Thermophilic Adaptation in Prokaryotes Is Constrained by Metabolic Costs of Proteostasis

Sergey V. Venev[1] and Konstantin B. Zeldovich*,[1]

[1]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 368 Plantation St, Worcester, MA

*Corresponding author: E-mail: konstantin.zeldovich@umassmed.edu.

Associate editor: Banu Ozkan

Abstract

Prokaryotes evolved to thrive in an extremely diverse set of habitats, and their proteomes bear signatures of environmental conditions. Although correlations between amino acid usage and environmental temperature are well-documented, understanding of the mechanisms of thermal adaptation remains incomplete. Here, we couple the energetic costs of protein folding and protein homeostasis to build a microscopic model explaining both the overall amino acid composition and its temperature trends. Low biosynthesis costs lead to low diversity of physical interactions between amino acid residues, which in turn makes proteins less stable and drives up chaperone activity to maintain appropriate levels of folded, functional proteins. Assuming that the cost of chaperone activity is proportional to the fraction of unfolded client proteins, we simulated thermal adaptation of model proteins subject to minimization of the total cost of amino acid synthesis and chaperone activity. For the first time, we predicted both the proteome-average amino acid abundances and their temperature trends simultaneously, and found strong correlations between model predictions and 402 genomes of bacteria and archaea. The energetic constraint on protein evolution is more apparent in highly expressed proteins, selected by codon adaptation index. We found that in bacteria, highly expressed proteins are similar in composition to thermophilic ones, whereas in archaea no correlation between predicted expression level and thermostability was observed. At the same time, thermal adaptations of highly expressed proteins in bacteria and archaea are nearly identical, suggesting that universal energetic constraints prevail over the phylogenetic differences between these domains of life.

Key words: **thermophiles, evolution, protein folding, homeostasis, chaperones.**

Introduction

Over the 4 billion years of evolution, life has colonized an extreme diversity of physical environments on Earth, ranging from volcanic hot vents in the oceans to permafrost to hypersaline lakes. Adaptations to these conditions allow proteins and nucleic acids to function in a wide range of physical and chemical environments, resulting in specific patterns of nucleotide and amino acid usage (Galtier and Lobry 1997; Kreil and Ouzounis 2001; Zeldovich et al. 2007b; Berezovsky et al. 2007; England et al. 2003; Fukuchi et al. 2003; Sghaier et al. 2013; Sabath et al. 2013). Although the variation of amino acid frequencies across species is relatively constrained for a given genomic composition (Krick et al. 2014; Goncearenco and Berezovsky 2014), amino acid compositions of prokaryotic proteomes are sensitive to the temperature and salinity of their natural environments (Fukuchi et al. 2003; Kreil and Ouzounis 2001). Unraveling the evolutionary origins of amino acids usage in proteomes involves two main questions: first, what are the origins of the generally similar average amino acid usage across multiple highly divergent species, and second, what biological mechanisms drive adaptation of amino acid frequencies to environmental conditions.

Correlations between nucleotide and amino acid frequencies have been revealed simultaneously with the discovery of the genetic code (Sueoka 1961; Jukes et al. 1975; King and Jukes 1969). Subsequent genome-wide studies found that genomic composition strongly affects the patterns of amino acid and codon usage at the organismal level (Kreil and Ouzounis 2001; Knight et al. 2001; Lightfield et al. 2011; Goncearenco and Berezovsky 2014). At the same time, mutational patterns cannot fully explain the genome composition (Rocha et al. 2010). Closely related species adapted to different environments demonstrate variation in amino acid usage unaccounted for by their similar genomic compositions (Singer and Hickey 2003; McDonald 2010; Haney et al. 1999; Fukuchi et al. 2003). Therefore, selection at the level of nucleotide frequencies does not fully explain the variation of amino acid composition, and multiple mechanisms of protein-level selection have been proposed. In unicellular organisms, highly abundant proteins have a biased amino acid composition to decrease the metabolic cost of amino acid biosynthesis (Akashi and Gojobori 2002; Heizer et al. 2006; Seligmann 2003; Heizer et al. 2011). In this class of models, the best explanation of observed amino acid compositions is achieved in a phenomenological approach by Krick et al. (2014), who took into account the metabolic cost of amino acids synthesis and the rates of their chemical degradation.

Thermal adaptation is a particularly well-studied example of environmental adaptation that does not reduce to

Article

changing nucleotide frequencies. Maintenance of the pool of functional, properly folded proteins at elevated temperatures imposes constraints on protein structures (Szilagyi and Zavodszky 2000; England et al. 2003), as well as amino acid compositions (Zeldovich et al. 2007b; Singer and Hickey 2003; Kreil and Ouzounis 2001; Haney et al. 1999). The temperature span of life reaches almost 120 K (from -10° to 110°), a change in energy of 0.24 kcal/mol. As this value is comparable to the average effect of a single amino acid substitution in a folded protein $\Delta\Delta G \approx 1$ kcal/mol (Zeldovich et al. 2007c) and the typical energy of inter-residue van der Waals contacts, thermophilic proteins evolved sequence and structure features to increase their stability. Thermally adapted proteins utilize positive and negative design strategies, stabilizing their native folds and destabilizing unfolded conformations (Berezovsky et al. 2007). Increased fraction of hydrophobic residues contributes to protein core stability, whereas increased fraction of the charged residues enforces native fold uniqueness by destabilizing unfolded conformations. Destabilization of nonnative states can be achieved by an increased fraction of charged residues on protein surface and formation of ionic pairs (Szilagyi and Zavodszky 2000; Zhao et al. 2011). It is known that whereas salt bridge typically stabilizes the protein, longer range ion pairs are often destabilizing (Kumar and Nussinov 2002). Microscopic models of electrostatic effects in protein stability have been extensively developed (Loladze et al. 1999; Loladze and Makhatadze 2008; Strickler et al. 2006; Karshikoff et al. 2015; Sawle and Ghosh 2015), leading to in vitro validation by redesign of electrostatic interactions in ubiquitin and several other proteins. Overall, biophysical models provide a solid understanding of atomistic-level interactions in specific proteins, and, statistically, explain well the global temperature trends of amino acid frequencies in prokaryotes (Berezovsky et al. 2007; Venev and Zeldovich 2015).

Unfortunately, even state of the art models can only explain either the overall proteomic amino acid composition (Seligmann 2003; Heizer et al. 2011; Krick et al. 2014), or its temperature trends (Berezovsky et al. 2007; Venev and Zeldovich 2015), but not both. Here, we couple protein folding and protein homeostasis costs to bridge this gap and build a microscopic model explaining both amino acid composition and its temperature trends. As it is known, chaperone-assisted folding mechanisms evolved to repair misfolded proteins (Hartl et al. 2011), and even a moderate decrease in protein foldability imposes an organismal fitness cost (Drummond and Wilke 2008; Geiler-Samerotte et al. 2011). Chaperones require energy to function, which in turn creates an additional selective pressure on protein foldability, especially in the case of highly abundant proteins (Kepp et al. 2014). As proteostasis consumes up to 80% of total metabolic rate of unicellular free-living organisms (Kepp et al. 2014), adaptation towards energy efficiency is a significant driver of evolution. In fact, while the present work was under review, the Dill group published a kinetic model of proteostasis in *Escherichia coli*, showing that dynamic sorting of client proteins between chaperone systems is energy efficient for the cell. Specifically, it was found that the "sickest" proteins (ones with a stable misfolded state kinetically accessible from unfolded state) use the most energy-intensive GroEL chaperone (Santra et al. 2017).

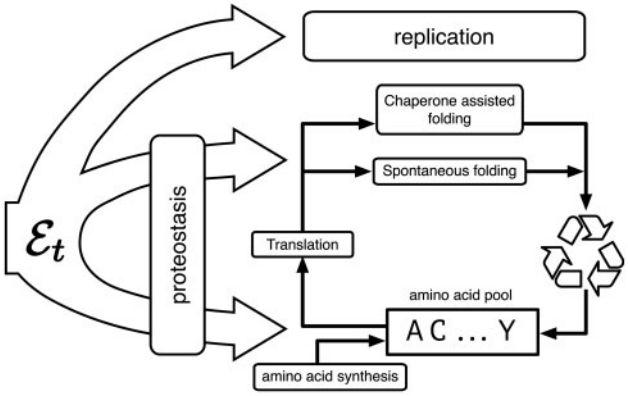

FIG. 1. Material and energy flux in proteostasis. Amino acid biosynthesis, translation and polypeptide synthesis, and chaperone assisted protein folding consume a significant fraction of energy $\mathcal{E}$ available to a prokaryote. Maintenance of steady state concentrations of every amino acid bears a known energy cost, with cheaper amino acids preferred in highly expressed proteins (Akashi and Gojobori 2002). We propose that energy cost of chaperone activity depends on amino acid composition of client proteins, as protein foldability is affected by amino acid composition (Dill 1985; Berezovsky et al. 2007; Venev and Zeldovich 2015). Therefore, amino acid composition evolves under the energetic constraint from two distinct processes, amino acid biosynthesis costs and chaperone activity.

Here, we propose that global amino acid composition evolved under the selective pressure of the total energetic cost of proteostasis. Following earlier studies, our model includes the cost of amino acid synthesis and maintenance of their constant concentrations in the presence of chemical degradation (fig. 1). The key new feature, however, comes from considering the energy cost of chaperone assisted protein folding. Protein stability against thermal unfolding depends on the amino acid composition, and amino acid compositions delivering highly foldable proteins require lower energy expenditures on repairing misfolded proteins by chaperones. As detailed below, minimization of the total energy spent on amino acid synthesis and maintenance of folded proteins by chaperones provides an accurate description of both average amino acid frequencies, and their trends with environmental temperature.

Results

Thermal Adaptation in Highly Expressed Proteins Is Similar in Bacteria and Archaea

Although archaea and bacteria have diverged early on during evolution, today they share many of the same environments, with both domains spanning wide temperature ranges. Thermal adaptations in the two domains provide a unique test case for comparing phylogenetically divergent responses to the same physical environment. To quantify thermal adaptation, we performed linear regressions between the frequencies f_a of each of the 20 amino acids in the archaeal and bacterial proteomes, and optimum growth temperature

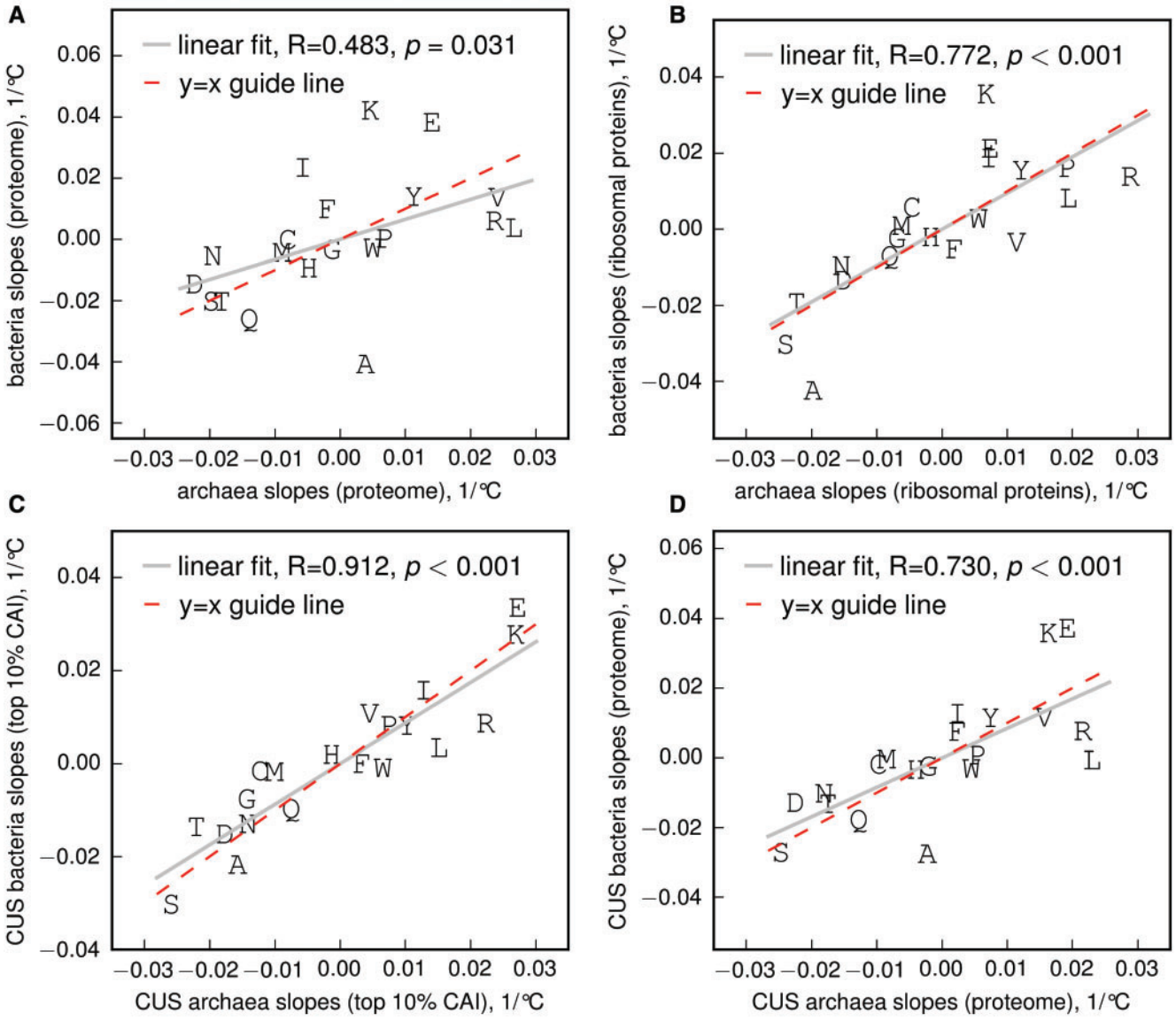

Fig. 2. Convergence of the archaeal and bacterial trends of thermal adaptation. Slopes of the amino acid frequencies versus OGT regressions are compared between archaea and bacteria. (*A*) Proteome-wide, the temperature trends of amino acid usage in bacteria and archaea are weakly correlated. (*B*) Ribosomal proteins of archaea and bacteria have similar patterns of thermal adaptation. (*C*) Predicted highly expressed proteins (top 10% CAI) in the organisms with CUS show identical patterns of thermal adaptation between bacteria and archaea. (*D*) Correlation of trends of thermal adaptation in complete proteomes of organisms with CUS is statistically insignificant in the organism bootstrap test (see text for details).

(OGT), and used the slopes of the regression df_a/dT as metric of adaptation (supplementary fig. S1A and *B*, Supplementary Material online). Amino acids with positive slopes are statistically overrepresented in thermophilic proteomes, whereas negative slopes reflect reduced usage of an amino acid in thermophiles.

The correlation between the temperature trends of amino acid frequencies in complete proteomes of bacteria and archaea is not very high, $R = 0.48$ (fig. 2A). Moreover, bacterial slopes are generally lower than archaeal ones. Therefore, phylogenetic divergence and ensuing biochemical differences had a profound effect on proteome-averaged amino acid usage in the two prokaryotic domains.

We hypothesized that for highly expressed proteins, energetic constraints on thermal adaptation may prevail over phylogenetic differences, leading to a greater similarity of archaea and bacteria. Highly expressed proteins are known to evolve slowly (Pál et al. 2001; Rocha and Danchin 2004), suggesting a stronger evolutionary constraint, which is partially reflected in more stringent folding requirements (Serohijos et al. 2012; Drummond et al. 2005; Drummond and Wilke 2008).

Ribosomal proteins serve as a particularly well-defined group of highly expressed proteins in both archaea and bacteria (Karlin et al. 2005). At the same time, differences in ribosome structures and sequences between the two domains are significant. Both domains of life exhibit very similar patterns in thermal adaptation of ribosomal proteins (fig. 2B), $R = 0.77$ (bootstrap to find a similar correlation in the same number of randomly selected proteins yields

$P < 0.001$ (supplementary fig. S2, Supplementary Material online). However, the specific function of ribosomal proteins may have limited their options for thermal adaptation. To identify other types of highly expressed proteins we used a sequence based approach using codon adaptation index (CAI; Sharp and Li 1987) in organisms with apparent codons usage selection (CUS), see Materials and Methods for details. Remarkably, for predicted highly expressed proteins from organisms with CUS, the trends in thermal adaptation are nearly identical. Figure 2C demonstrates $R = 0.91$, $P < 0.001$ for all proteins within the top 10% of CAI; excluding ribosomal proteins yields the same $R = 0.912$ (data not shown). The null hypothesis that the observed correlation can be explained by a random choice of a subset of organisms or proteins is safely rejected ($P = 0.015$, randomized CUS assignment, $P < 0.001$, randomized CAI ranking (supplementary fig. S3, Supplementary Material online). A greater similarity between thermal adaptations in highly expressed proteins from archaea and bacteria compared with the whole-proteome case comes mostly from the differences in usage of isoleucine, alanine, lysine, and glutamic acid (A, I, K, E), as shown in figure 2A and C and supplementary figure S1C and D, Supplementary Material online.

At the same time, CUS alone does not imply similarity in thermal adaptation for bacteria and archaea. Trends in thermal adaptation in complete proteomes of bacteria and archaea with CUS (fig. 2D), $R = 0.73$, appeared more similar than for complete proteomes of all species (fig. 2A). However, correlation of $R \geq 0.73$ could be achieved with probability 0.153 in a randomized selection of the same number of bacteria and archaea from the full data set (supplementary fig. S4, Supplementary Material online). Therefore, increased proteome-wide similarity of thermal adaptation between bacteria and archaea with CUS is not statistically significant.

This genomewide analysis clearly shows that highly expressed proteins in both archaea and bacteria share a common strategy of thermal adaptation, which becomes obscured at the level of complete proteomes. We propose that the common strategy may involve optimization of energetic costs of proteostasis by balancing amino acid metabolism and chaperone energy expenses, and present the results of the modeling below.

Simulated Amino Acid Frequencies Respond to Chaperone Energy Costs

We designed lattice model proteins of 64 residues each in a wide range of artificial temperatures $0.4 \leq T \leq 1.7$ in units of Miyazawa–Jernigan residue level potential (p.u.; Miyazawa and Jernigan 1999), following equations (8–10) (see Materials and Methods for details). The chaperone-adjusted synthesis cost w (see eq. 7), was varied from $w = 0$ to $w = 0.15$. The case of $w = 0$ corresponds to zero cost of maintaining the amino acid pool, so the energetic cost is completely defined by protein foldability. On the contrary, for large values of w, the energetic costs of amino acid maintenance prevail over energy expenditures by chaperones, reducing the effect of protein foldability on fitness. Proteins designed with no synthesis cost constraint, $w = 0$, mostly reproduced earlier results (Berezovsky et al. 2007; Venev and Zeldovich 2015). At low simulated temperatures, the folding constraint on protein sequences was weak. Starting from a random sequence with $\approx 1/20$ amino acid abundances, the design procedure was able to create well-folding sequences by swapping the residues while retaining the overall amino acid composition (supplementary fig. S5A, Supplementary Material online). As the temperature increased, relative amino acid abundances changed monotonically to allow designed proteins to increase their thermal stability (supplementary fig. S5A, inset, Supplementary Material online). Increasing frequencies of hydrophobic and charged residues extend the energy gap by decreasing the energy of the native state and raising the average decoy energy (Berezovsky et al. 2007).

The outcome of protein design changed significantly if the chaperone-adjusted synthesis cost w was introduced (supplementary fig. S5B, Supplementary Material online). In this case, frequent usage of "expensive" amino acids carried a significant penalty even if they were favorable for protein foldability. At $w = 0.05$, proteome-averaged frequencies of amino acids already diverged at low temperatures T, and the distribution of amino acid frequencies was largely determined by their relative metabolic maintenance costs, due to the smaller selective pressure on foldability. We then hypothesized that this interplay between the costs of raw materials (amino acid pool) and maintenance of product (chaperone-assisted folding) can explain both the average amino acid composition and its temperature trends.

Simulated Trends Correlate with Biological Data

The amino acid frequencies produced by our model are controlled by two parameters, temperature T and the chaperone-adjusted synthesis cost w. To assess the fit of the model to observed frequencies for given values of w and T, we used the Jensen–Shannon divergence (JSD) between the frequency distributions of the 20 types of amino acid in the simulated and biological data. The JSD between two probability (frequency) distributions is zero if the distributions are identical, and equals one if the two distributions are completely unrelated. Using the Pearson correlation coefficient between amino acid frequencies as the measure of similarity of the distributions produced qualitatively similar results (data not shown). Prokaryotic genomes were separated into mesophilic ($20 \leq \mathrm{OGT} \leq 50°$) and thermophilic ($\mathrm{OGT} > 50°$) groups and average amino acid frequencies from these groups were used for comparison with the simulated data. This analysis has been performed separately for bacteria and archaea. In bacteria, the JSD between simulated and real data reaches minima at specific values of T and w (fig. 3). Our model correctly segregated thermophilic and mesophilic genomes. The value of JSD_M reached its minimum at $T_M = 0.7$ p.u. whereas JSD_T reached the minimum at a higher temperature $T_T = 0.9$ (fig. 3A and B). Both JSD_M and JSD_T reached their minima at the same value of chaperone-adjusted synthesis cost $w^* = 0.05$, so the energetic balance between chaperone activity and costs of amino acid maintenance appeared similar between thermophiles and mesophiles. Remarkably, the value of w^* was the same for archaea and bacteria; this finding

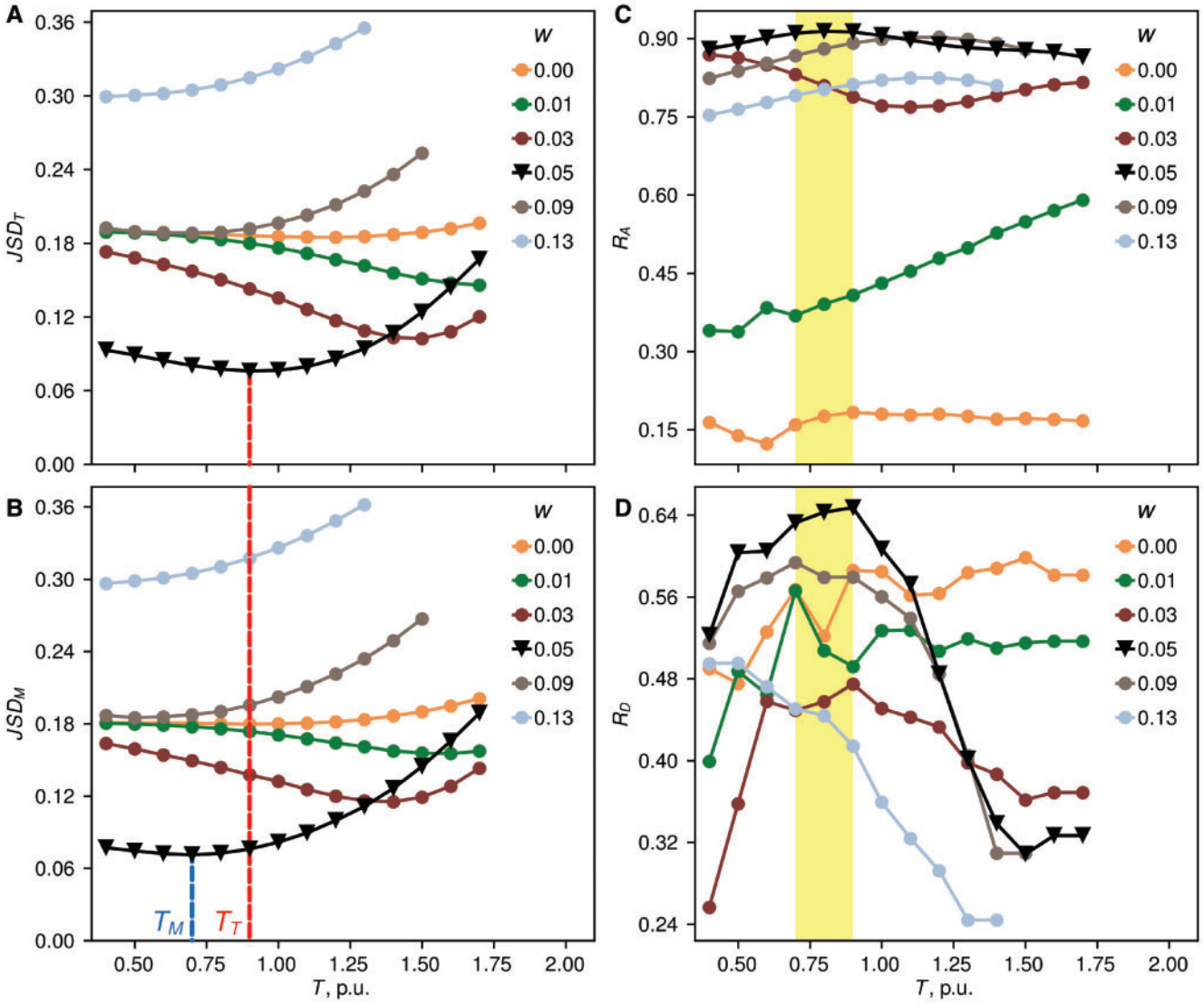

FIG. 3. Simulated frequencies of amino acids compared with the naturally evolved ones for bacteria. (*A*, *B*) Jensen–Shannon divergence between amino acid frequencies in simulated data and thermophilic (*A*) and mesophilic (*B*) proteomes exhibits clear minima with respect to the temperature T and shaperone-adjusted synthesis costs w. The best match between the model and mesophilic proteomes is achieved at a lower temperature than the best match to the thermophilic ones. (*C*) Pearson correlation coefficient R_A between amino acid frequencies in simulated data and all bacterial proteomes reaches $R \sim 0.9$ for $0.7 < T < 0.9$. (*D*) In the same temperature range, the temperature trends of amino acid in simulated data are strongly correlated with those in bacterial proteomes, $R = 0.64$.

was robust upon changes of proteostasis cost cutoff Π^* (supplementary table S1, Supplementary Material online).

The temperature range $[T_M, T_T]$ and the cost w^* successfully describe the complete data set of both mesophiles and thermophiles in terms of amino acid composition and its temperature trends. To simplify comparison with earlier works, figure 3C presents the Pearson correlation coefficient $R_A = R(f \rightarrow_{\text{model}}, f \rightarrow_{\text{bio}})$ between average amino acid frequencies in 262 bacteria (mesophiles and thermophiles combined), and simulated data. The very high correlation, $R_A \approx 0.9$ is similar to the predictions of the current-best phenomenological model (Krick et al. 2014). Figure 3D shows the correlation R_D between the amino acid temperature trends (slopes df_a/dT) in the model and biological data, $R_D = R(df \rightarrow_{\text{model}}/dT, df \rightarrow_{\text{bio}}/dT)$, in the same 262 bacterial species. Difference quotients of simulated amino acid frequencies $\Delta f_a/\Delta T, a = 1 \ldots 20$ in the $[T_M, T_T]$ range were used to calculate the simulated temperature trends. For the real frequencies of amino acids, the slopes were derived from the linear regression over the entire OGT range (supplementary fig. S1A–D, Supplementary Material online). Similar to the values of R_A, the value of R_D exhibits a clear maximum with respect to both w and T, reaching $R_D \approx 0.60$, similar to earlier findings (Venev and Zeldovich 2015).

Interestingly, the relative temperature range in the model, $(T_T - T_M)/T_M \approx 29\%$ compared well with the actual temperature range of prokaryotes, thriving between ~280 and 370 K, a 30% change in absolute temperature. Comparison between simulated data and amino acid frequencies in archaea is presented in supplementary figure S6, Supplementary

Material online, and shows similar values of the optimum temperature range and chaperone-adjusted synthesis cost w. Qualitatively similar results are obtained by comparing the simulation with amino acid frequencies derived from predicted highly expressed proteins (top 10% of CAI for organisms with CUS; supplementary figs. S7 and S8, Supplementary Material online).

To prove that these results are not a numerical artifact, we have shuffled the values of amino acid maintenance cost C_a and repeated the simulations. The reshuffling breaks the connection between the biochemistry of amino acids, reflected in their costs C_a, and their physical properties, such as interaction energies ϵ_{ab}. As shown in supplementary figure S9, Supplementary Material online, the values of JSD_A and R_D obtained from reshuffled C_a are almost always lower than those for the initial true C_a, yielding $p \leq 0.01$. Therefore, we demonstrated that the interplay between amino acid synthesis costs and protein folding is internally consistent, and our model produces correct amino acid frequencies only when the connection between the organism-level biochemical and physical properties of amino acids is preserved.

The predicted temperature trends of amino acid frequencies had a statistically significant correlation with biological data for bacteria, $R = 0.64$ (supplementary fig. S10B, Supplementary Material online). A much weaker correlation was observed for archaea, $R = 0.35$ (supplementary fig. S10A, Supplementary Material online). We have then considered only highly expressed proteins (top 10% of CAI for organisms with CUS) from either domain, expecting that the selective pressure of proteostasis is stronger for this group of proteins. In bacteria, consideration of highly expressed proteins did not significantly affect the agreement between simulated and biological data, $R = 0.55$ (supplementary fig. S10D, Supplementary Material online). However, temperature trends in highly expressed proteins in archaea were strongly correlated with our simulation, $R = 0.61$ (supplementary fig. S10C, Supplementary Material online). Therefore, in archaea, highly expressed proteins experience a selective pressure that is well-described by the model, while the complete archaeal proteomes apparently evolved under a variety of constraints yet to be identified.

Results of the simulations were generally robust with the respect to changes of the model parameters, such as proteostasis cost cutoff Π^*, equation (10) as shown in supplementary table S1, Supplementary Material online. The correlation coefficient R_D between the simulated and biological temperature derivatives of amino acid frequencies was not sensitive to Π^*. However, at low values of Π^*, the selection was not strong enough, leading to the optimum parameter w being different for thermophiles and mesophiles in certain cases. This artifact vanished at $\Pi^* \geq 0.7$, the value ultimately chosen for production simulations.

Consistent with previous findings (Venev and Zeldovich 2015), the temperature trend of leucine frequency was not captured well by the model. Leucine is a very hydrophobic residue, as reflected by the Miyazawa and Jernigan interaction potential. Accordingly, the frequency of leucine rapidly increased with temperature in simulated proteomes, as leucine participates in hydrophobic interactions in the protein core. As it is known (Goncearenco et al. 2014), the frequency of leucine does not increase with temperature in bacteria, although it does so in archaea (supplementary fig. S1A and B, Supplementary Material online). Combined with the fact that leucine is relatively simple to synthesize, and is coded by six different codons, these observations clearly point to the biochemical differences between archaea and bacteria, and to the limitations of current biophysical models. Our choice of the Miyazawa and Jernigan amino acid interaction potential, which overemphasizes attractive forces, may be one of the factors contributing to the leucine being an outlier. Alternative derivations of the knowledge based potential, such as (Thomas and Dill 1996), may alleviate this issue. Aspartic acid is an another known outlier (Goncearenco et al. 2014). This charged amino acid is predicted to increase in frequency as the temperature rises, just as glutamic acid, lysine, and arginine. However, while glutamic acid and lysine consistently increase in frequency in both bacteria and archaea, aspartic acid is surprisingly depleted in natural thermophilic proteomes.

Amino Acid Synthesis Costs Increase with Environmental Temperature

Metabolic costs of amino acid synthesis are negatively correlated with protein expression levels across the three domains of life (Akashi and Gojobori 2002; Swire 2007). In figure 4A and C, we plotted the proteome-averaged Akashi–Gojobori amino acid synthesis cost against environmental temperature for 140 archaea and 262 bacteria, assuming equal expression levels of all proteins. We found a statistically significant positive correlation, confirming that thermal stability requires heavier usage of synthetically "expensive" proteins, in agreement with an earlier observation made on *Thermus thermophilus* genome (Swire 2007). In contrast with the amino acid synthesis cost, the amino acid maintenance cost, which combines synthesis and decay (Krick et al. 2014), is not significantly correlated with the environmental temperature (OGT; fig. 4B and D). These observations are generally consistent with our simulations. Supplementary figure S11A–H, Supplementary Material online, demonstrates that for the optimum temperature range $[T_M, T_T]$ and chaperone-adjusted synthesis cost $w^* = 0.05$, amino acid synthesis costs increased with temperature, whereas amino acid maintenance costs weakly decreased, similar to figure 4.

Highly Expressed Proteins Are Similar to Thermophilic Ones in Bacteria, Not in Archaea

In full agreement with earlier findings by Akashi and Gojobori (2002) and Swire (2007), our analysis found a statistically significant negative correlation between the amino acids synthesis costs and CAI (proxy for expression) in sets of 167 bacterial and 65 archaeal genomes (supplementary fig. S12, Supplementary Material online). At the same time, it has been proposed that amino acid composition of highly expressed proteins is similar to the composition of thermophilic proteins (Cherry 2010). As noticed earlier (Serohijos et al. 2012), this is somewhat contradictory to the findings

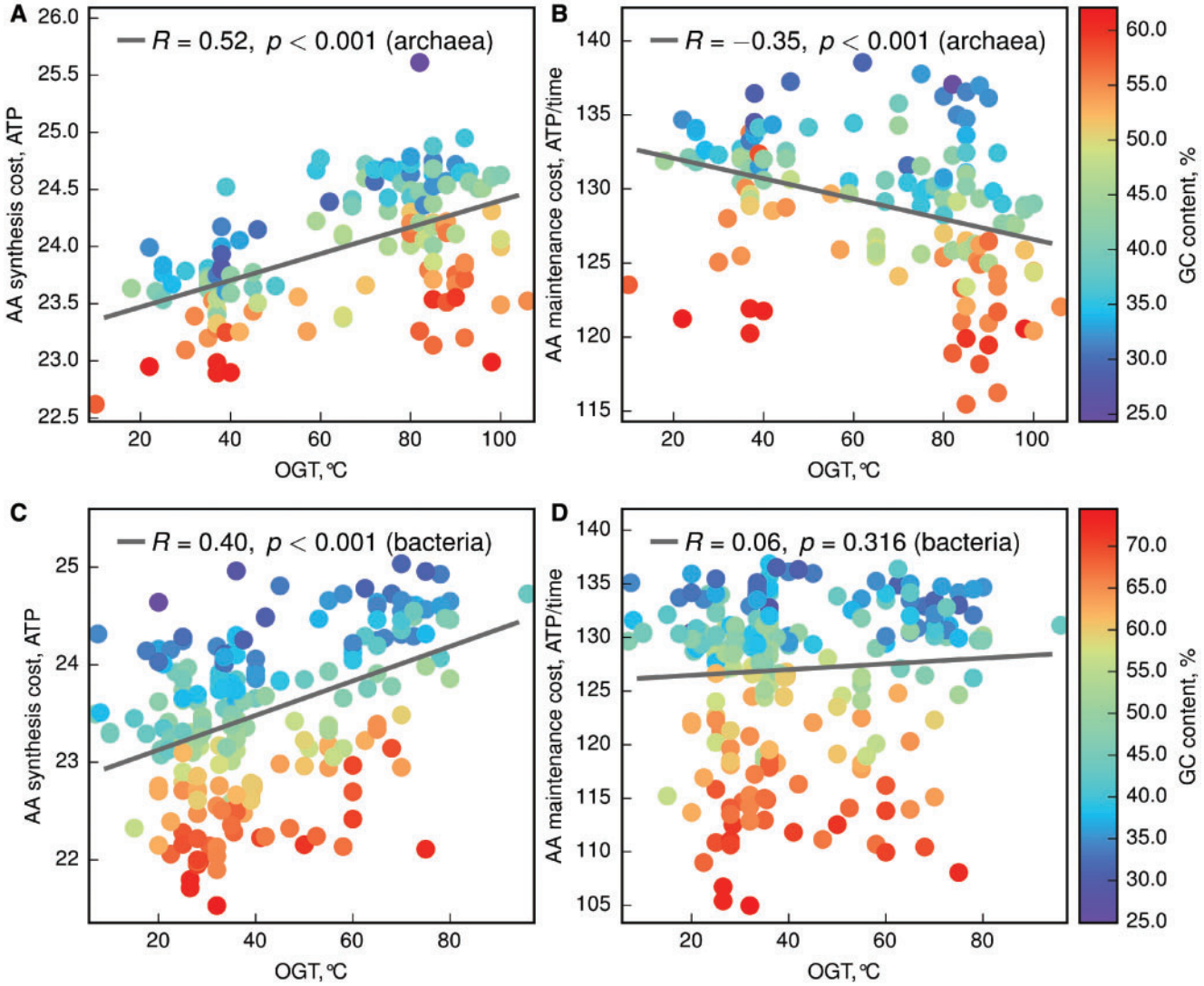

FIG. 4. Temperature trends of the amino acid synthesis and maintenance costs in prokaryotes. Proteome average cost of amino acid synthesis and amino acid maintenance for archaeal species (*A*, *B*) and bacterial species (*C*, *D*). Marker color represents the genome-wide GC content of each specie; as it is well-established, genome-wide GC content is not correlated with OGT, see also supplementary figure S1*A* and *B*, Supplementary Material online.

of Akashi et al and Swire, as thermophilic proteins tend to be more synthetically "expensive".

To look for the origins of this controversy, we compared the protein expression levels, approximated by CAI, with their thermostability in bacteria and archaea. To assess protein thermostability of a group of proteins, we used the JSD_T between their average amino acid composition and the average amino acid composition of 92 thermophilic archaea or 66 thermophilic bacteria (OGT $> 50^{\circ}$). For each organism, proteins were split into twenty bins according to their CAI, and average amino acid usage of each bin was compared with the thermophilic composition using JSD_T. For archaea, we did not observe a significant correlation between CAI and JSD_T (fig. 5A). However, bacteria exhibited a statistically significant negative correlation between JSD_T and CAI bin (fig. 5B): bacterial proteins appeared more similar to thermophilic ones (lower values of JSD) as their CAI increased. As a control, we have reshuffled synonymous codons within each genome, destroying the codon bias and thus CAI metric of each protein, but leaving amino acid composition intact. No correlation was observed in reshuffled data for bacteria (fig. 5B). To rule out the possible effects of binning on the observed correlations, we repeated the analysis using 5 and 50 bins of CAI values, and observed the same trends (data not shown). These results partially support Cherry's findings and demonstrate the immense flexibility of the 20-dimensional space of protein sequence composition to satisfy multiple physical and phylogenetic constraints.

We have also tried to approximate protein thermostability by the fraction of IVYWREL amino acids, which is strongly correlated with OGT at the proteomic level (Zeldovich et al. 2007b). Although a strong negative correlation between IVYWREL and CAI was observed, the same trend persisted upon synonymous codon reshuffling, in both bacteria and archaea (supplementary fig. S13, Supplementary Material online). Therefore, an intrinsic connection between the amino

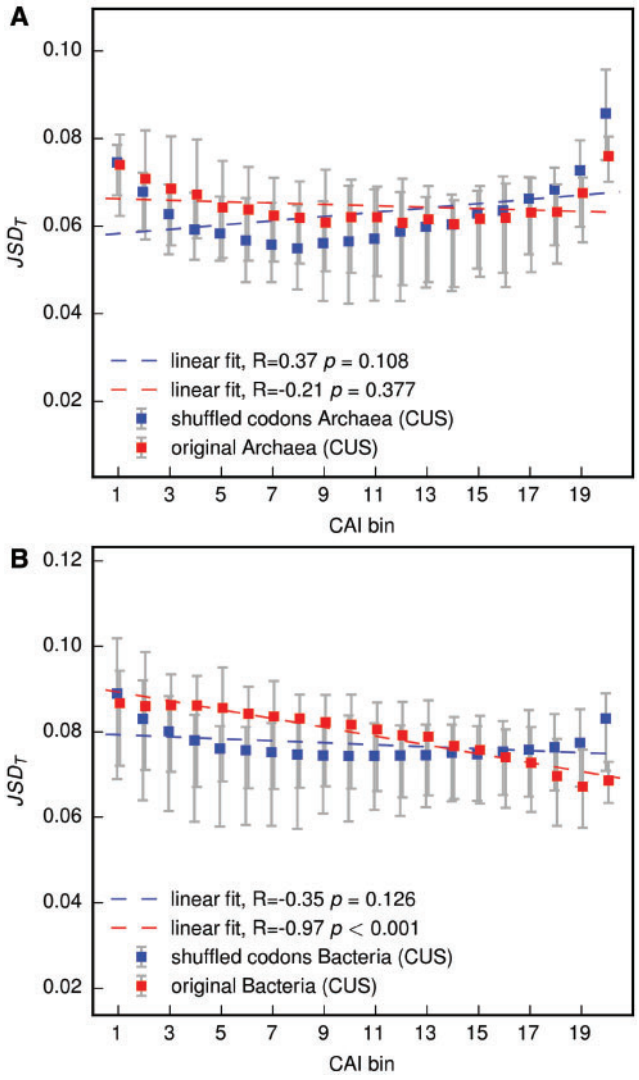

FIG. 5. Similarity between proteins with high CAI and thermophilic proteins by comparing amino acid composition. Proteins in each organism are grouped into 20 bins according to CAI, and amino acid composition within each groups is compared with the average thermophilic composition using Jensen–Shannon divergence, separately for archaea (*A*), and bacteria (*B*). In bacteria, the higher is protein expression, the more similar is amino acid composition to a thermophilic one (decreasing JSD_T, red line, statistically significant). No such trend was observed for archaea. As a control, codon reshuffling was used to destroy the relation between CAI and amino acid composition of proteins. For both archaea and bacteria, the correlation between JSD_T and CAI for reshuffled codons was not significant. Error bars represent the 30% and 70% percentiles of the underlying distributions.

acid and nucleotide frequencies via the genetic code prevents the use of IVYWREL metric to compare protein expression (CAI) and thermostability.

To check if our simulation can capture an increased similarity between highly expressed and thermophilic proteins, we have divided the model proteins into three bins of low, medium, and high abundance, positing that the metabolic cost is proportional to the total amount of protein in each group, equation (11) (see Materials and Methods for details). We have run the simulation for $w^* = 0.05$, corresponding to the best global fit of model with experimental data, and found that highly abundant model proteins were indeed consistently closer in amino acid composition to natural thermophilic bacteria, compared with low-abundant model protein (supplementary fig. S14, Supplementary Material online). Although this finding is encouraging, the effect is relatively small, $R_T = 0.82$ for low abundance versus $R_T = 0.88$ for high abundance. Further study of this phenomenon warrants development of a more detailed cellular fitness model, as in our approach the contribution of proteins to fitness was exclusively defined by stability, metabolic cost and abundance, without considering essentiality or connectivity in the cell's metabolic network.

Discussion

Statistically significant correlations between environmental conditions and amino acid usage are well-established, dating back at least to 1982, when amino acid usage was quantitatively linked to environmental temperature by Ponnuswamy et al. (1982). An early physical model of amino acid composition has been proposed by Dill (1985), who derived the ratio of hydrophobic to polar residues conferring highest stability to a globular protein. The interest in the statistical understanding of thermal adaptation increased as microscopic simulations of protein evolution became possible (Taverna and Goldstein 2002; Bloom et al. 2006; Goldstein 2008). Modeling of lattice proteins (Berezovsky et al. 2007; Venev and Zeldovich 2015) showed that although the temperature trends in amino acid frequencies can be explained by purely physical models, the frequencies themselves are weakly correlated with genomic data. This discrepancy suggests that either the physical models are still not precise enough to resolve individual amino acids beyond their rough classification by hydrophobicity, or other factors contribute significantly to amino acid usage.

Complementary to protein folding constraints, metabolic costs and overall energy balance of a cell have been long identified as powerful evolutionary drivers (Pál et al. 2006), as exemplified, for example, by the success of quantitative flux based metabolic models (Varma and Palsson 1994; Price et al. 2004). Akashi and Gojobori (2002) estimated the energy expended on the synthesis of each of the 20 types of amino acid molecules, and found that highly expressed proteins are enriched in "cheap", easily synthesized amino acids. These findings highlighted the importance of proteostasis as the major cellular process, coupling energy and material fluxes in a cell. The flux models were further advanced by an estimate of the amino acid decay rates within a cell (Krick et al. 2014). By combining the amino acid synthesis cost, decay rate, and sequence entropy into an empiric cost function, Krick et al. (2014) made successful predictions of amino acid frequencies. However, this model did not explicitly address protein folding or other physical considerations, and so is difficult to extend to the study of thermal adaptation.

To bridge this gap, we proposed that proteostasis is not limited to the chemical turnover of amino acid molecules,

but, crucially, maintains appropriate levels of functional, correctly folded proteins. Molecular chaperones are an integral part of this process, attempting to refold proteins in an ATP-dependent manner. Invoking quality control systems in response to misfolded proteins causes a fitness penalty proportional to the fraction of misfolded proteins, their expression level and is largely function-independent (Geiler-Samerotte et al. 2011). Moreover, further experiments suggested that it is indeed the metabolic cost of chaperone activity that imposes the fitness penalty, rather than the consequences, for example, toxicity, of the presence of abundant misfolded proteins (Tomala et al. 2014). Chaperone function provides a feedback to the genotype, by accelerating its evolution while serving as a capacitor for otherwise deleterious phenotypic mutations (Bogumil and Dagan 2012; Çetinbaş et al. 2013).

Following Kepp et al. (2014) and in parallel with (Santra et al. 2017), we hypothesized that the energy consumed by chaperones is nonnegligible and must be taken into account together with other metabolic costs. Specifically, we assumed that the total energy cost of proteostasis includes contributions from both amino acid turnover and chaperone activity. The key feature of the model is the statistical dependence between foldability of a protein and its amino acid composition (Dill 1985; Berezovsky et al. 2007; Venev and Zeldovich 2015). Indeed, well-folded proteins typically contain a balanced mix of charged and hydrophobic residues, while intrinsically unfolded proteins do not (Uversky et al. 2000). Statistically, proteins with an unbalanced amino acid composition are less stable and so may require more frequent chaperone intervention. Therefore, we posited that amino acid compositions have evolved to minimize the total energy spent on amino acid homeostasis and chaperone activity, and tested this hypothesis by simulations.

By incorporating protein folding and metabolic cost in a single model, we were able to capture average amino acid composition and its temperature trends simultaneously (fig. 3), significantly improving upon purely physical models (Berezovsky et al. 2007; Venev and Zeldovich 2015). These models are captured in our study as a limiting case $w = 0$. As demonstrated in figure 3, the predictive power of the model dramatically increases by considering a balance between protein folding requirement and the metabolic cost constraints, $w \neq 0$. It has been long posited that stringent protein folding requirements are associated with increased fitness (Taverna and Goldstein 2002; Bloom et al. 2006; Zeldovich et al. 2007a; Lobkovsky et al. 2010). More or less explicitly, these works assumed that fitness is related to metabolic rates which in turn depend on the amount of folded, functional enzymes to catalyze chemical reactions. This assumption led to various models where fitness was positively correlated with stability, as in equation (8). In the present model, however, the positive contribution of stability to fitness does not stem from the metabolic rate argument. Rather, it emerges from the energy balance of the cell: more stable proteins require less energy expense in the chaperone system, leaving more ATP and other resources for replication. Therefore, our model suggests a novel and independent biological mechanism leading to the fitness being an increased function of protein stability. Further experimental studies will be needed to elucidate the relative contributions of metabolic versus energetic correlates of protein stability to organism fitness.

While comparing the model findings with experimental data, we have made several novel statistical observations. First, we showed that the trends of thermophilic adaptation of highly expressed proteins are very similar in archaea and bacteria, while no strong correlation is observed at the whole-proteome level (fig. 2). We interpret this finding as manifestation of convergent response to a selective pressure acting on highly expressed proteins irrespective of their phylogenetic history, and suggest the energetics of proteostasis as mechanistic explanation. Second, we clearly demonstrate that proteome-wide amino acid synthesis cost, according to Akashi–Gojobori scale, increases with OGT in both archaea and bacteria (fig. 4). This observation supports our hypothesis of synthetically expensive amino acids being crucial for protein stability and thermal adaptation. At the same time, it is well-established that highly expressed proteins are "cheap" to synthesize (Akashi and Gojobori 2002; Seligmann 2003; Heizer et al. 2006; Raiford et al. 2008 and supplementary fig. S12, Supplementary Material online). Therefore, we find that thermophilic proteins are expensive to synthesize but highly expressed ones are biosynthetically cheap. At the same time, it has been suggested that amino acid composition of highly expressed proteins is similar to that of thermophilic proteins (Cherry 2010), which creates a logical inconsistency. We attempted to address this issue by estimating the expression levels using CAI and correlating it with various composition-based predictors of thermostability in a large set of bacterial and archaeal proteomes. In the bacterial data set, we observed that highly expressed proteins had amino acid compositions more similar to the average composition of thermophilic proteomes. This finding parallels earlier results by Cherry (2010). However, no significant correlation was found in archaea (fig. 5). Apparent inconsistencies in the empirically observed cost-expression-stability triangle require further study, and suggest a surprising flexibility of amino acid usage evolving to satisfy different constraints. Observed statistical differences between archaea and bacteria in the cost-expression-stability space may complicate comparisons of evolutionary simulations (Drummond et al. 2005; Serohijos et al. 2012) with experimental data. Further development of high-throughput experimental methods for characterizing protein expression levels and thermostability, such as limited proteolysis and mass spectrometry, LiP-MS (Leuenberger et al. 2017) will make it possible to transition away from sequence-based predictors, and will stimulate the next generation of predictive, organism-level models of metabolism and selection.

Materials and Methods

Model of Protein Homeostasis and Simulation of Adapted Proteomes

Our model for protein homeostasis costs closely follows Kepp et al. (2014), together with the hypothesis that a specific

amino acid composition generates an additional, folding-related energy demand due to chaperone activity required to refold misfolded or unstable proteins. As in Kepp et al. (2014), we assume that the cellular fitness increases with the amount of energy spent on replication. As the total energy supply is limited, reduction of the energetic costs of proteostasis confers fitness advantage on the cell.

The energy expenditures of proteostasis, figure 1, can be approximated by

$$\mathcal{E}_p \approx \mathcal{E}_s + \mathcal{E}_f + \mathcal{E}_d, \quad (1)$$

where subscripts s, f, d denote synthesis, chaperone-assisted folding and degradation respectively. We assume that protein synthesis cost $\mathcal{E}_s$ depends on the protein sequence only via amino acid composition and sequence length L. Thus, the total protein synthesis cost $\mathcal{E}_s$ can be approximated by

$$\mathcal{E}_s \approx \sigma L + \sum_{a=1}^{20} C_a n_a, \quad (2)$$

where the first term is the cost of translating L codons (σ per each), and the second term represents the total energy of synthesizing n_a amino acids of each kind in a protein, $\sum_{a=1}^{20} n_a = L$. The vector $C_a, a = 1, \ldots, 20$ in equation (2) is the amount of energy required per unit time to maintain a constant concentration of each type of amino acid monomers, as they are consumed by protein synthesis and also being chemically degraded at different rates, as derived in Krick et al. (2014).

Maintenance of constant concentration of folded, functional proteins involves the action of chaperones, which help refold improperly folded proteins. Chaperone-assisted refolding consumes energy primarily on conformational transitions required to form the hydrophobic cavity (Hartl et al. 2011). We assume that the cost of chaperone activity is proportional to the fraction of unfolded client proteins, which in turn depends on the amino acid composition of the client. Denoting the fraction of chaperone clients in the native state as P_{nat}, one can express the energy costs of chaperone activity and protein degradation as

$$\mathcal{E}_f + \mathcal{E}_d \approx (F + D_F) \cdot P_{nat} + (U + D_U) \cdot (1 - P_{nat}), \quad (3)$$

where F is the energy spent per unit time to assist successful folding of the P_{nat} fraction of natively-folded protein, U is energy consumed by chaperones to refold $1 - P_{nat}$ fraction of client proteins that fail to fold spontaneously, and D_F, D_U are proteasome energy expenditures of degrading natively folded and nonnatively folded proteins, respectively. We assume that $F + D_F < U + D_U$, that is, maintenance of poorly folding proteins is costlier than maintenance of the well-folding ones. This assumption is supported by the evidence of dosage-dependent fitness penalty induced by misfolding mutations in a protein unrelated to cellular metabolism (Geiler-Samerotte et al. 2011), and by recent modeling of *E. coli* chaperone network (Santra et al. 2017). As we describe proteostasis as steady state phenomenon, temporal effects such as differences in protein and amino acid molecule lifetime, kinetics of protein folding and refolding can be neglected.

To model folding of chaperone client proteins, we used a lattice model (Shakhnovich and Gutin 1990; Sikosek and Chan 2014) of compact polymers on a $4 \times 4 \times 4$ cubic lattice, with a randomly generated subset of $N = 10^4$ conformations. Choosing a different set of 10^4 conformations did not affect the results of the simulation. A residue level knowledge-based potential (Miyazawa and Jernigan 1999) ϵ_{ab}, $a, b = 1, \ldots, 20$ was used to calculate energy of nonlocal contacts in each conformation:

$$E_{i,S} = \sum_{k,l=1}^{L} \epsilon_{S_k S_l} \delta^i_{kl}, \quad (4)$$

where $i = 1, \ldots, N$ is the index of conformation, S_k is the type of the amino acid at position k of the sequence S, and δ^i_{kl} is a contact map of the conformation i. Dependence of amino acid interaction potentials on temperature (Goldstein 2007; Pucci et al. 2014) is not considered in the model. The equilibrium fraction of natively-folded proteins P_{nat} was calculated from the Boltzmann distribution:

$$P_{nat} = \frac{\exp(-E_{nat}/k_B T)}{\sum_{i=1}^{N} \exp(-E_i/k_B T)}, \quad (5)$$

where E_{nat} is the lowest energy among all conformations and k_B is the Boltzmann constant.

Substituting equations (2, 3) into equation (1), we derive the proteostasis cost,

$$\mathcal{E}_p \approx \sigma L + \sum_{a=1}^{20} C_a n_a + (F + D_F) \cdot P_{nat} + (U + D_U) \cdot (1 - P_{nat}), \quad (6)$$

which can be rewritten as

$$\mathcal{E}_p = \alpha - \beta(P_{nat} - w \cdot \sum_{a=1}^{20} C_a n_a), \quad (7)$$

where α, $\beta > 0$ and $w > 0$ are constants, and C_a is the vector of amino acid maintenance costs. Importantly, foldability P_{nat} of chaperone clients depends on their amino acid composition, allowing for a nontrivial interplay between the two terms in equation (7). The parameter w controls the relative fitness costs of protein (mis)folding (implicit via chaperone activity) and amino acid biosynthesis in our model. The limiting case of $w = 0$ corresponds to a purely physical model where fitness is proportional to protein foldability, as in Taverna and Goldstein (2002), Bloom et al. (2006), Zeldovich et al. (2007a), Lobkovsky et al. (2010), whereas the opposite case of large w recapitulates flux based energetic models (Akashi and Gojobori 2002; Krick et al. 2014; Kepp et al. 2014). In the following, w will be referred to as chaperone-adjusted synthesis cost.

To simulate proteomes evolved to minimize the costs of protein homeostasis, we design lattice proteins using the following scoring function for sequence S at temperature T:

$$\Pi(S,T,w) = P_{\text{nat}} - w \cdot \sum_{a=1}^{20} C_a n_a. \quad (8)$$

Minimization of the proteostasis costs $\mathcal{E}_p$ is equivalent to maximizing the score $\Pi(S,T,w)$. Furthermore, for the organism to be viable, all of its proteins must be sufficiently stable ($P_{\text{nat}} > 0.5$), which leads to the additional constraint:

$$P_{\text{nat}}(S_i,T) > 0.5, \qquad i = 1 \ldots M. \quad (9)$$

The simulation introduces mutations in sequences until a sufficiently high proteostasis score is satisfied:

$$\frac{1}{M}\sum_{i=1}^{M} \Pi(S_i,T,w) > \Pi^*, \quad (10)$$

where Π^* is the threshold proteostasis score defining a viable organism. In the limiting case of $w = 0$, the Π^* is similar to the parameters previously used in Zeldovich et al. (2007a). For production simulations, $\Pi^* = 0.7$ was used, see Results and supplementary table S1, Supplementary Material online for details.

The protein design simulation starts with $M = 10^5$ random sequences. At every step, one amino acid mutation is introduced in each sequence, and a change of its score Π is calculated. Mutations that increase Π are always accepted, whereas mutations that decrease Π are accepted according to Metropolis Monte Carlo (MC) scheme with probability $P = \exp\left(-(\Pi_{\text{old}} - \Pi_{\text{new}})/T_{\text{MC}}\right)$, with the design "temperature" $T_{\text{MC}} = 10^{-4}$. The stability condition (9) is then checked, and if all sequences satisfy it, the proteostasis cost criterion (10) is evaluated. The simulation stops if the criterion (10) is met or the number of iterations exceeds 10^3.

To introduce protein abundance in the model, we assign each protein with abundance level a_i for the three groups of low, medium and high abundance ($a = 0.005, 0.1, 1$ respectively). We assign proteins to abundance groups in 25:50:25 ratio, so medium-abundance proteins represent half of the proteome, and low- and high-abundant group are one quarter of the proteome each. While this is a strong assumption, we believe that it captures the nonuniform protein abundance distribution at a coarse-grained level. Since proteostasis costs scale linearly with abundance, we can rewrite the design criterion (10) as

$$\frac{\sum_{i=1}^{M} a_i \Pi(S_i,T,w)}{\sum_i a_i} > \Pi^*. \quad (11)$$

This expression reduces to equation (10) if abundances a_i are all equal to each other. Furthermore, the MC criterion for accepting mutations becomes $P = \exp\left(-a_i(\Pi_{\text{old}} - \Pi_{\text{new}})/T_{\text{MC}}\right)$, so proteins of low abundance are less constrained in their sequences as long as the stability criterion (9) is still satisfied. This feature of the model is supported by our analysis of real data, where temperature trends in bacteria and archaea were similar for highly abundant proteins but less so for complete proteomes. The effective decrease of MC temperature for highly abundant proteins also provides for their slower evolution, established, for example, by Drummond et al. (2005). The distribution of protein stabilities P_{nat} of all sequences generated in the simulation is shown in supplementary figure S15, Supplementary Material online. As expected, low-abundance proteins have a lower average stability and a wider distribution of stabilities.

Therefore, for each combination of the two input parameters, environmental temperature T and chaperone-adjusted synthesis cost w, we were able to generate simulated proteomes of 10^5 sequences of length 64 each, optimizing the proteostasis costs (8) or (11) subject to stability condition (9). The GPU-accelerated lattice protein folding library GaleProt (Venev and Zeldovich 2015) was used for massively parallel evaluation of P_{nat}. Frequencies of amino acids found in the simulated proteomes were used for comparison with genomics data.

Data Sets

We used RefSeq and BioProject databases at NCBI to retrieve 543 completely sequenced, annotated, single-chromosome bacterial genomes with known OGT or a specified environmental temperature. A Python script (Cock et al. 2009) was used to retrieve OGT data from NCBI Entrez. If only a temperature range was specified, the average temperature was used as OGT. Following (Goncearenco et al. 2014), we removed 281 overrepresented species with the values of OGT of 27.5°, 30°, 37° as they represent plant and animal pathogens and experience diverse selective pressures unrelated to environmental temperature. The bacterial data set covers the OGT range of 15–90° and genome-wide GC content (GC) of 30–70% (supplementary fig. S16*B*, Supplementary Material online). As archaea are much less represented in the BioProject database, we performed a manual literature search for OGT of 617 species of archaea available in GenBank. The search yielded 223 species with known OGT and sufficient annotation (whole genome shotgun assemblies were included if at least 600 protein coding sequences were annotated). Genomes of 83 halophiles have been excluded from analysis, as they experience a strong evolutionary pressure of hypersaline environment (Fukuchi et al. 2003), and appear as outliers on the overall monotonous OGT trends of amino acid usage. The scatter plots in genomic GC-OGT coordinates for archaea (supplementary fig. S16*A*, Supplementary Material online) reveal a relatively homogeneous coverage, with the GC range 30–70% and OGT 25–110° with a lower coverage at ~60° OGT, which may be attributed to the lack of corresponding environments. The analyzed data set comprised 262 bacteria and 140 archaea with sufficient annotation and known OGT, accessions numbers and OGT are listed in supplementary tables S2 and S3, Supplementary Material online. Analysis scripts and protocols are available at http://github.com/sergpolly/Thermal_adapt_scripts, last accessed November 1, 2017.

Identification of Highly Abundant Proteins

Protein abundance and expression level are important factors to consider when calculating energetic costs. Unfortunately, for most of prokaryotes with completely sequenced genomes neither protein abundance nor expression have been directly characterized, for example, there are only two archaeal entries in the major protein abundance database, PaxDB (Wang et al. 2015). We used a sequence based approach to identify putatively highly expressed proteins using CAI (Sharp and Li 1987). Ribosomal proteins were used as a reference of highly expressed proteins (Pedersen et al. 1978; Srivastava and Schlessinger 1990) to establish the codon usage pattern. We selected all species with at least 25 annotated ribosomal proteins, and used CAI as a proxy for expression and abundance level.

Previously, it has been shown that CAI has its limitations as a predictor of gene expression (Botzman and Margalit 2011), as in some species the CAI distribution is very narrow and codon usage of ribosomal protein genes is nearly indistinguishable from other genes (supplementary fig. S17*B*, Supplementary Material online). To address this issue, we selected a group of genomes where at least 85% of ribosomal protein genes are within the 25% of all genes with the highest CAI rank. This empirical criterion selects genomes with wide distributions of CAI and a marked difference in codon usage between ribosomal and other proteins (supplementary fig. S17*A*, Supplementary Material online), which in turn implies strong codon usage selection (CUS). We assume that in organisms with CUS, CAI can be used as a proxy for gene expression and, statistically, abundance (Sharp and Li 1987; Jansen et al. 2003; Supek and Vlahovicek 2005; Maier et al. 2009). CUS was identified in ~50% of species used in this study, 167 bacteria out of 262 and 65 archaea out of 140. Our CUS criterion is compatible with the criteria proposed in Botzman and Margalit (2011) (supplementary fig. S18, Supplementary Material online), and preserves a relatively uniform GC-OGT distribution of species (supplementary fig. S16, Supplementary Material online). For CUS organisms, highly expressed genes (abundant proteins) were defined as the genes within the top 10% of CAI values. Thus we avoided using CAI-ranking for individual genes, which in turn mitigates the problem of poor expression versus abundance correlation (Maier et al. 2009).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors gratefully acknowledge the help of Alexey Shaytan and NCBI helpdesk for assistance with NCBI databases. This work was supported in part by the Defense Advanced Research Project Agency (DARPA), Prophecy Program (Contracts Nos. HR0011-11-C-0095 and D13AP00041) and by NIH P01 GM109767.

References

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A*. 99(6):3695–3700.

Berezovsky IN, Zeldovich KB, Shakhnovich EI. 2007. Positive and negative design in stability and thermal adaptation of natural proteins. *PLoS Comput Biol*. 3(3):0498–0507.

Bloom JD, Labthavikul ST, Otey CR, Arnold FH. 2006. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A*. 103(15):5869–5874.

Bogumil D, Dagan T. 2012. Cumulative impact of chaperone-mediated folding on genome evolution. *Biochemistry* 51(50):9941–9953.

Botzman M, Margalit H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol*. 12(10):R109.

Çetinbaş M, Shakhnovich EI, Pande VS. 2013. Catalysis of protein folding by chaperones accelerates evolutionary dynamics in adapting cell populations. *PLoS Comput Biol*. 9(11):e1003269.

Cherry JL. 2010. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol*. 27(3):735–741.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.

Dill KA. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24(6):1501–1509.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102(40):14338–14343.

England JL, Shakhnovich BE, Shakhnovich EI. 2003. Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc Natl Acad Sci U S A*. 100(15):8727–8731.

Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. 2003. Unique amino acid composition of proteins in halophilic bacteria. *J Mol Biol*. 327(2):347–357.

Galtier N, Lobry J. 1997. Relationships between genomic g+c content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*. 44(6):632–636.

Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, Drummond DA. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci U S A*. 108(2):680–685.

Goldstein RA. 2007. Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: insights from the quasi-chemical approximation. *Protein Sci*. 16(9):1887–1895.

Goldstein RA. 2008. The structure of protein evolution and the evolution of protein structure. *Curr Opin Struct Biol*. 18(2):170–177.

Goncearenco A, Berezovsky IN. 2014. The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. *Biol Direct*. 9(1):29.

Goncearenco A, Ma B-G, Berezovsky IN. 2014. Molecular mechanisms of adaptation emerging from the physics and evolution of nucleic acids and proteins. *Nucleic Acids Res*. 42(5):2879–2892.

Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic methanococcus species. *Proc Natl Acad Sci U S A*. 96(7):3578–3583.

Hartl FU, Bracher A, Hayer-Hartl M. 2011. Molecular chaperones in protein folding and proteostasis. *Nature* 475(7356):324–332.

Heizer EM, Raiford DW, Raymer ML, Doom TE, Miller RV, Krane DE. 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Mol Biol Evol*. 23(9):1670–1680.

Heizer EM, Raymer ML, Krane DE. 2011. Amino acid biosynthetic cost and protein conservation. *J Mol Evol*. 72(5–6):466–473.

Jansen R, Bussemaker HJ, Gerstein M. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res.* 31(8):2242–2251.

Jukes TH, Holmquist R, Moise H. 1975. Amino acid composition of proteins: selection against the genetic code. *Science (New York, NY)* 189(4196):50–51.

Karlin S, Mrázek J, Ma J, Brocchieri L. 2005. Predicted highly expressed genes in archaeal genomes. *Proc Natl Acad Sci U S A.* 102(20):7303–7308.

Karshikoff A, Nilsson L, Ladenstein R. 2015. Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS J.* 282(20):3899–3917.

Kepp KP, Dasmeh P, Sanchez-Ruiz JM. 2014. A model of proteostatic energy cost and its use in analysis of proteome trends and sequence evolution. *PLoS ONE* 9(2):1–12.

King JL, Jukes TH. 1969. Non-darwinian evolution. *Science (New York, NY)* 164(881):788–798.

Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and gc composition within and across genomes. *Genome Biol.* 2(4):RESEARCH0010.

Kreil DP, Ouzounis CA. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* 29(7):1608–1615.

Krick T, Verstraete N, Alonso LG, Shub DA, Ferreiro DU, Shub M, Sánchez IE. 2014. Amino acid metabolism conflicts with protein diversity. *Mol Biol Evol.* 31(11):2905–2912.

Kumar S, Nussinov R. 2002. Relationship between ion pair geometries and electrostatic strengths in proteins. *Biophys J.* 83(3):1595–1612.

Leuenberger P, Ganscha S, Kahraman A, Cappelletti V, Boersema PJ, von Mering C, Claassen M, Picotti P. 2017. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* 355(6327):eaai7825.

Lightfield J, Fram NR, Ely B, Otto M. 2011. Across bacterial phyla, distantly-related genomes with similar genomic gc content have similar patterns of amino acid usage. *PLoS ONE* 6(3):1–12.

Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci U S A.* 107(7):2983–2988.

Loladze VV, Makhatadze GI. 2008. Removal of surface charge-charge interactions from ubiquitin leaves the protein folded and very stable. *Protein Sci.* 11(1):174–177.

Loladze VV, Ibarra-Molero B, Sanchez-Ruiz JM, Makhatadze GI. 1999. Engineering a thermostable protein via optimization of charge–charge interactions on the protein surface. *Biochemistry* 38(50):16419–16423.

Maier T, Güell M, Serrano L. 2009. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583(24):3966–3973.

McDonald JH. 2010. Temperature adaptation at homologous sites in proteins from nine thermophile–mesophile species pairs. *Genome Biol Evol.* 2(1):267–276.

Miyazawa S, Jernigan RL. 1999. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* 34(1):49–68.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158(2):927–931.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7(5):337–348.

Pedersen S, Bloch PL, Reeh S, Neidhardt FC. 1978. Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. *Cell* 14(1):179–190.

Ponnuswamy P, Muthusamy R, Manavalan P. 1982. Amino acid composition and thermal stability of proteins. *Int J Biol Macromol.* 4(3):186–190.

Price ND, Reed JL, Palsson BØ. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol.* 2(11):886–897.

Pucci F, Dhanani M, Dehouck Y, Rooman M, Zhang Y. 2014. Protein thermostability prediction within homologous families using temperature-dependent statistical potentials. *PLoS ONE* 9(3):1–8.

Raiford D, Heizer Esley M, Miller J, Akashi R, Raymer H, Krane MD. 2008. Do amino acid biosynthetic costs constrain protein evolution in *Saccharomyces cerevisiae*? *J Mol Evol.* 67(6):621–630.

Rocha EPC, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol.* 21(1):108–116.

Rocha EPC, Feil EJ, Nachman MW. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *PLoS Genet.* 6(9):1–4.

Sabath N, Ferrada E, Barve A, Wagner A. 2013. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol Evol.* 5(5):966–977.

Santra M, Farrell DW, Dill KA. 2017. Bacterial proteostasis balances energy and chaperone utilization efficiently. *Proc Natl Acad Sci U S A.* 114(13):E2654–E2661.

Sawle L, Ghosh K. 2015. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J Chem Phys.* 143(8):085101.

Seligmann H. 2003. Cost-minimization of amino acid usage. *J Mol Evol.* 56(2):151–161.

Serohijos AWR, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* 2(2):249–256.

Sghaier H, Thorvaldsen S, Saied NM. 2013. There are more small amino acids and fewer aromatic rings in proteins of ionizing radiation-resistant bacteria. *Ann Microbiol.* 63(4):1483–1491.

Shakhnovich E, Gutin A. 1990. Enumeration of all compact conformations of copolymers with random sequence of links. *J Chem Phys.* 93(8):5967.

Sharp PM, Li W-H. 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15(3):1281–1295.

Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface.* 11(100):20140419.

Singer GA, Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 317:39–47.

Srivastava AK, Schlessinger D. 1990. Mechanism and regulation of bacterial ribosomal RNA processing. *Ann Rev Microbiol.* 44(1):105–129.

Strickler SS, Gribenko AV, Gribenko AV, Keiffer TR, Tomlinson J, Reihle T, Loladze VV, Makhatadze GI. 2006. Protein stability and surface electrostatics: a charged relationship. *Biochemistry* 45(9):2761–2766. PMID: 16503630.

Sueoka N. 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc Natl Acad Sci U S A.* 47(8):1141–1149.

Supek F, Vlahovicek K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics* 6(1):182.

Swire J. 2007. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J Mol Evol.* 64(5):558–571.

Szilagyi A, Zavodszky P. 2000. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* 8(5):493–504.

Taverna DM, Goldstein RA. 2002. Why are proteins so robust to site mutations? *J Mol Biol.* 315(3):479–484.

Thomas PD, Dill KA. 1996. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A.* 93(21):11628–11633.

Tomala K, Pogoda E, Jakubowska A, Korona R. 2014. Fitness costs of minimal sequence alterations causing protein instability and toxicity. *Mol Biol Evol.* 31(3):703–707.

Uversky VN, Gillespie JR, Fink AL. 2000. Why are natively unfolded proteins unstructured under physiologic conditions? *Proteins* 41(3):415–427.

Varma A, Palsson BØ. 1994. Metabolic flux balancing: basic concepts, scientific and practical use. *Nat Biotechnol*. 12(10):994–998.

Venev SV, Zeldovich KB. 2015. Massively parallel sampling of lattice proteins reveals foundations of thermal adaptation. *J Chem Phys*. 143(5):055101.

Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of paxdb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15(18):3163–3168.

Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI. 2007a. A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. *PLoS Comput Biol*. 3(7):e139.

Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007b. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol*. 3(1):0062–0072.

Zeldovich KB, Chen P, Shakhnovich EI. 2007c. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci U S A*. 104(41):16152–16157.

Zhao N, Pang B, Shyu CR, Korkin D. 2011. Charged residues at protein interaction interfaces: unexpected conservation and orchestrated divergence. *Protein Sci*. 20(7):1275–1284.