppi}}$ when protein stabilities are set at their minimum.
(TIF)

Figure S3 The histogram of interaction-induced stabilities $\Delta\Delta G_{\mathrm{ppi}}$ when all dissociation constants are set at 5 nM.
(TIF)

Figure S4 The histogram of estimated disorder in the proteins of the yeast proteome.
(TIF)

Table S1 A table for the parameters and topology of the toy proteome.
(PDF)

Table S2 A table reporting correlations between stability and interaction using TANGO [24].
(PDF)

Table S3 A table reporting correlations between stability and interaction using AGGRESCAN [26].
(PDF)

Table S4 A table reporting correlations between stability and interaction when protein stabilities depend on their chain length.
(PDF)

Table S5 A table reporting correlations between stability and interaction when protein stabilities are set to their minumum.
(PDF)

Table S6 A table reporting correlations between stability and interaction when all dissociation constants are set at 5 nM.
(PDF)

Text S1 An inventory of population genetics terms, additional information about the toy model, and misc. information about the analysis.
(PDF)

Author Contributions

Conceived and designed the experiments: PDD SM. Performed the experiments: PDD. Analyzed the data: PDD. Contributed reagents/materials/analysis tools: PDD. Wrote the paper: PDD SM.

References

1. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci 102: 14338–14343.
2. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.
3. Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, et al. (2010) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci 108: 680–685.
4. Wilke CO, Drummond DA (2010) Signatures of protein biophysics in coding sequence evolution. Curr Opin Struc Biol 20: 385–389.
5. Olzscha H, Schermann SM, Woerner AC, Pinkert S, Hecht MH, et al. (2011) Amyloid-like Aggregates Sequester Numerous Metastable Proteins with Essential Cellular Functions. Cell 144: 67–78.
6. Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci 104: 16152–16157.

7. Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO reports 8: 737–742.

8. Alberts B, Bray D, Lewis J, Raff M, Roberts K, et al. (2002) Molecular biology of the cell. New York: Garland Science.

9. Branden C, Tooze J (1998) Introduction to protein structure. New York: Garland Science.

10. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The Stability Effects of Protein Mutations Appear to be Universally Distributed. J Mol Bio 369: 1318–1332.

11. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Bio 271: 511–523.

12. Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Bio Evol 18: 1283–1292.

13. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41–42.

14. Lynch M (2011) The evolution of multimeric protein assemblages. Mol Bio Evol 29: 1353–1366 doi: 10.1093/molbev/msr300.

15. Masino L, Nicastro G, Calder L, Vendruscolo M, Pastore A (2011) Functional interactions as a survival strategy against abnormal aggregation. The FASEB journal 25: 45–54.

16. Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M (2009) Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci 106: 10159–10164.

17. Fernández A, Lynch M (2011) Non-adaptive origins of interactome complexity. Nature 474: 502–505.

18. Rutherford SL, Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. Nature 396: 336–342.

19. Rutherford S, Swalla BJ (2007) The Hsp90 Capacitor, Developmental Remodeling, and Evolution : The Robustness of Gene Networks and the Curious Evolvability of Metamorphosis. Critical Reviews in Biochemistry 42: 355–372.

20. Geiler-samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL (2010) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci 108: 680–685.

21. Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. Proc Natl Acad Sci 102: 9541–9546.

22. Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Sys Bio 6: 421–435.

23. Niwa T, Ying BW, Saito K, Jin WZ, Takada S, et al. (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins. Proc Natl Acad Sci 106: 4201–4206.

24. Fernandez-Escamilla AM, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nature Biotech 22: 1302–1306.

25. Yang JR, Liao BY, Zhuang SM, Zhang J (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci 109: 831–840.

26. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, et al. (2007) Aggrescan: a server for prediction and evaluation of "hot spots" of aggregation in polypeptides. Bioinfo 8: 65.

27. Khersonsky O, Roodveldt C, Tawfik DS (2006) Enzyme promiscuity : evolutionary and mechanistic aspects. Curr Opin Chem Biol 10: 498–508.

28. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. Bioinfo 2: 288–289.

29. Bershtein S, Mu W, Shakhnovich E (2012) Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. Proc Natl Acad Sci 109: 4857–4862.

30. Ward J, Sodhi J, McGuffin L, Buxton B, Jones D (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. Journal of molecular biology 337: 635–645.

31. Dyson H, Wright P (2002) Coupling of folding and binding for unstructured proteins. Current opinion in structural biology 12: 54.

32. Ignatova Z, Gierasch LM (2004) Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling. Proc Natl Acad Sci 101: 523–528.

33. Guo M, Xu Y, Gruebele M (2012) Temperature dependence of protein folding kinetics in living cells. Proc Natl Acad Sci 109: 1–5.

34. Stark C, Breitkreutz B, Reuly T, Boucher L, Breitkreutz A, et al. (2008) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: 535–539.

35. Maslov S, Ispolatov I (2007) Propagation of large concentration changes in reversible proteinbinding networks. Proc Natl Acad Sci 104: 13655–13660.

36. Startk C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The biogrid interaction database: 2011 update. Nucleic Acids Res 39: 698–704.

37. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, et al. (2003) Global analysis of protein localization in budding yeast. Nature 425: 685–691.

38. Ghaemmaghami S, Huh W, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in budding yeast. Nature 425: 737–741.

39. Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. Curr Opin Struc Biol 19: 596–604.

40. Ghosh K, Dill KA (2009) Computing protein stabilities from their chain lengths. Proc Natl Acad Sci 106: 10649–10654.

41. Dill KA, Ghosh K, Schmit JD (2011) Physical limits of cells and proteomes. Proc Natl Acad Sci 108: 17876–17882.

42. Ghosh K, Dill KA (2010) Cellular proteomes have broad distributions of protein stability. Biophys J 99: 3996–4002.

43. Chen P, Shakhnovich EI (2010) Thermal adaptation of viruses and bacteria. Biophys J 98: 1109–1118.

44. Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI (2007) A first-principles model of early evolution: Emergence of gene families, species, and preferred protein folds. PLoS Comp Biol 3: e139.

45. Heo M, Maslov S, Shakhnovich E (2011) Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci 108: 4258–4263.

46. Chen P, Shakhnovich EI (2010) Thermal adaptation of viruses and bacteria. Biophys J 98: 1109–1118.

47. Wylie CS, Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci 108: 9916–9921.

48. Heo M, Kang L, Shakhnovich EI (2008) Emergence of species in evolutionary "simulated annealing". Proc Natl Acad Sci 106: 1869–1874.

49. Chen Y, Dokholyan NV (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. Mol Bio Evol 25: 1530–3.