ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Proteins from Thermophilic *Thermus thermophilus* Often Do Not Fold Correctly in a Mesophilic Expression System Such as *Escherichia coli*

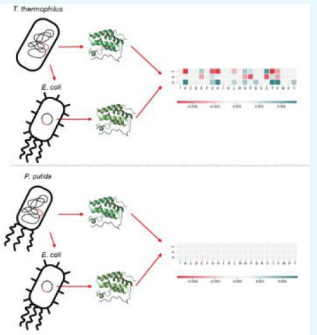
Alibek Kruglikov, Yulong Wei, and Xuhua Xia*

Cite This: *ACS Omega* 2022, 7, 37797−37806

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Majority of protein structure studies use *Escherichia coli* (*E. coli*) and other model organisms as expression systems for other species' genes. However, protein folding depends on cellular environment factors, such as chaperone proteins, cytoplasmic pH, temperature, and ionic concentrations. Because of differences in these factors, especially temperature and chaperones, native proteins in organisms such as extremophiles may fold improperly when they are expressed in mesophilic model organisms. Here we present a methodology of assessing the effects of using *E. coli* as the expression system on protein structures. We compare these effects between eight mesophilic bacteria and *Thermus thermophilus* (*T. thermophilus*), a thermophile, and found that differences are significantly larger for *T. thermophilus*. More specifically, helical secondary structures in *T. thermophilus* proteins are often replaced by coil structures in *E. coli*. Our results show unique directionality in misfolding when proteins in thermophiles are expressed in mesophiles. This indicates that extremophiles, such as thermophiles, require unique protein expression systems in protein folding studies.



1. INTRODUCTION

Identification of protein structure is a major requirement in studies on protein functions[1,2] and interactions[3−5] and in the design of novel enzymes.[6−8] As of July 2022, there were more than 190000 records in Protein Data Bank (PDB),[9] the largest database of protein structures,[10] of which almost 180000 were protein structure entries. The deposited structures can be used in various fields of research. For example, PDB data had been recently used in research related to COVID-19,[11−14] protein evolution,[15−17] computational enzyme,[18] and drug design,[19] as well as for training computational protein structure prediction algorithms.[20−24]

While PDB holds a relatively large variety of protein types and source species, diversity is much lower for expression systems used in structure determination experiments. Protein source species are the species that the protein-coding sequences were taken from, and protein expression systems are the species that these proteins were grown in (Figure 1). A majority of PDB experiments use *Escherichia coli* (*E. coli*) as the expression system. For example, out of over 1800 PDB entries with *Bacillus subtilis* (*B. subtilis*) as the source organism, more than 1700 were grown in *E. coli*. For most other species, the proportion of proteins that were grown in *E. coli* is even higher.

While using recombinant model organisms—those with genetically recombined genes—for protein production is generally effective, lower protein activity and solubility can be observed in many cases, including the formation of inclusion bodies.[25−27] Various protocols and methodologies have been developed in an effort to improve recombinant protein production quality; however, their effectiveness is not uniform across different protein source species. For many thermophilic species, production of recombinant protein in *E. coli* in active form is still a major challenge. For example, multiple studies show that a large fraction of *Thermus thermophilus* (*T. thermophilus*) proteins are formed in insoluble and/or in inactive form when grown in recombinant *E. coli*,[28−31] potentially because *T. thermophilus* has optimal growth temperatures around 70−80 °C, which are much higher than that of *E. coli* (37 °C). It has been previously reported that soybean Late Embryogenesis Abundant proteins became more hydrated upon heating,[32] suggesting that protein solubility is influenced by the cellular environment; therefore what is not soluble in mesophilic *E. coli* may well be soluble in *T. thermophilus*. In fact, induction temperature is one of the most critical growing conditions to produce soluble protein.[33]

These findings suggest that proteins may misfold in recombinant expression systems having dissimilar cellular environments. Nevertheless, expression system is not always considered important or even reported in structural studies. Moreover, protein structure prediction models, including the most advanced ones, such as AlphaFold2,[24] do not use

Received: July 28, 2022
Accepted: October 7, 2022
Published: October 14, 2022

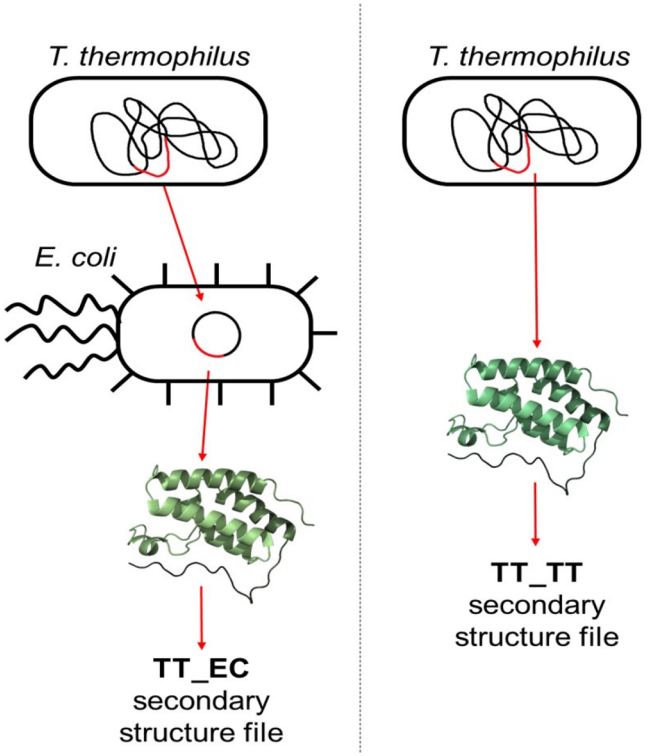


Figure 1. Example of protein structures in different expression systems. TT_EC file is formed using proteins that have *T. thermophilus* as source organism and *E. coli* as expression system. TT_TT is formed using proteins that have *T. thermophilus* as both source and expression system (sometimes referred to as "native" expression system in this research).

expression system and cellular environment as factors in training and validation. This disparity highlights the need for a methodology to assess the effect of different expression systems on protein structure determination.

Here we present a metric to evaluate protein secondary structure (SS) differences and analyze how changing the expression system from "native" to *E. coli* affects protein SS. We used PDB data to create AA/SS data sets for *T. thermophilus* (TT) and seven nonthermophilic bacteria species (where AA means amino acid and SS means secondary structure). For each species (say XX), there are two sets of protein structure data, one obtained with XX as both the protein source and the expression system and the other with XX as the source species but *E. coli* as the expression system. These two sets of protein structures are represented as XX_XX and XX_EC (Figure 1, where XX is TT). We then processed the data into probability matrices and calculated Jensen–Shannon divergence (JSD) between the XX_XX matrix and the XX_EC matrix. This JSD measured the difference in protein structure between XX_XX and XX_EC. We found that JSD was significantly higher between TT_TT and TT_EC than between XX_XX and XX_EC (where XX is a mesophilic bacterial species). This implies that *T. thermophilus* proteins in nonthermophilic species do not fold in the same way as in their "native" thermophilic expression system, while for the other species the expression system did not affect protein folding.

2. MATERIALS AND METHODS

For a fair comparison, we need the same protein from a source species but expressed in different expression systems, one being the source system (native) and the other being *E. coli*. An overview of our methodology generating such data is summarized in Figures 2 and 3.

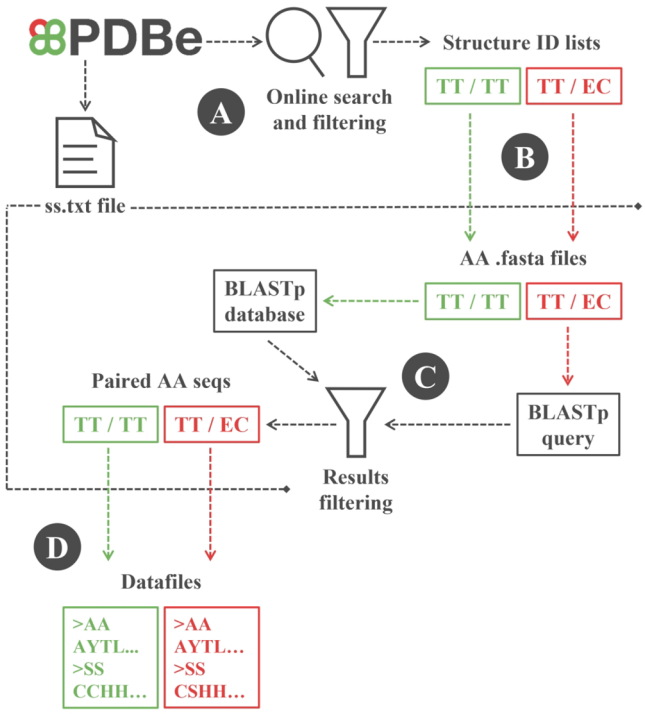


Figure 2. Overview of methodology (part 1). Relevant structure IDs were found through a search on PDBe (A) and processed into AA.fasta files (B). The found structures with the same protein origin species and different expression species were paired on the basis of AA sequences using BLASTp (C), and the paired AA/SS were saved as datafiles (D).

2.1. Data Collection. Bacterial species from which we collected the data are listed in Table 1 and were selected because they have sufficient data available on PDB both as protein source and expression systems. The PDB online advanced searching tool was used to create separate ID lists for each source/expression pair. More specifically, PDB Europe (PDBe) was used because of its more advanced filtering interface; however, its data are the same as on PDB. Advanced search can be accessed through this link: https://www.ebi.ac.uk/pdbe/entry/search/index/?advancedSearch:true=. We used several filters to obtain protein ID lists: (1) organism name, (2) expression host name, (3) resolution, and (4) molecule type. Protein source and expression system were set in accordance with Table 1, molecular type was set to "protein" and experimental method was set to X-ray diffraction only with resolution between 0 and 2.5 Å. Note that proteins are purified in their naturally folded state in the expression system, ideally in their functional forms, before crystallization and X-ray diffraction.

Structure IDs were taken for each protein origin/expression system pair found using the online search and used to obtain the corresponding SS and AA sequences. We used a PDB Secondary Structure file in FASTA format (latest version is accessible at https://cdn.rcsb.org/etl/kabschSander/ss.txt.gz—sometimes multiple refreshes of the page are required to obtain the file; alternatively, a copy of the file is available at https://github.com/alibekk93/project-protein_folding_

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806

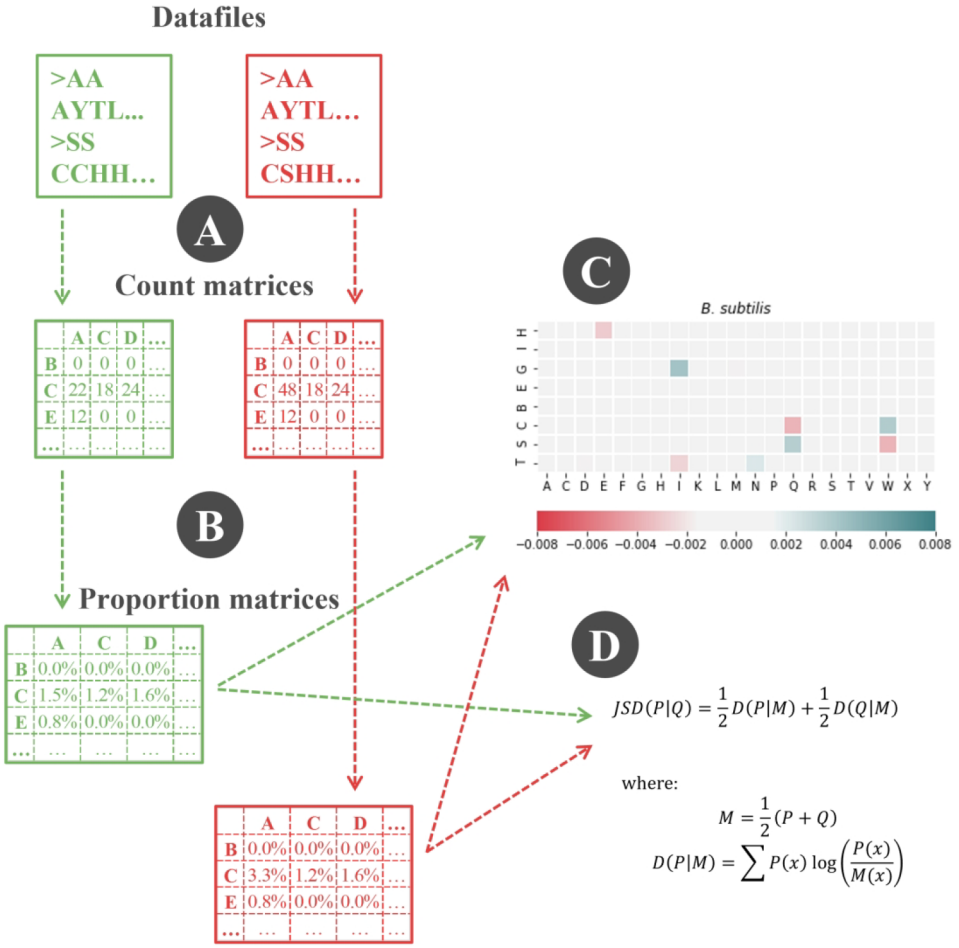


Figure 3. Overview of methodology (part 2). Datafiles were used to construct AA/SS count (A) and proportion matrices (B). For each studied species, JSD was calculated between the "native" proportion matrix and the recombinant *E. coli* proportion matrix (D). The differences between the matrices were also visualized using heat maps (C).

distances/ss.txt.gz) to collect SS and AA. Each structure in this file is represented by AA and SS strings, where each single AA corresponds to a single SS at the same position. The numbers of structure IDs identified for each ID list are shown on Table 1. Identical AA sequences from the same expression systems were allowed as their corresponding SS sequences could differ and provide relevant data.

2.2. Filtering the Data Using BLASTp. To identify proteins whose structure has been determined when they were expressed in both the source species (native environment) and in *E. coli*, we processed the AA sequences obtained from the previous step into BLASTp database files using DAMBE.[34] For each pair of compared protein databases a BLASTp search was performed using AA sequences of the proteins, with the recombinant *E. coli* expression system database as query and the native database as BLASTp database.

We used ungapped BLASTp with three-letter words to match proteins from different databases. Minimal matching accuracy was set to 95% to allow for small variations of AA sequence due to point mutations and random errors in experiment without allowing different proteins to be matched. Minimal matching length was set to 50 to remove short matching sequences, and the maximal *E*-value was set to 0.01. BLASTp parameters were set so that only proteins with very similar AA chains and those likely to be the same protein were kept in both databases and so that only matching parts could contribute to the analysis. For each pair obtained, sequence start and end were used to cut out the matching parts of the sequences and not include the nonmatching parts.

Filtering protein databases using BLASTp removed most of the proteins from the original Protein ID lists. This happened because most entries on PDB are only available with one expression system (grown in *E. coli*). On the other hand, those entries available in the native expression system are also not always available with the *E. coli* expression system. For example, the original BS_BS database contained eight entries and the original BS_EC database had 1145 entries; however, after BLASTp only four entries from BS_BS matched with 17 entries from BS_EC. The full list of numbers of entries in the databases before and after BLASTp is shown in Table 1.

2.3. Construction of Probability Matrices. First, tabular BLASTp results were used to create data files in FASTA format for each protein source species/expression system combination, each containing the AA and SS sequences. Structure IDs were used to obtain AA and SS sequences from the PDB file. Query and database start and end positions were used to cut the matching parts from the obtained sequences. This way only matching parts of the proteins would be left.

In many cases a single-query structure would match more than one database structure and vice versa; therefore, it was necessary to multiply AA and SS sequences in those cases to make sure that matching parts are the same length. In this way

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806

Table 1. Species Used in the Analysis[a]

protein source species	expression system	structure IDs list	entries before BLASTp	entries after BLASTp
Bacillus subtilis	*B. subtilis*	BS_BS	8	4
Bacillus subtilis	*E. coli*	BS_EC	1145	17
Desulfovibrio vulgaris	*D. vulgaris*	DV_DV	26	4
Desulfovibrio vulgaris	*E. coli*	DV_EC	67	20
Lactococcus lactis	*L. lactis*	LL_LL	25	5
Lactococcus lactis	*E. coli*	LL_EC	166	6
Pseudomonas fluorescens	*P. fluorescens*	PF_PF	13	4
Pseudomonas fluorescens	*E. coli*	PF_EC	275	2
Pseudomonas putida	*P. putida*	PP_PP	3	1
Pseudomonas putida	*E. coli*	PP_EC	542	37
Salmonella enterica	*S. enterica*	SE_SE	34	29
Salmonella enterica	*E. coli*	SE_EC	860	17
Streptomyces rubiginosus	*St. rubiginosus*	SR_SR	17	17
Streptomyces rubiginosus	*E. coli*	SR_EC	11	11
Thermus thermophilus	*T. thermophilus*	TT_TT	19	7
Thermus thermophilus	*E. coli*	TT_EC	1091	3

[a]For each protein source species there are two expression systems used—that of the source species ("native" expression system) and *E. coli*. Lists of structure IDs were created for each source/expression pair. For example, BS_BS contains IDs of structures with *B. subtilis* as both protein source organism and protein expression system, while BS_EC has IDs of structures with *B. subtilis* as protein source organism and *E. coli* as protein expression system.

we obtained chains from the BLAST database with identical AA sequences, but possibly different SS sequences, and had each variation of SS in correct proportions. AA and SS sequences were concatenated for each of the databases so that all AA sequences were in one line and all SS sequences were in another line of the resulting file.

Data files were processed into count matrices to count each AA/SS combination. We converted count matrices into probability matrices by simple division of each count value by count matrix sum. Protein SS can be described with three SS types: H (helix), E (sheet), and C (coil) or with eight SS types: H (α-helix), I (π-helix), G (3_{10}-helix), E (β-sheet), B (β-bridge), C (coil), S (bend), and T (turn).. We refer to the two classification systems as 3-SS (three types of SS) and 8-SS (eight types of SS) in this work. While many models and studies, especially the earlier ones, use 3-SS, using 8-SS can give more details. In order to work with both 8-SS and 3-SS, we transformed the original 8-SS matrices to 3-SS by adding up the matrix values.

2.4. Jensen–Shannon Divergence Calculation and Statistical Analysis. After the matrices were created, they were compared to each other, so that for each protein source species the two matrices compared were with *E. coli* and "native" expression systems. Jensen–Shannon divergence (JSD) was used to evaluate differences between matrices. A common measure of probability distribution differences is Kullback–Leibler divergence (KLD). It is nonsymmetric, which means that KLD of distribution P from distribution Q does not have to be equal to KLD of distribution Q from distribution K. KLD is defined as

$$\mathrm{KLD}(P||Q) = \sum P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

where $P(x)$ and $Q(x)$ are discrete probability distributions.

JSD is a metric similar to KLD, and it could be called a symmetrized and smoothed version of it. It is defined as

$$\mathrm{JSD}(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

where $P(x)$ and $Q(x)$ are discrete probability distributions, $M = \frac{1}{2}(P + Q)$, and D is KLD.

In our case JSD is a more suitable measure than KLD due to the large number of zeros in our data. KLD requires a division of one probability by another, and when the denominator probability is zero (which is very often the case in our data), calculation results in infinity. Standard practice is to drop zeros completely, but in our case that would be dropping a very significant amount of our data, because probability matrices contain a lot of zeros. Contrary to KLD, JSD does not have this problem thanks to its symmetric nature. Probability distributions are compared not with each other but with their average distribution, which means that only positions where both probability matrices are zero need to be dropped—those positions are exact in any case and therefore dropping them is not an issue.

Larger JSD would mean that changing the protein expression system from "native" to *E. coli* had more effects and that a recombinant protein SS is a worse representative of native SS. Moreover, in order to deduce possible mechanisms of how the change of the expression system affects protein folding, we calculated difference between the matrices by subtracting recombinant proportion matrix values from native proportion matrix. Subtraction results would not be a good evaluation of matrix divergence but can show at which AA/SS positions the differences between matrices are large.

We estimated the statistical significance of calculated JSD using resampling techniques. Bootstrapping was used to calculate 95% confidence intervals of JSD—AA/SS positions were randomly resampled with replacement 1000 times from the original data files and JSD were calculated for them. In addition, we used permutation to test the significance of differences between species' JSD values by combining all data files into a uniform distribution and resampling positions from it 10000 times to form data of 1000 residues in length.

3. RESULTS

3.1. Magnitude of Expression System Effect. Calculated results are summarized on Table 2. In addition, we visualize bootstrapping results on box plots (Figures 4 and 5) and permutation results on histograms (Figure 6). The largest JSD was found between *T. thermophilus* matrices, and this was the case using both 8-SS and 3-SS. Moreover, bootstrapping results show that *T. thermophilus* JSD is the only one significantly higher than other species' JSD for both 8-SS and 3-SS.

3.2. Directionality of Expression System Effect. Heat maps in Figures 7 and 8 show differences between proportion matrices used in JSD calculations. These figures help in

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806

Table 2. Mean Jensen−Shannon Divergences between Native and Recombinant AA/SS Matrices[a]

protein source species	JSD8	JSD8 *p*-value	JSD3	JSD3 *p*-value
Bacillus subtilis	0.010	0.082	0.001	0.697
Desulfovibrio vulgaris	0.008	0.343	0.003	0.004
Lactococcus lactis	0.007	0.594	0.003	0.012
Pseudomonas fluorescens	0.009	0.270	0.002	0.101
Pseudomonas putida	0.002	1.000	0.000	1.000
Salmonella enterica	0.003	0.999	0.000	1.000
Streptomyces rubiginosus	0.004	0.995	0.001	0.992
Thermus thermophilus	0.033	0.000	0.014	0.000

[a] *p*-values are from bootstrapping analysis testing the null hypothesis that different expression systems have no effect on protein structure. JSD8 and JSD3 are the calculated JSD using 8 or 3 types of SS. Both JSD8 and JSD3 are the greatest for *T. thermophilus* matrices, and that is the only species where both metrics are significantly different from bootstrapped distributions. This indicates that switching the protein expression system from "native" to *E. coli* affects the folding of *T. thermophilus* proteins more than other species' proteins.

identifying which particular elements of matrices were most different and display potential directionality of the differences. In line with JSD results, *T. thermophilus* matrices show more differences than any other species' matrix pair. It can be seen that using *E. coli* as an expression system for thermophilic proteins leads to lower frequencies in helices and higher frequencies of coils. The 8-SS heat map shows that this effect is particularly strong on 3_{10}-helices (structure G). Other protein expression systems considered in this study show smaller differences from that of *E. coli* as there are no large JSD values detected for them.

No strong patterns have been discovered in terms of variability of secondary structures between different amino acids (Figures 7 and 8). While distances between matrices of *T. thermophilus* proteins seem to be high with hydrophobic alanine, valine, and glycine, that is also the case for histidine (charged) and threonine (polar and uncharged).

4. DISCUSSION

4.1. Lack of Required Chaperones. There are several possible explanations for variations in JSD for different species. Because *T. thermophilus* is a thermophile, it is adapted to protein denaturation, partially through chaperone-dependent protein folding. It is possible that when *E. coli* is used as an expression system, certain helices are not formed or repaired due to a lack of these chaperones and coils are formed instead. For example, DnaK chaperone expression requires less ATPase activity in *T. thermophilus* than in *E. coli* and it participates in protein folding mediation.[35] Trigger factor proteins also show differences in structure and activity between the species.[36] Such effects are likely to be particularly important for any intrinsically unstructured proteins that require chaperone activity for correct folding and structural stability.[37] While a common way around this problem is to co-express required folding chaperones together with the studied protein, it is not always clear whether that was done on PDB because not all structures there have publications and, even if they do, it is not always clearly explained whether co-expression of chaperones was performed. Moreover, co-expression of chaperones does not always provide the desired effects as differences in other cell-specific factors may lead to chaperones losing their activity or even becoming toxic for the host.[38]

4.2. Suboptimal Cellular Environment. Due to thermophilic adaptations of *T. thermophilus*, it is possible that the tendency toward coil structures instead of helical structures in *E. coli* as an expression system is a result of differences in cellular conditions of the expression systems, namely, nonoptimal folding temperatures. This would explain why this effect is more apparent for 3_{10}-helices than α-helices, as the latter are more stable.[39] Regardless, future research directions may prompt researchers to study varying environmental conditions of expression systems and their effect on protein folding using data with varying temperature, pH, and salinity.

Differences in protein solubility due to temperatures could have affected protein crystallization and thus structure identification.[40−42] To assess this possibility, an analysis of structures with different crystallization techniques performed might be necessary. That kind of analysis would help to determine whether effects observed in this experiment are due to differences in cellular environments and chaperones or due to experimental design.

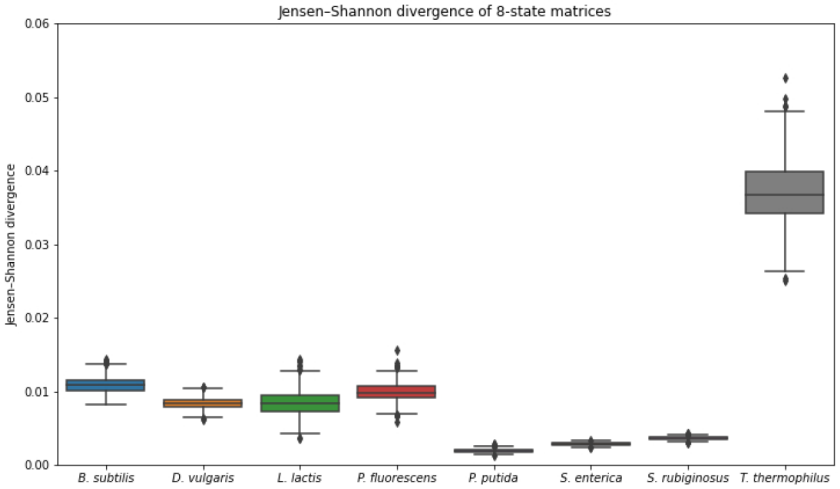


Figure 4. Box plots of bootstrapped JSD (8 SS types). High JSD indicates larger differences between "native" and *E. coli* expression systems. *T. thermophilus* JSD are much higher than those of other bacteria.

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806

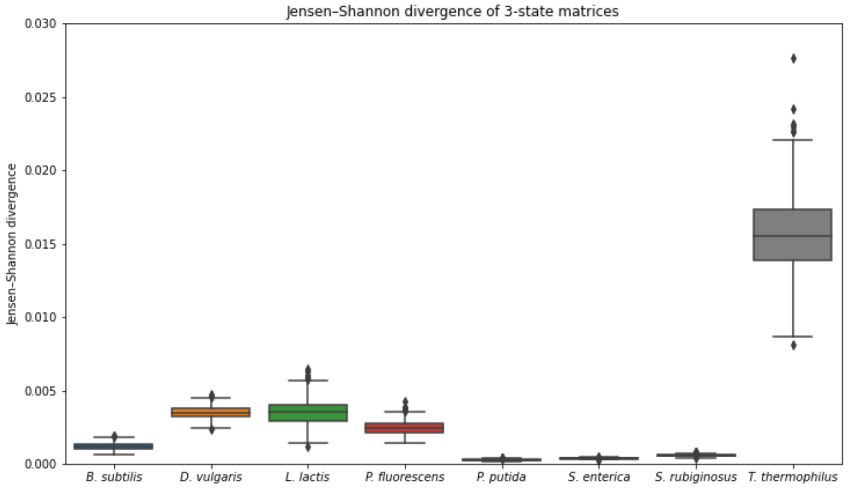


Figure 5. Box plots of bootstrapped JSD (3 SS types). High JSD indicates larger differences between "native" and *E. coli* expression systems. *T. thermophilus* JSD are much higher than those of other bacteria.

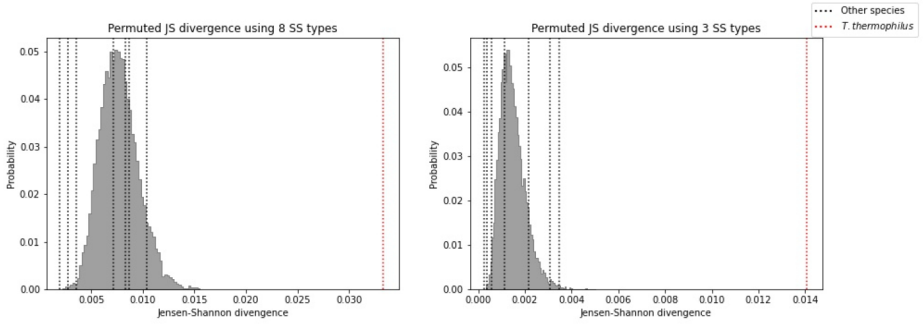


Figure 6. Distributions of permuted JSD results. *T. thermophilus* JSD (red line) is much larger than JSDs of the other species (gray lines) and the nonspecific JSD (gray histogram).

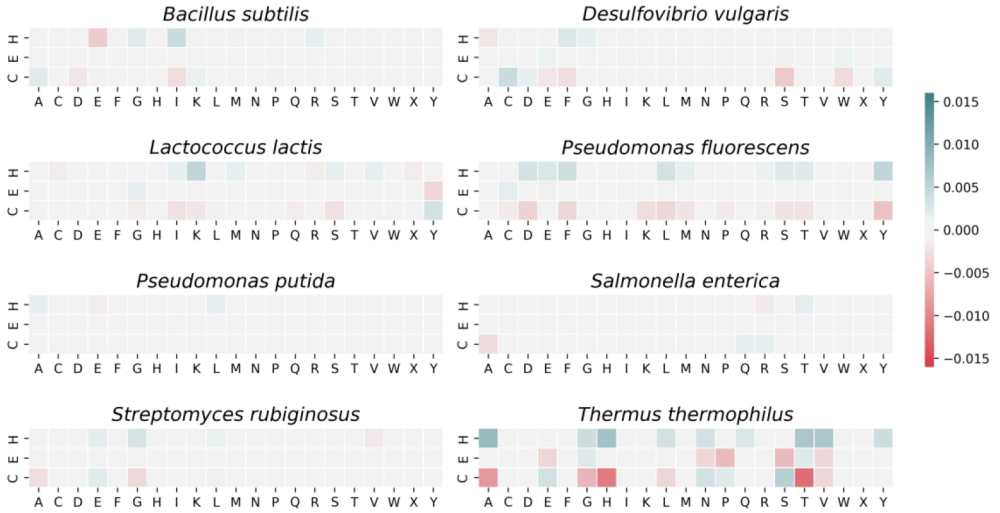


Figure 7. Heat maps of proportion matrices differences with 3 SS types showing directionality of the effect induced by using *E. coli* as the expression system. More negative values (red) indicate larger proportions in *E. coli* as the expression system; more positive values (green) indicate larger proportions in "native" expression systems. The effects were most visible in *T. thermophilus*, where helices (H) were observed more frequently when proteins were expressed in *T. thermophilus* and coils (C) were instead more abundant when proteins were expressed in *E. coli*. No such effect nor directionality of differences could be seen in other species. The three SS types are H (helix), E (sheet), and C (coil).

4.3. Codon Optimization. Protein folding can be affected by the rate of protein synthesis, which can be controlled by codon usage.[43,44] Assuming equal translation initiation rates, nonoptimal codon usage in the recombinant expression system will lead to slower rates of protein production, which in turn may lead to protein misfolding and aggregation.[45] Unfortunately, the PDB itself has no information about codon optimization in experiments and nucleotide sequences are

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806

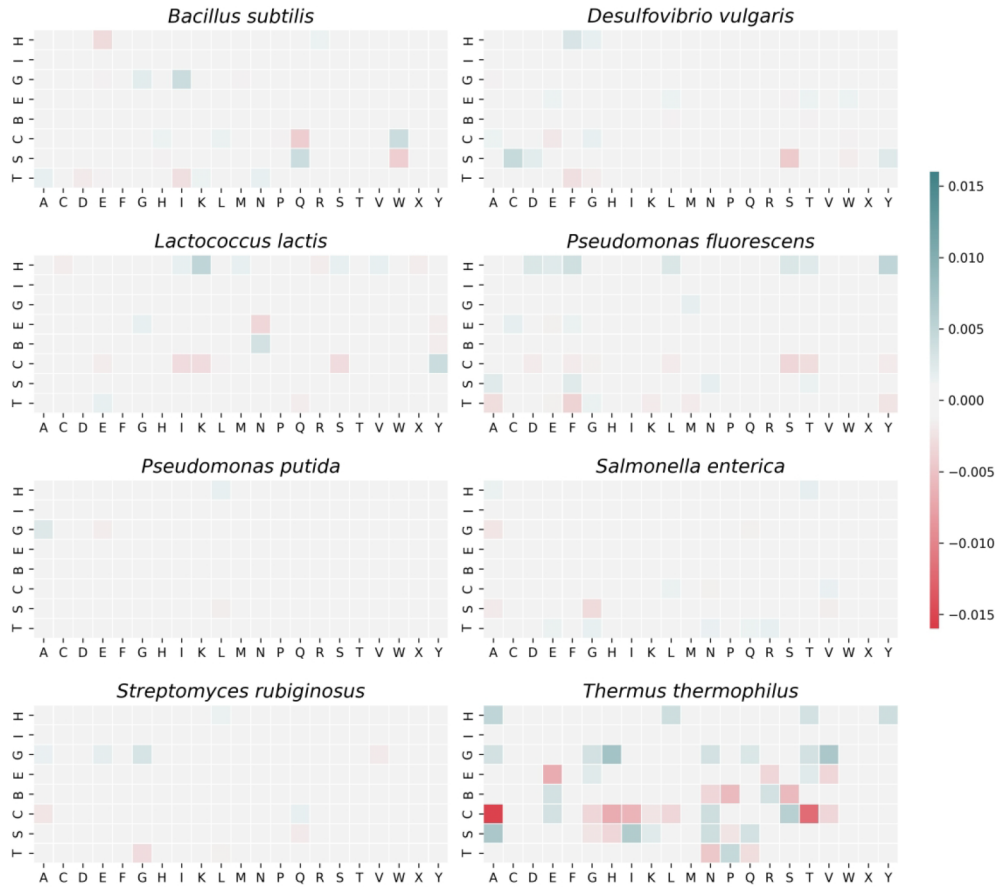


Figure 8. Heat maps of proportion matrices differences with 8 SS types showing directionality of effect induced by using *E. coli* as the expression system. More negative values (red) indicate larger proportions in *E. coli* as the expression system; more positive values (green) indicate larger proportions in "native" expression systems. Strong effects could be observed in *T. thermophilus*, where α-helices (H) and 3_{10}-helices (G) were observed more frequently when proteins were expressed in *T. thermophilus* and coils (C) were instead more abundant when proteins were expressed in *E. coli*. No such effect nor directionality of differences can be seen in other species. The eight SS types are H (α-helix), I (π-helix), G (3_{10}-helix), E (β-sheet), B (β-bridge), C (coil), S (bend), and T (turn).

not provided. Moreover, not all records have corresponding publications with full description of experimental and even the ones that had been published often do not have information on whether codons were optimal. This means that lack of codon optimization is a potential factor that caused high differences between *E. coli* and *T. thermophilus* expression systems in our analysis.

4.4. Protein Crowding. Macromolecular crowding is another potential explanation of our results. As concentrations of proteins and other macromolecules inside the cells increase, the volume available for new proteins being produced falls.[46−48] Crowding leads to an increase in protein thermodynamic activity, which affects folding, among other processes.[49−51] In prokaryotes this effect is more profound than in eukaryotes.[46]

Naturally, protein crowding may occur in both *E. coli* and *T. thermophilus* expression systems, as well as other systems, but recombinant expression systems are more likely to have this effect.[52,53] *E. coli* natural cellular concentration could be unsuitable for *T. thermophilus* protein folding and may lead to increased crowding and inclusion body formation.[26] In addition, chaperone-assisted misassemble prevention mechanisms may be compromised in recombinants as they would lack the required chaperones.[52,54]

4.5. Significance. The connection between protein structure and protein function is established, as structure directly dictates function.[55] Improperly folded proteins often lose their initial functions and can even gain novel toxic functions in their place. For example, many misfolded proteins related to Parkinson's and Alzheimer's diseases have neurotoxic functions.[56] It is possible that thermophilic proteins grown in *E. coli* lose their functions entirely or partially. Previous studies identified that *T. thermophilus* enzyme activity is reduced when using mesophilic recombinant hosts, such as *E. coli*.[28,31,57,58] Additionally, *E. coli* had been previously shown to be an inadequate expression system for thermophilic proteins in functional metagenomic[59−61] and directed evolution studies.[62−64] Our results are in line with the previous findings; we also expand on them, showing how a mesophilic expression system can affect secondary structures of thermophile proteins and that the change has directionality toward less helices and more coils.

Additionally, helical structures have been shown to be more common in thermostable proteins and are associated with thermostability.[65,66] The higher tendency for helix formation for *T. thermophilus* proteins when grown in "native" expression system could be a mechanism of protein stabilization under higher temperatures. Using *E. coli* as an expression system led

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806

to higher proportions of coils and therefore might have reduced thermostability adaptation.

In some cases, thermophile protein misfolding can be removed by subunit rearrangement caused by heating of the protein.[58] However, that is not very common and more often the problem of misfolding can be solved by using thermophiles as expression systems for thermophilic proteins can be a solution to the problem of misfolding.[28,57] These facts suggest that the protein expression system cellular environments need to be matched with those of protein source organisms in order to facilitate correct folding.

4.6. Study Limitations. Our study has multiple limitations which we should address here as well. First, it is evident that the number of protein structures which remain after all filtering procedures is very small for many of the species, including *T. thermophilus*. While we attempt to lower the impact of the low number of structures with resampling, it is still possible that the differences that we observe are related to specific protein structures or even by some errors in PDB experiments. In addition to the main results, described previously, we calculated JSD between matrices without BLASTp filtering. The rest of the procedures were kept the same as before. This way we could greatly increase the data size; however, the drawback is that proteomes now consisted of very different proteins and therefore these results cannot be fully conclusive either. Nevertheless, we found that JSD between *T. thermophilus* matrices is much higher than for all other species, the same as with the main results. We provide these additional results in the Supporting Information (Figures S1 and S2).

Second, our approach of using matrices in calculating JSD is double-edged. On one hand, using matrices allows us to compare entire proteomes rather than single proteins in a simple and computationally efficient way. This way we can compare proteomes which consist of very different proteins. On the other hand, only a pairwise comparison would show what effect protein type has on differences in folding. Ideally, all proteomes in our study should consist of the same proteins and in that case a pairwise comparison would be highly advantageous. Additionally, for the sake of simplicity and easier interpretation, our matrices were computed using one-to-one AA/SS pairing. This approach neglects potential effects that neighbor AA has on SS. It may be beneficial to use windows of several AA/SS to compute matrices in future research.

As we allowed AA sequences to differ by 5% during our BLASTp filtering step, the datafile AA sequences had some level of variation and this could have an effect on SS and JSD8/JSD3. We looked at how AA similarity affected JSD8 and JSD3, and there seems to be no relationship (Figures S3 and S4). *T. thermophilus* large JSD8 and JSD3 results are highly unlikely to be explained by AA differences; however, this is still possible due to the complex nature of the AA/SS relationship and this possibility should not be ignored completely.

We believe that obtaining more structural data would be essential in order to design a study which would not have the limitations that we discussed. This is especially the case with data of expression systems other than *E. coli*. Often predicted structures could be used when PDB does not have sufficient data; however, to our knowledge, no protein structure prediction model has an expression system as a feature.

5. CONCLUSION

In conclusion, our results show that thermophilic protein folding in mesophilic *E. coli* introduces significant changes on the structure level. Misfolding of thermophilic proteins grown using mesophilic hosts can lead to loss or change of protein functions which will harm both research and industrial applications. While there can be many possible explanations for the reasons of misfolding, it is important to study *T. thermophilus* and other extremophiles protein expression with protein source species as protein expression systems in order to minimize expression system effects. It is also evident that a much higher diversity of expression systems on PDB is essential for more thorough understanding of expression system effects on protein folding.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c04786.

Figures for additional analysis of proteomes without BLASTp filtering and links to raw secondary structure data and our code (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Xuhua Xia − *Department of Biology, University of Ottawa, Ottawa, Canada K1N 6N5; Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Canada K1N 6N5;* orcid.org/0000-0002-3092-7566; Email: xxia@uottawa.ca

Authors

Alibek Kruglikov − *Department of Biology, University of Ottawa, Ottawa, Canada K1N 6N5;* orcid.org/0000-0001-6074-8764

Yulong Wei − *Department of Biology, University of Ottawa, Ottawa, Canada K1N 6N5*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c04786

Author Contributions

A.K. carried out the experiment and wrote the manuscript with support from Y.W. and X.X.; X.X. supervised the project.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC, RGPIN/2018-03878) of Canada to X.X.

■ REFERENCES

(1) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12* (1), 7.

(2) Whisstock, J. C.; Lesk, A. M. Prediction of Protein Function from Protein Sequence and Structure. *Q. Rev. Biophys.* **2003**, *36* (3), 307−340.

(3) Brady, G. P.; Sharp, K. A. Entropy in Protein Folding and in Protein—Protein Interactions. *Curr. Opin. Struct. Biol.* **1997**, *7* (2), 215−221.

(4) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295* (2), 337−356.

(5) Zhang, Q. C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C. A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; et al. Structure-

Based Prediction of Protein-Protein Interactions on a Genome-Wide Scale. *Nature* **2012**, *490* (7421), 556−560.
(6) Wood, J. M.; Maibaum, J.; Rahuel, J.; Grütter, M. G.; Cohen, N.-C.; Rasetti, V.; Rüger, H.; Göschke, R.; Stutz, S.; Fuhrer, W.; et al. Structure-Based Design of Aliskiren, a Novel Orally Effective Renin Inhibitor. *Biochem. Biophys. Res. Commun.* **2003**, *308* (4), 698−705.
(7) Koga, N.; Tatsumi-Koga, R.; Liu, G.; Xiao, R.; Acton, T. B.; Montelione, G. T.; Baker, D. Principles for Designing Ideal Protein Structures. *Nature* **2012**, *491* (7423), 222.
(8) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. Computational Enzyme Design. *Angew. Chem., Int. Ed.* **2013**, *52* (22), 5700−5725.
(9) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(10) Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. In *Protein Crystallography*; Springer, 2017; pp 627−641. DOI: 10.1007/978-1-4939-7000-1_26.
(11) Wibmer, C. K.; Ayres, F.; Hermanus, T.; Madzivhandila, M.; Kgagudi, P.; Oosthuysen, B.; Lambson, B. E.; de Oliveira, T.; Vermeulen, M.; van der Berg, K.; Rossouw, T.; Boswell, M.; Ueckermann, V.; Meiring, S.; von Gottberg, A.; Cohen, C.; Morris, L.; Bhiman, J. N.; Moore, P. L. SARS-CoV-2 501Y.V2 Escapes Neutralization by South African COVID-19 Donor Plasma. *Nat. Med.* **2021**, *27* (4), 622−625.
(12) Khan, S. A.; Zia, K.; Ashraf, S.; Uddin, R.; Ul-Haq, Z. Identification of Chymotrypsin-like Protease Inhibitors of SARS-CoV-2 via Integrated Computational Approach. *J. Biomol. Struct. Dyn.* **2021**, *39* (7), 2607−2616.
(13) Khan, R. J.; Jha, R. K.; Amera, G. M.; Jain, M.; Singh, E.; Pathak, A.; Singh, R. P.; Muthukumaran, J.; Singh, A. K. Targeting SARS-CoV-2: A Systematic Drug Repurposing Approach to Identify Promising Inhibitors against 3C-like Proteinase and 2′-O-Ribose Methyltransferase. *J. Biomol. Struct. Dyn.* **2021**, *39* (8), 2679−2692.
(14) Coutard, B.; Valle, C.; de Lamballerie, X.; Canard, B.; Seidah, N. G.; Decroly, E. The Spike Glycoprotein of the New Coronavirus 2019-NCoV Contains a Furin-like Cleavage Site Absent in CoV of the Same Clade. *Antiviral Res.* **2020**, *176*, 104742.
(15) Schüler, A.; Bornberg-Bauer, E. Evolution of Protein Domain Repeats in Metazoa. *Mol. Biol. Evol.* **2016**, *33* (12), 3170−3182.
(16) Konaté, M. M.; Plata, G.; Park, J.; Usmanova, D. R.; Wang, H.; Vitkup, D. Molecular Function Limits Divergent Protein Evolution on Planetary Timescales. *eLife* **2019**, *8*, No. e39705.
(17) Sharir-Ivry, A.; Xia, Y. The Impact of Native State Switching on Protein Sequence Evolution. *Mol. Biol. Evol.* **2017**, *34* (6), 1378−1390.
(18) Harrington, L. B.; Jha, R. K.; Kern, T. L.; Schmidt, E. N.; Canales, G. M.; Finney, K. B.; Koppisch, A. T.; Strauss, C. E. M.; Fox, D. T. Rapid Thermostabilization of *Bacillus Thuringiensis* Serovar Konkukian 97−27 Dehydroshikimate Dehydratase through a Structure-Based Enzyme Design and Whole Cell Activity Assay. *ACS Synth. Biol.* **2017**, *6* (1), 120−129.
(19) Durairaj, D. R.; Shanmughavel, P. In Silico Drug Design of Thiolactomycin Derivatives Against Mtb-KasA Enzyme to Inhibit Multidrug Resistance Of *Mycobacterium Tuberculosis*. *Interdiscip. Sci.: Comput. Life Sci.* **2019**, *11* (2), 215−225.
(20) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577* (7792), 706−710.
(21) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46* (W1), W296−W303.
(22) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins: Struct., Funct., Bioinf.* **2009**, *77* (4), 778−795.
(23) Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved Protein Structure Prediction Using Predicted Interresidue Orientations. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (3), 1496−1503.
(24) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(25) Rosano, G. L.; Ceccarelli, E. A. Recombinant Protein Expression in *Escherichia Coli*: Advances and Challenges. *Front. Microbiol.* **2014**, *5*, 172.
(26) Sørensen, H. P.; Mortensen, K. K. Advanced Genetic Strategies for Recombinant Protein Expression in *Escherichia Coli*. *J. Biotechnol.* **2005**, *115* (2), 113−128.
(27) Kaur, J.; Kumar, A.; Kaur, J. Strategies for Optimization of Heterologous Protein Expression in E. Coli: Roadblocks and Reinforcements. *Int. J. Biol. Macromol.* **2018**, *106*, 803−822.
(28) Hidalgo, A.; Betancor, L.; Moreno, R.; Zafra, O.; Cava, F.; Fernández-Lafuente, R.; Guisán, J. M.; Berenguer, J. *Thermus Thermophilus* as a Cell Factory for the Production of a Thermophilic Mn-Dependent Catalase Which Fails To Be Synthesized in an Active Form in *Escherichia Coli*. *Appl. Environ. Microbiol.* **2004**, *70* (7), 3839−3844.
(29) López-López, O.; Cerdán, M.-E.; González-Siso, M.-I. *Thermus Thermophilus* as a Source of Thermostable Lipolytic Enzymes. *Microorganisms* **2015**, *3* (4), 792−808.
(30) Niehaus, F.; Bertoldo, C.; Kähler, M.; Antranikian, G. Extremophiles as a Source of Novel Enzymes for Industrial Application. *Appl. Microbiol. Biotechnol.* **1999**, *51* (6), 711−729.
(31) Krefft, D.; Papkov, A.; Zylicz-Stachula, A.; Skowron, P. M. Thermostable Proteins Bioprocesses: The Activity of Restriction Endonuclease-Methyltransferase from *Thermus Thermophilus* (RM.TthHB27I) Cloned in *Escherichia Coli* Is Critically Affected by the Codon Composition of the Synthetic Gene. *PLoS One* **2017**, *12* (10), e0186633.
(32) Soulages, J. L.; Kim, K.; Walters, C.; Cushman, J. C. Temperature-Induced Extended Helix/Random Coil Transitions in a Group 1 Late Embryogenesis-Abundant Protein from Soybean. *Plant Physiol.* **2002**, *128* (3), 822−832.
(33) Kim, Y.; Bigelow, L.; Borovilos, M.; Dementieva, I.; Duggan, E.; eschenfeldt, W.; Hatzos, C.; Joachimiak, G.; Li, H.; Maltseva, N.; Mulligan, R.; Quartey, P.; Sather, A.; Stols, L.; Volkart, L.; Wu, R.; Zhou, M.; Joachimiak, A. High-Throughput Protein Purification for X-Ray Crystallography and NMR. In *Advances in Protein Chemistry and Structural Biology*; Joachimiak, A., Ed.; Structural Genomics, Part A; Academic Press, 2008; Vol. 75, pp 85−105. DOI: 10.1016/S0065-3233(07)75003-9.
(34) Xia, X. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Mol. Biol. Evol.* **2018**, *35* (6), 1550−1552.
(35) Schlee, S.; Reinstein, J. The DnaK/ClpB Chaperone System from *Thermus Thermophilus*. *Cell. Mol. Life Sci.* **2002**, *59* (10), 1598−1606.
(36) Godin-Roulling, A.; Schmidpeter, P. A.; Schmid, F. X.; Feller, G. Functional Adaptations of the Bacterial Chaperone Trigger Factor to Extreme Environmental Temperatures. *Environ. Microbiol.* **2015**, *17* (7), 2407−2420.
(37) Gsponer, J.; Futschik, M. E.; Teichmann, S. A.; Babu, M. M. Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science* **2008**, *322* (5906), 1365−1368.

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806

(38) Sahdev, S.; Khattar, S. K.; Saini, K. S. Production of Active Eukaryotic Proteins through Bacterial Expression Systems: A Review of the Existing Biotechnology Strategies. *Mol. Cell. Biochem.* **2007**, *307* (1), 249−264.
(39) Rohl, C. A.; Doig, A. J. Models for the 310-helix/Coil, Π-helix/Coil, and A-helix/310-helix/Coil Transitions in Isolated Peptides. *Protein Sci.* **1996**, *5* (8), 1687−1696.
(40) Mikol, V.; Giegé, R. Phase Diagram of a Crystalline Protein: Determination of the Solubility of Concanavalin A by a Micro-quantitation Assay. *J. Cryst. Growth* **1989**, *97* (2), 324−332.
(41) Chayen, N.; Akins, J.; Campbell-Smith, S.; Blow, D. M. Solubility of Glucose Isomerase in Ammonium Sulphate Solutions. *J. Cryst. Growth* **1988**, *90* (1−3), 112−116.
(42) McPherson, A. [7] Crystallization of Proteins by Variation of PH or Temperature. In *Methods in Enzymology*; Elsevier, 1985; Vol. *114*, pp 125−127. DOI: 10.1016/0076-6879(85)14009-7.
(43) Plotkin, J. B.; Kudla, G. Synonymous but Not the Same: The Causes and Consequences of Codon Bias. *Nat. Rev. Genet.* **2011**, *12* (1), 32−42.
(44) Zylicz-Stachula, A.; Zolnierkiewicz, O.; Sliwinska, K.; Jezewska-Frackowiak, J.; Skowron, P. M. Modified 'One Amino Acid-One Codon' Engineering of High GC Content TaqII-Coding Gene from Thermophilic *Thermus Aquaticus* Results in Radical Expression Increase. *Microb. Cell Fact.* **2014**, *13* (1), 7.
(45) Nedialkova, D. D.; Leidel, S. A. Optimization of Codon Translation Rates via TRNA Modifications Maintains Proteome Integrity. *Cell* **2015**, *161* (7), 1606−1618.
(46) Ellis, R. J. Macromolecular Crowding: An Important but Neglected Aspect of the Intracellular Environment. *Curr. Opin. Struct. Biol.* **2001**, *11* (4), 500.
(47) Kinjo, A. R.; Takada, S. Effects of Macromolecular Crowding on Protein Folding and Aggregation Studied by Density Functional Theory: Dynamics. *Phys. Rev. E* **2002**, *66* (5), 051902.
(48) Miklos, A. C.; Sarkar, M.; Wang, Y.; Pielak, G. J. Protein Crowding Tunes Protein Stability. *J. Am. Chem. Soc.* **2011**, *133* (18), 7116−7120.
(49) Kim, J. S.; Yethiraj, A. Crowding Effects on Protein Association: Effect of Interactions between Crowding Agents. *J. Phys. Chem. B* **2011**, *115* (2), 347−353.
(50) Samiotakis, A.; Cheung, M. S. Folding Dynamics of Trp-Cage in the Presence of Chemical Interference and Macromolecular Crowding. I. *J. Chem. Phys.* **2011**, *135* (17), 175101.
(51) Kuznetsova, I. M.; Turoverov, K. K.; Uversky, V. N. What Macromolecular Crowding Can Do to a Protein. *Int. J. Mol. Sci.* **2014**, *15* (12), 23090−23140.
(52) Westphal, A. H.; Geerke-Volmer, A. A.; van Mierlo, C. P. M.; van Berkel, W. J. H. Chaotropic Heat Treatment Resolves Native-like Aggregation of a Heterologously Produced Hyperthermostable Laminarinase. *Biotechnol. J.* **2017**, *12* (6), 1700007.
(53) Ninh, P. H.; Honda, K.; Sakai, T.; Okano, K.; Ohtake, H. Assembly and Multiple Gene Expression of Thermophilic Enzymes in *Escherichia Coli* for in Vitro Metabolic Engineering. *Biotechnol. Bioeng.* **2015**, *112* (1), 189−196.
(54) Hartl, F. U.; Bracher, A.; Hayer-Hartl, M. Molecular Chaperones in Protein Folding and Proteostasis. *Nature* **2011**, *475* (7356), 324−332.
(55) Berg, J. M.; Tymoczko, J. L.; Stryer, L. Protein Structure and Function. *Biochemistry*, 5th ed.; W. H. Freeman, 2002.
(56) Winklhofer, K. F.; Tatzelt, J.; Haass, C. The Two Faces of Protein Misfolding: Gain- and Loss-of-Function in Neurodegenerative Diseases. *EMBO J.* **2008**, *27* (2), 336−349.
(57) Fujino, Y.; Goda, S.; Suematsu, Y.; Doi, K. Development of a New Gene Expression Vector for *Thermus Thermophilus* Using a Silica-Inducible Promoter. *Microb. Cell Fact.* **2020**, *19* (1), 126.
(58) Goda, S.; Kojima, M.; Nishikawa, Y.; Kujo, C.; Kawakami, R.; Kuramitsu, S.; Sakuraba, H.; Hiragi, Y.; Ohshima, T. Intersubunit Interaction Induced by Subunit Rearrangement Is Essential for the Catalytic Activity of the Hyperthermophilic Glutamate Dehydrogenase from Pyrobaculum Islandicum. *Biochemistry* **2005**, *44* (46), 15304−15313.
(59) Angelov, A.; Mientus, M.; Liebl, S.; Liebl, W. A Two-Host Fosmid System for Functional Screening of (Meta)Genomic Libraries from Extreme Thermophiles. *Syst. Appl. Microbiol.* **2009**, *32* (3), 177−185.
(60) Angelov, A.; Loderer, C.; Pompei, S.; Liebl, W. Novel Family of Carbohydrate-Binding Modules Revealed by the Genome Sequence of *Spirochaeta Thermophila* DSM 6192. *Appl. Environ. Microbiol.* **2011**, *77* (15), 5483−5489.
(61) Leis, B.; Angelov, A.; Mientus, M.; Li, H.; Pham, V. T. T.; Lauinger, B.; Bongen, P.; Pietruszka, J.; Gonçalves, L. G.; Santos, H.; Liebl, W. Identification of Novel Esterase-Active Enzymes from Hot Environments by Use of the Host Bacterium *Thermus Thermophilus*. *Front. Microbiol.* **2015**, *6*, 275.
(62) Chautard, H.; Blas-Galindo, E.; Menguy, T.; Grand'Moursel, L.; Cava, F.; Berenguer, J.; Delcourt, M. An Activity-Independent Selection System of Thermostable Protein Variants. *Nat. Methods* **2007**, *4* (11), 919−922.
(63) Mate, D. M.; Rivera, N. R.; Sanchez-Freire, E.; Ayala, J. A.; Berenguer, J.; Hidalgo, A. Thermostability Enhancement of the *Pseudomonas Fluorescens* Esterase I by in Vivo Folding Selection in *Thermus Thermophilus*. *Biotechnol. Bioeng.* **2020**, *117* (1), 30−38.
(64) Bosch, S.; Sanchez-Freire, E.; del Pozo, M. L.; Ćesnik, M.; Quesada, J.; Mate, D. M.; Hernández, K.; Qi, Y.; Clapés, P.; Vasić-Rački, Đ.; Findrik Blažević, Z.; Berenguer, J.; Hidalgo, A. Thermostability Engineering of a Class II Pyruvate Aldolase from *Escherichia Coli* by in Vivo Folding Interference. *ACS Sustainable Chem. Eng.* **2021**, *9* (15), 5430−5436.
(65) Miotto, M.; Olimpieri, P. P.; Di Rienzo, L.; Ambrosetti, F.; Corsi, P.; Lepore, R.; Tartaglia, G. G.; Milanetti, E. Insights on Protein Thermal Stability: A Graph Representation of Molecular Interactions. *Bioinformatics* **2019**, *35* (15), 2569−2577.
(66) Vogt, G.; Argos, P. Protein Thermal Stability: Hydrogen Bonds or Internal Packing? *Folding Des.* **1997**, *2*, S40−S46.

https://doi.org/10.1021/acsomega.2c04786
ACS Omega 2022, 7, 37797−37806