Structured proteins are abundant in unevolved sequence space

Vyacheslav Tretyachenko[1,2], Jiří Vymětal[3], Tereza Neuwirthová[1#], Jiří Vondrášek[3], Kosuke Fujishima[4,5] and Klára Hlouchová*[1,3]

[1] Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague, 12843, Czech Republic

[2] Department of Biochemistry, Faculty of Science, Charles University, Prague, 12843, Czech Republic

[3] Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, 16610, Czech Republic

[4] Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, 1528550, Japan

[5] Graduate School of Media and Governance, Keio University, Fujisawa, Japan, 2520882

[#] Current address: R&D Informatics Solutions, MSD Czech Republic s.r.o., Prague, Czech Republic

* To whom correspondence may be addressed: klara.hlouchova@natur.cuni.cz

Abstract

Natural proteins represent numerous but tiny structure/function islands in a vast ocean of possible protein sequences not challenged by biological evolution and are yet to be explored by research. Recent studies have suggested this uncharted sequence space endows a surprisingly high structural propensity but understanding of this phenomenon has been awaiting a systematic high-throughput approach.

Here we designed, prepared, and characterized two combinatorial protein libraries consisting of randomized proteins, each 105 residues in length. The first library constructed proteins from the entire canonical alphabet of 20 amino acids. The second library used a subset of only 10 residues (A,S,D,G,L,I,P,T,E,V) that represent a consensus view of plausibly available amino acids through prebiotic chemistry. Based on libraries *in silico* analyses and bulk protease resistance/solubility screening, we report that both canonical and "early" proteins have a similar structure content. While the inherent solubility of the early library is higher than that of the canonical library, only the latter can be increased by chaperone supplementation. On the contrary, we hypothesize that the early library solubility and folding is enabled by salts and cofactors in the cell-like milieu where these assays were performed. While the early library proteins are inherently more thermostable, stability of both libraries can be elevated by chaperone activity. Interestingly, their structure content remains unchanged. These observations suggest that chaperones play a more significant role with the fully evolved alphabet.

In conclusion, our study shows that compact structure occurrence (i) is (up to 40%) abundant in random sequence space, (ii) independent of the general Hsp60 chaperone system activity, and (iii) is not granted solely by the "late" and complex amino acid additions.

Keywords

Protein sequence space, protein structure, amino acid alphabet, genetic code evolution, random sequence space

Introduction

Today's biological systems are anchored in the universal genetic coding apparatus, relying on coded amino acids that were likely selected in the first 10-15% of Earth's history [1]. While sources of prebiotic organic material provided a broad selection of amino acids, only about half of the canonical amino acids were detected in this pool [2]. There is substantial evidence that this set formed an early version of the genetic code and that the "late" amino acids were recruited only after an early metabolism was in existence. The boundary between these two sets is blurry. However, large meta-analyses of these studies agree that "early", i.e. the smaller and less complex amino acids (Gly, Ala, Asp, Glu, Val, Ser, Ile, Leu, Pro, Thr) were a fixture in the genetic code before its evolution [3,4].

The factors which drove the selection of 20 coded amino acids remain puzzling. Solubility, ease of biosynthesis, racemization stability, un/reactivity with tRNA and potential peptide product stability seem to explain some "choices" but not others [5,6] . Most recently, analysis of the *set* of amino acids revealed that the canonical alphabet shows an unusually good repertoire of the chemical property space when compared to plausible alternatives [7,8]. Such studies lead to speculations that similar amino acid selection would be expected on other Earth-like planets [5,8,9].

In extant proteins, significant part of the "late" amino acids (Arg, Lys, His, Cys, Trp and Tyr) belong to the essential catalytic residues, i.e. they are associated with catalysis in almost all of the enzyme classes [10]. At the same time, the putatively early amino acids have been related to protein disorder and lack of 3D structure [11]. However, scarce sampling of random sequences composed of early amino acids suggests that such proteins have a higher solubility than the full canonical alphabet [12,13]. Moreover, computational and experimental mutational studies removing or reducing the late amino acids in selected proteins imply that the early amino acids comprise a non-zero folding potential [14–18]. If prone to tertiary structure formation, it has been hypothesized that the early alphabet could more probably form molten globules rather than tightly packed structures, mainly due to the lack of aromatic and positively charged amino acids. According to this hypothesis, the addition of late amino acids would be required to increase protein stability and catalytic activity [11,17,19]. Interestingly, it was shown that while positively charged amino acids are more compatible with protein folding, they also promote protein aggregation if their position within the sequence is not optimized or assisted by molecular chaperones. Thus it was hypothesized that chaperone emergence coincided with the incorporation of basic residues into the amino acid alphabet leading to the increase in the plasticity of natural folding space [20].

To assess the intrinsic structural and functional properties of the full amino acid alphabet, semi high-throughput studies using combinatorial sequence libraries have been performed previously [21–25]. Surprisingly, secondary structure occurrence in random sequence libraries has been recorded with similar frequency as in biological proteins, while folding (or more precisely, occurrence of collapsed conformations) has been reported in up to 20% of tested proteins [21,24,25]. However, more systematic and high-throughput screening is still awaited to confirm these observations. Moreover, it remains unclear how much these properties are a result of the full alphabet fine-tuning, whether structured molecules emerge spontaneously and independently in the canonical amino acid sequence space and/or if the early amino acids could provide similar structural traits.

To fill this knowledge gap, we characterized libraries of 10^{12} randomized protein sequences from the full and early amino acid alphabets to assess their collective biochemical characteristics. While the bioinformatic prediction revealed similar secondary structure potential in both libraries and lower aggregation propensity of the full alphabet, the early alphabet is significantly more soluble and thermostable under cell-like experimental conditions. The full alphabet sequences were found to interact with molecular chaperones that can compensate for their otherwise poor solubility. Up to ~40% folding occurrence is observed in both studied libraries. The results therefore agree with previous scarce sampling observations and in addition, the folding frequency and inducibility of some properties in a cell-like environment are systematically mapped. Moreover, this study provides a unique synthetic biology pipeline that could be used to survey properties of any other protein alphabets associated with different biological phenomena of interest.

Results

Library expression and quality control

The combinatorial protein libraries studied in this work consisted of 105 amino acid long proteins with an 84 amino acid long variable parts, FLAG/HIS tag sequences on N'/C' ends and the thrombin cleavage site in the middle of the protein construct (Supplementary Fig. S1). The variable region was designed by the CoLiDe algorithm and consisted of a specific set of degenerate codons in order to match the natural canonical (full alphabet, 20F) and prebiotically plausible (A,S,D,G,L,I,P,T,E,V; early alphabet, 10E) amino acid distributions (Supplementary Table S1) [26]. The amino acid ratios for both libraries corresponded to natural amino acid distribution from the UniProt database [27]. The libraries were assembled from two overlapping oligonucleotides, transcribed into their corresponding mRNA and translated using an in vitro translation system (Supplementary Fig. S2). In order to verify the designed library variability and amino acid distribution, we sequenced the assembled degenerate oligonucleotide DNA library and performed a mass spectrometric analysis of the purified library protein product. The root mean squared error (RMSE) from the target amino acid distribution was ~0.06 in both libraries 20F and 10E (Supplementary Table S2, Supplementary Fig. S3). The variability analysis of the sequenced library showed that 96% of sequences were unique, no significant sequence enrichment was observed (Fig. 1, Supplementary Table S3). Due to the synthesis errors, STOP codons were introduced into 12% of the library sequences. However, their products were not observed in the following western blot protein analyses (Supplementary Fig. 5/7/9). The variability of the purified protein product was validated by MALDI-TOF mass spectrometry, the mean and spread of the experimental spectra closely matching the predicted distributions (Supplementary Fig. S4).

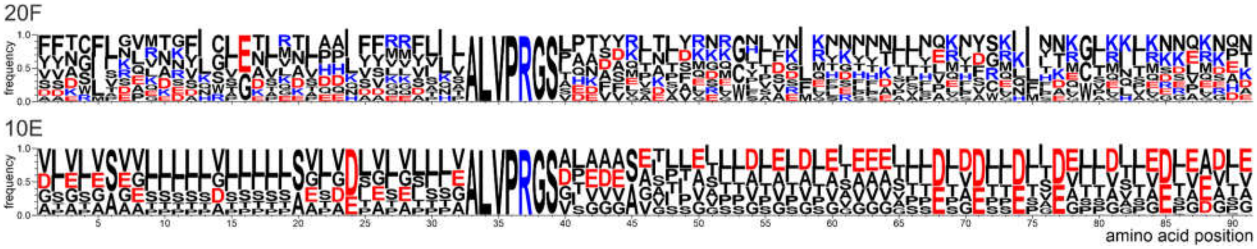

Figure 1. Sequence logo representation of full (top) and early (bottom) alphabet libraries variability constructed from the corresponding sequenced DNA templates.

Secondary structure and aggregation propensity predictions

Sequences of both 20F and 10E libraries acquired by the high throughput sequencing were analyzed by a consensus protein secondary structure prediction [28]. 200 000 sequences were analyzed from each library. Interestingly, despite the different amino acid distributions, comparable alpha helix and beta sheet forming tendencies were reported in both libraries with only a slight increase in alpha helical content in the 20F library (33 % vs. 30% in 10E) (Fig. 2A). The overall alpha helical and beta sheet content correlate well among the individual predictors used for both studied libraries which is not necessarily the case for other alternative and more artificial alphabets (unpublished observation). The prediction of aggregation propensity of the same set of sequences indicated significantly higher aggregation tendency of 10E library proteins in comparison to 20F library (Fig. 2B).

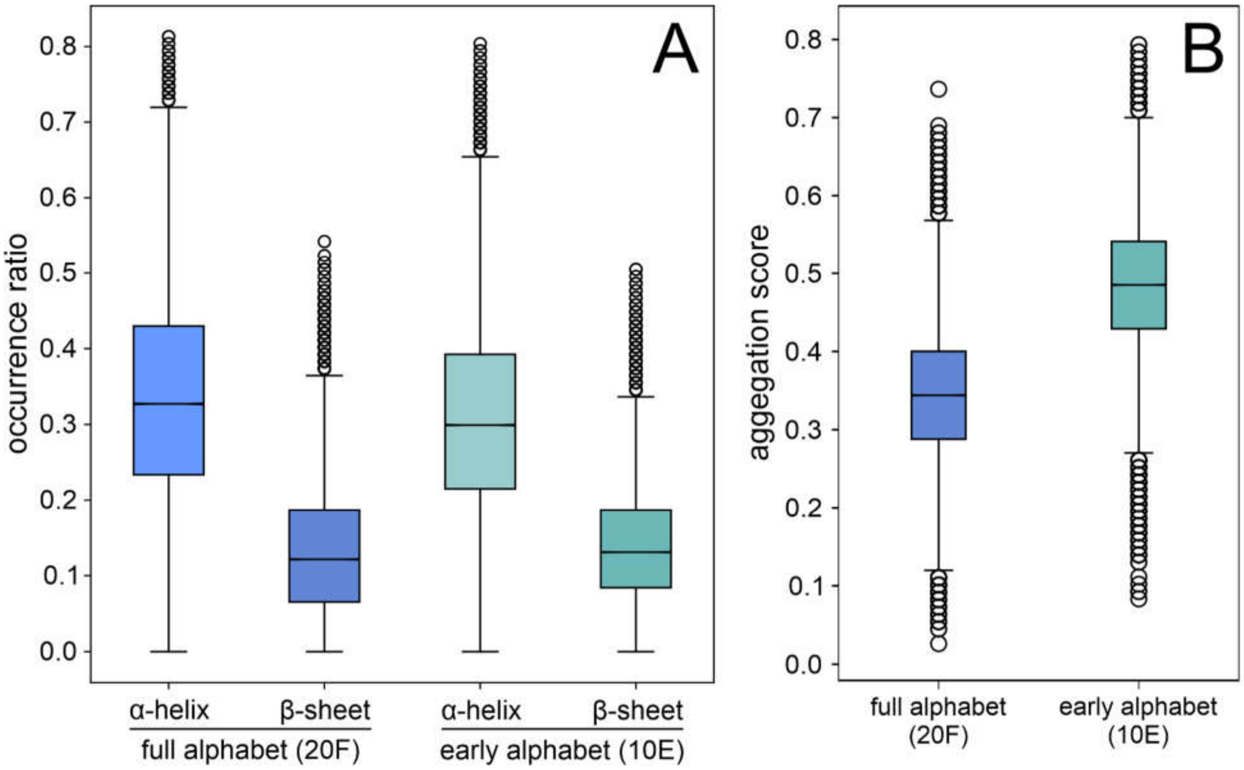

Figure 2. Bioinformatic prediction of alpha helical and beta sheet content (A) and aggregation propensity (B) of a sample of 200 000 sequences acquired by high throughput sequencing of the early (green) and full (blue) alphabet libraries DNA templates. Aggregation score is defined as ratio of predicted aggregation-prone residues per sequence

Expression and solubility analysis in the absence and presence of DnaK chaperone system

To systematically assess the expression profiles of the libraries, a quantitative western blot analysis was performed with the library products expressed at different temperatures (25, 30 and 37 °C) and with/without DnaK/DnaJ/GrpE chaperone system supplementation (further referred as to DnaK). The analysis was carried out in triplicates and western blot signals of both total expression and soluble fractions were quantified with ImageJ [29]. For both 20F and 10E libraries the expression yields grow with the increasing temperature with the overall yield being mildly lower in the chaperone supplemented reactions at 37 °C (Fig. 3). In the case of the 20F library, the solubility of the library is relatively poor but is significantly improved by chaperone supplementation. While in the chaperone supplemented reaction the soluble fraction grew with expression temperature proportionally with the total expression, in the chaperone absent condition, the soluble fraction yields did not significantly change with the transition from 30 to 37 °C. On the other hand, chaperone supplementation did not have any sound effect on the 10E library expression and solubility (Fig. 3).

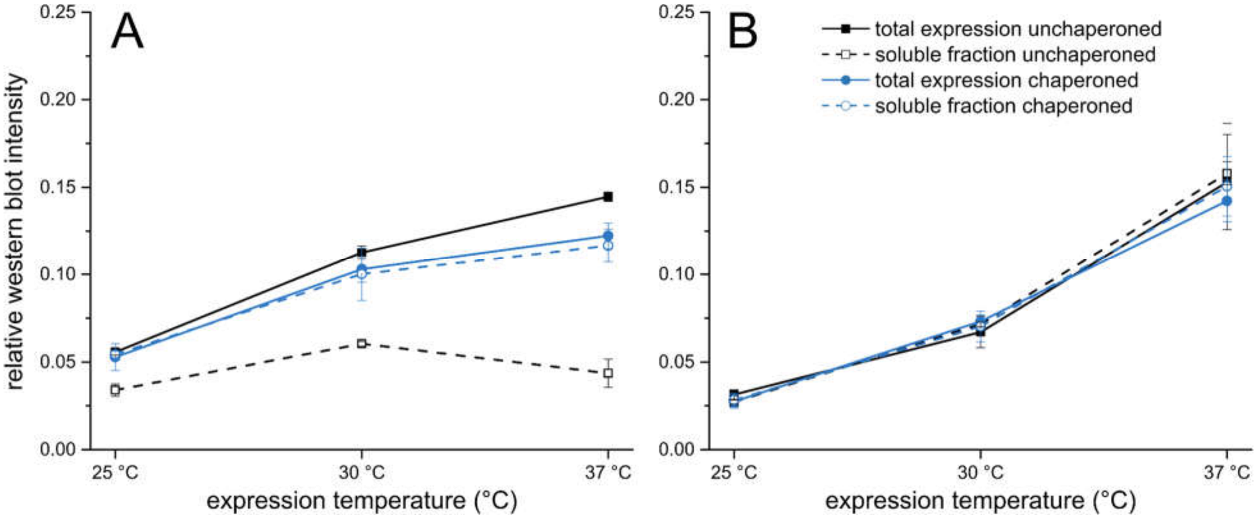

Figure 3. A summary of expression and solubility analysis of the full (A) and early (B) alphabet libraries at three different temperatures. Total expression (solid line) and soluble fraction (dashed line) were compared in chaperoned (blue line) and unchaperoned (black line) conditions. For original data see Supplementary Fig. S5/S6 and Supplementary Table S4.

Assessment of proteolytic resistance

The structural potential of random protein libraries was assessed by proteolysis. The digestion was performed in triplicates by Lon and thrombin proteases in co-translational and post-translational conditions, respectively (Fig. 4). The Lon protease is a part of the E. coli protein misfolding system and known to specifically digest unfolded proteins in exposed hydrophobic regions [30]. Here we adapted a previously published protocol on single protein structure assessment for combinatorial library characterization [31]. The method is used to separate and quantify distinct protease sensitive parts of the library within both the soluble and insoluble fractions of the expressed libraries.The thrombin protease assay was adapted from the study of Chiarabelli et al, where the structure occurrence is derived from the cleaved/uncleaved ratio of proteins with an engineered thrombin cleavage site situated in the middle of the sequence [21]. The unstructured proteins are expected to be quickly degraded on the exposed cleavage site.

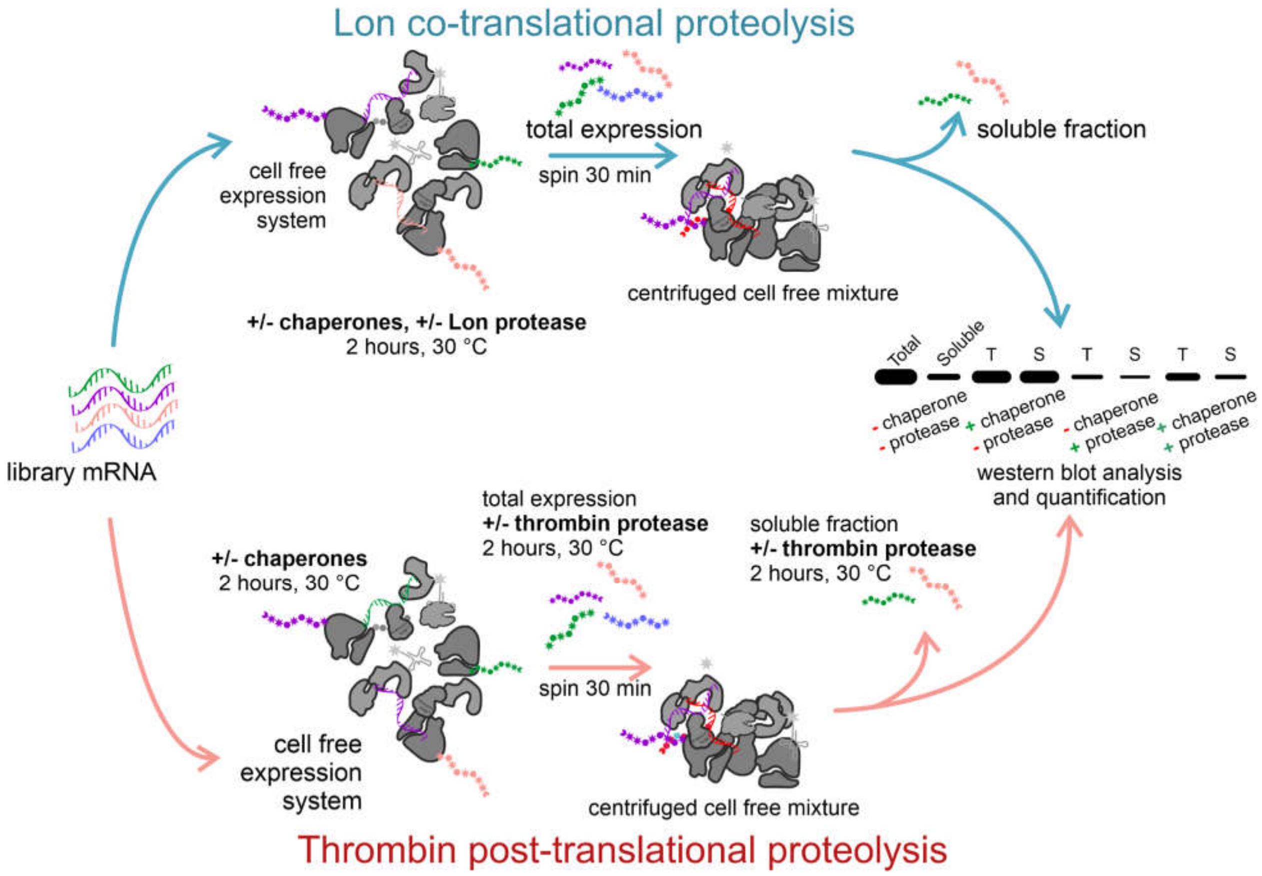

Figure 4. Scheme of the proteolytic resistance experimental pipeline. In co-translational proteolytic assay (top) the Lon protease is present during the cell-free expression, in post-translational proteolytic assay (bottom) thrombin protease is added to the separated total and soluble fraction of the expressed library after the translation is quenched by puromycin addition

According to the 20F library analysis, the soluble/undegradable structured proteins represent ~30-40% of the total product (Fig. 5A). Upon the addition of DnaK chaperone, most of the library solubilizes but the structured content does not increase significantly and occupies ~40-50% of the total product. In comparison, chaperone addition does not have any impact on the solubility or structure content in the 10E library (Fig. 5B). Interestingly, the structured content (soluble undegradable) in the 10E library is similar as in 20F library after the addition of chaperones.

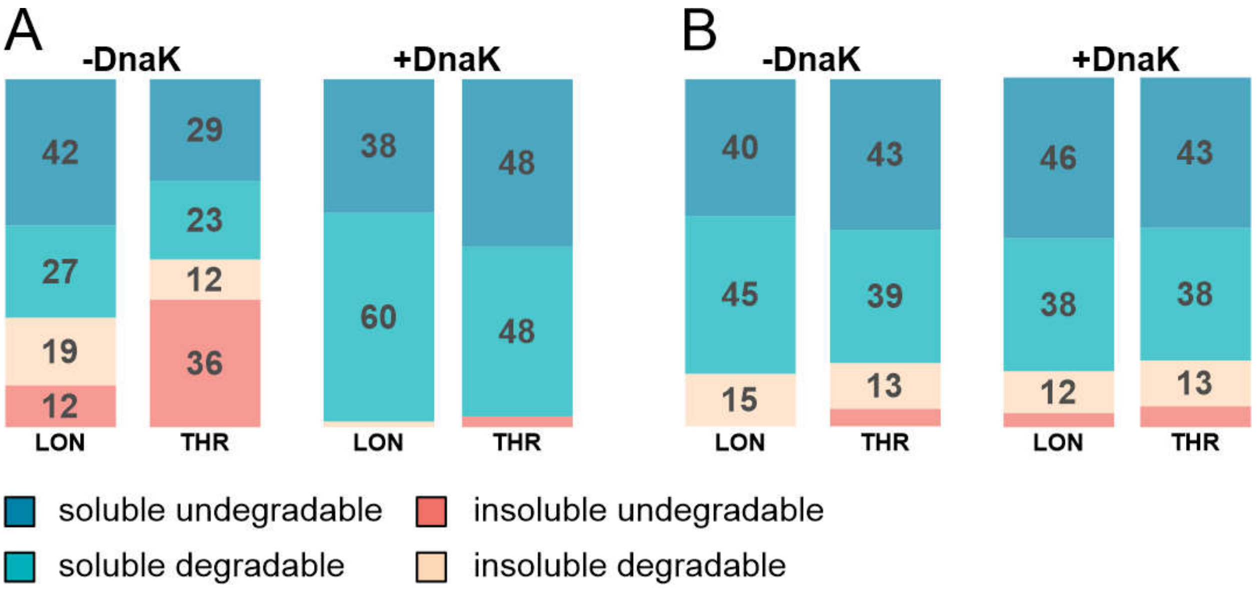

Figure 5. An integrated solubility/proteolysis resistance analysis of the full (A) and early (B) alphabet libraries. Libraries were expressed either in the absence (left double column) or presence (right double column) of the DnaK chaperone system. Proteolysis was performed by protease Lon (left columns) in a co-translational regime or by thrombin protease (right columns) in a post-translational mode. Values in the boxes represent the percentage ratios of the total expressed library per fraction. For original data see Supplementary Fig. S7/S8/S9/S10 and Supplementary Table S5/S6.

Thermostability characterization

Following the expression, solubility and structural content assessment, we analyzed the temperature sensitivity of the 20F and 10E proteins. The libraries expressed with and without chaperone supplementation were subjected to 15 minutes/42 °C heat shock. The aggregated fraction was removed by centrifugation and the soluble fraction was compared with and without thrombin treatment (Fig. 6).

The 10E library is intrinsically more thermostable than 20F (~60 vs ~30% of the libraries remain soluble after the heat shock, respectively) while the DnaK chaperone system induces the thermostability of both. The protease resistant fraction of the soluble part of libraries remains the same (~40%) as before the heat shock treatment apart from unchaperoned 20F library which demonstrates slight decrease in both soluble and degradation resistant fraction (Fig. 6).

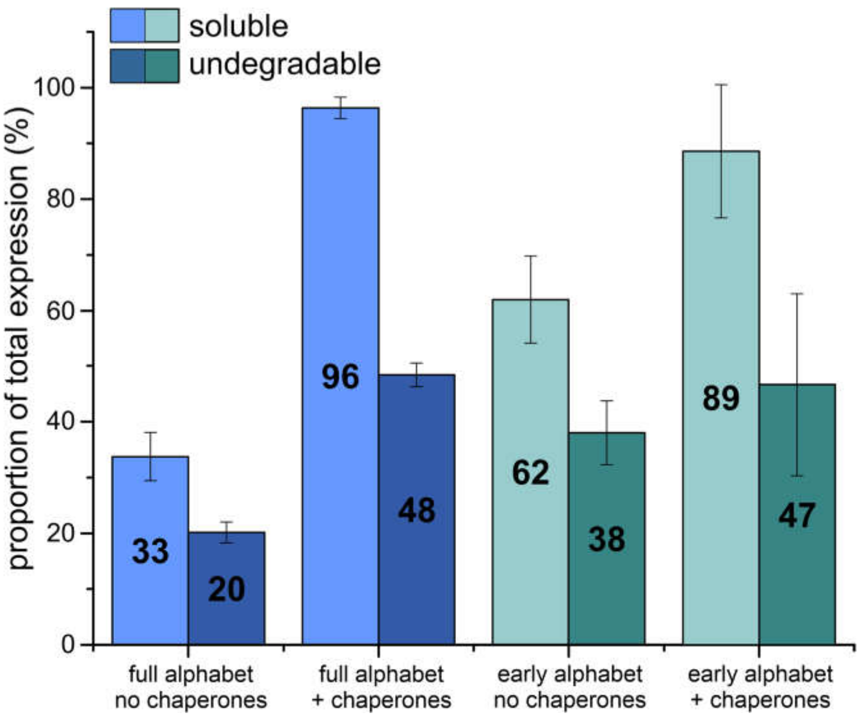

Figure 6. Thermostability analysis showing soluble proportions (light blue and green) of the total expression of the full and early alphabet libraries after a heat shock (42 °C / 15 min) treatment and their respective thrombin resistant proportions (dark blue and green) of the total expression in unchaperoned and chaperoned conditions. Numbers in the bars represent the percentage fraction of the total expressed library. For original data see Supplementary Fig. S9/S10 and Supplementary Table S6.

Discussion

In this study, a high-throughput systematic approach was used to experimentally analyze the structural properties of the vast protein sequence space. Random sequences have been proposed as proxies for both (i) precursors of *de novo* emerged proteins in current evolution as well as (ii) sources of peptide/protein birth at the earliest stages of life preceding templated proteosynthesis [32,33]. However, the structural properties of random sequences have so far remained uncomprehended while a few recent bioinformatic and coarse-grained studies have pointed to their surprising properties, such as high secondary structure propensity and *in vivo* tolerance [24,25,34]. Here, two combinatorial protein libraries encompassing upto 10^{12} individual sequences from two distinct alphabets (representing hypothetical stages of genetic code evolution) have been characterized.

Solubility of the natural alphabet random proteins can be induced by chaperones

The first “full” alphabet library is based on the amino acid composition of the Uniprot database reflecting on the properties of today’s proteomes. It has been previously shown that similar constructs have limited solubility but a similar secondary structure potential to biological proteins [12,13,25]. Our study confirms these results and in addition, we specify that 20-50% of the overall diverse library appears in the soluble fraction in the 30-37 °C temperature range. While previous studies of similar construct size evaluated the solubility of individual proteins that were overexpressed (many of them with partial solubility) at different *E. coli* strains and under different conditions, our library was expressed using a reconstituted cell-free protein synthesis (CFPS) system and its large diversity (contrasting with overexpression of individual proteins) was confirmed by MALDI. Therefore, we cannot make direct comparison to previous studies of individual proteins but rather report the "fingerprint" properties of the full alphabet domain-size proteins.

Interestingly, this library of unevolved sequences was observed to interact productively with the natural molecular chaperone system DnaK/DnaJ/GrpE which was supplemented to the CFPS in another experiment. This interaction caused almost total solubilization of the otherwise insoluble proteins over the studied temperature range. While the solubility traits may be quite different for much shorter polymer lengths, our previous study showed that random domain-size sequences cope with significant aggregation, especially if they are rich in secondary structure content [25]. To characterize the library folding potential without introducing potential bias, we used an *in situ* double proteolysis experiment adapting two previously reported approaches [21,31]. The experiment combined co-translational proteolysis by disorder-specific Lon protease and a post-translational cleavage by thrombin designed to cut a potentially exposed cleavage site engineered in the center of random proteins. Besides the increased robustness of the structure content estimation, such a combined approach brings a unique insight into the library translation dynamics.

The double proteolysis experiment revealed that ~30-40% of library 20F proteins are protease resistant during proteolysis. Upon the addition of chaperones (which solubilizes the library as described above), the ratio of protease resistant species rose only mildly to ~40-50%. The preferentially unstructured nature of the full alphabet library echoes the nature of naturally evolved *de novo* proteins, i.e. proteins that emerge in current biology from previously non-coding DNA (summarized in [35]).

Overall, these results show that while inherent protein solubility is limited in random sequence space made of the natural alphabet, it can be induced significantly by the activity of molecular chaperones. At the same time, DnaK chaperone system has only a minor effect on structure formation, suggesting that the majority of the potentially solubilized sequences are devoid of tertiary structure arrangements. Nevertheless, the ~40% natural abundance of soluble and yet protease-resistant sequences in unevolved sequence space may be surprising in the light of earlier hypotheses and exceeds the estimates of folding frequency reported by previous coarse-grained studies [21,36]. Nevertheless, major differences in the experimental setups (cell-free vs cell-based expression, low-level vs overexpression, high- vs. low-throughput methodology, library design and sequence length) prevent the possibility of direct comparisons among these studies. A direct comparison of the full library properties can however be made with another library of proteins studied here under the same experimental conditions.

Structure formation is comparable in proteins from the full canonical alphabet and its early subset, unaffected by chaperones

Second "early" alphabet library was constructed from a 10 amino acid subset of the full alphabet which was proposed to constitute an earlier version of the genetic code and be reflected in the composition of early proteins [3]. We would first like to emphasize here that by this study, we do not try to establish that there was necessarily a time in life's evolution where domain-size proteins were composed entirely of this amino acid subset. Our analysis rather deals with the inherent physico-chemical properties of such an alphabet, were it to form or dominate protein-like structures. We also acknowledge that the earliest stages of peptide/protein formation (preceding templated proteosynthesis and perhaps also its early less specific versions) probably utilized a whole plethora of prebiotically plausible amino acids or similar chemical entities but inclusion of such non-canonical amino acids in the studied alphabets is beyond the scope of this study [1,37,38].

Although the overall secondary structure propensity of the early alphabet is comparable to the full alphabet according to the bioinformatic prediction, the occurrence of alpha-helix is mildly (~3%) lower. While these differences are statistically borderline, they may have interesting implications for the evolution of protein structural properties. Brack&Orgel proposed that beta-sheet structures were prebiotically significant and later significance of alpha-helices in protein folds was also recently implied by the structural analysis of the ribosomal protein content showing that the most ancient protein protein fragments of this molecular fossil are mostly disordered and of beta-sheet formation [39–41]. Despite the similar secondary structure propensities of the full and early alphabets, the 10E library proteins are significantly more soluble (~90%) upon expression. They retain similar solubilities in chaperoned/unchaperoned condition unlike the 20F library proteins. This observation supports the previously stated hypothesis of chaperone co-evolution with the incorporation of the first positively charged amino acids into the early amino acid alphabets [20].
A significantly higher solubility of the 10E library proteins (and similar protein compositions) is in an agreement with previous studies [12,13]. This phenomenon could be related to the lower complexity of 10E library proteins given by the more compact amino acid alphabet. While 20F proteins represent a highly variable sample of protein folding space with plenty of opportunities for aggregation initiation, the 10E proteins display a narrower subspace with much more uniform sequence and physicochemical characteristics distributions. In addition, their overall negative charge and absence of positively charged/aromatic amino acids are conditions which were previously shown to suppress both nonspecific aggregations as well as independent protein folding formation [20]. At the same time though, the 10E alphabet contains a significant proportion of hydrophobic amino acids. Using the ProA bioinformatic predictor of protein aggregations, the 10E library would be expected to be intrinsically less soluble, contradicting our and previous empirical observations. However, contrasting with the intrinsic behavior of the protein alone, our (and previous experimental) assays were performed in a cell-like environment, rich in different salts and other small molecules/cofactors.

Interestingly, the 10E library also displays a significant amount of tertiary structure formation. In the absence of chaperones, the ratio of the protease resistant fraction is 40-50% in both the co- and post-translational digestion assay, i.e. similar to the 20F protease resistant fraction when supplemented with chaperones.

Such a high occurrence of structure formation within the 10E library is non-intuitive and unexpected purely from its amino acid composition. However, several folders have been recently identified from the same or similar protein composition in experiments reducing extant protein compositions [15,16,18,42,43]. Where characterized in more detail, assistance of salts, metal ions or cofactor binding were found to explain the folding properties [15,18,42,44]. In addition, Despotovic et al. recently confirmed that folded conformations of a highly acidic 60-residue protein can be induced by positively charged counterions, in case of Mg2+ corresponding roughly to its concentration in the CFPS reaction (~10mM) [45]. These studies allow us to hypothesize that the high structural propensity of the 10E alphabet could result from the cation/cofactor-rich environment, where the lack of hydrophobic and electrostatic interactions is compensated by these chemical entities. Alternatively or concurrently, the library solubility and protease resistance could be partly explained by tertiary structure formation induced by oligomerization as previously hypothesized by Yadid et al. in a study using 100 amino acid long fragments (albeit from different amino acid composition) [46]. Our study presented here cannot unambiguously differentiate between these two possible scenarios or their contribution as the highly variable library sample of a limited amount prevents more sophisticated physico-chemical analyses that could be used to address these phenomena in follow-up studies.

Early alphabet proteins are inherently more thermostable in a cell-like milieu

One of the notable assumed characteristics of the early prebiotic Earth is an elevated temperature of its environment [47]. The temperature-induced aggregation propensity of random protein libraries was investigated by their exposure to a 15-min heat shock at 42 °C. Interestingly, the quantity of thermostable fractions in chaperone unexposed proteins was approximately two times higher in the early alphabet library (~30% vs ~60% for 20F and 10E libraries, respectively) which might indicate its natural tendency to withstand an elevated temperature environment. On the other hand, addition of chaperones improved the thermal stabilities of both 20F and 10E libraries up to almost full solubility upon the heat shock treatment. This observation confirms our previous conclusions about strong dependence of the canonical amino acid alphabet proteins on the chaperone activity and extends it to stability support of the early alphabet proteins. Additionally, the fraction of protease resistant proteins remains unchanged (~40%) upon the heat shock for both libraries, suggesting that the proteins destabilized by the elevated temperature do not belong to this category.

While most of the above referenced studies reducing the composition of extant proteins towards the early set of amino acids did not observe an increase in their thermostability [15,16,18,42,44], we are here concerned with a comparison of unevolved sequences from the two amino acid repertoires and therefore their inherent properties. Unlike the studies performed with purified protein samples, our thermostability assay was performed in the CFPS reaction milieu, i.e. in an environment rich in salts and other small molecules, indicating innate thermostability properties in an access of such chemical entities.

Concluding remarks

In summary, while our study confirms some of the previously reported properties of the random sequences space (such as its surprisingly high secondary structure potential and relative ease of expression), we expand on this knowledge using a systematic high-throughput approach using diverse combinatorial libraries composed of two different alphabets. Escaping the restraints of scarce sampling, our study maps the tertiary structure, solubility and thermostability potential in random sequences composed of the natural vs. the evolutionary early canonical alphabets. The analyses were performed in a cell-like environment (rich in salts and cofactors) that may better represent protein birth conditions during both origins of life and in extant biology. Under such conditions, the early alphabet sequences are inherently more soluble and thermostable while the natural alphabet proteins can reach similar properties by the activity of natural chaperones. Interestingly, our study reports that both alphabets frequently give rise to proteolysis resistant soluble structures, occupying upto ~40% of all sequences. Because the intrinsic properties of the prebiotically plausible amino acids do not imply such properties, we hypothesize that the protein solubility and folding within this library are enabled by the cell-like milieu, assisted by salts, metal cation and cofactors. However, follow up studies are suggested to further explore these findings.

Methods

Design of libraries from early and full amino acid alphabet

Two 105 amino acid long random sequence libraries were designed using the CoLiDe algorithm for combinatorial library design [26], using the amino acid ratios listed in Supplementary Table S1. The randomized part of the libraries consisted of 84 amino acids, the rest is attributed to the FLAG affinity purification site on the N-end of the construct, hexahistidine tag on the C-end and thrombin protease recognition site (ALV**PRG**S) in the middle of the construct (Supplementary Figure S1).

Bioinformatic analysis of secondary structure potential

Prediction of secondary structure potential of the studied libraries was performed by a consensus predictor as described previously [28]. It combines outputs of the spider3, psipred, predator, jnet, simpa and GOR IV secondary structure predictors [48–53]. None of the predictors was allowed to use homology information that might prevent high-throughput processing of protein sequences. The final assignment of secondary structure followed the most frequently predicted secondary structure element at each amino acid position. Protein aggregation was predicted by the ProA algorithm in a protein prediction mode [54].

Preparation of experimental libraries

20F and 10E DNA libraries were synthesized commercially as two overlapping degenerate oligonucleotides (see Supplementary information for the sequences) that were designed by the CoLiDe algorithm to follow the natural canonical (full alphabet, 20F) and prebiotically plausible (A,S,D,G,L,I,P,T,E,V; early alphabet, 10E) amino acid distributions (Supplementary Table S1). The overlapping oligonucleotides were annealed and extended by Klenow fragment to form a double-stranded DNA (dsDNA). Annealing was performed by heating complementary oligonucleotide mixture (48 µl total reaction volume, 2 µM final concentration of each) in NEB2 buffer provided with 200 µM dNTPs to 90 ºC for 2 minutes and cooling down to 32 ºC with 1 ºC/min temperature gradient. The Klenow extension was performed by Klenow polymerase (NEB): 10 U of Klenow polymerase was added to annealed oligonucleotides, incubated for 5 minutes at 25 °C, 37°C for 1 hour (polymerization step) and at 50 °C for 15 minutes (inactivation step). Final dsDNA libraries were further column purified using the DNA Clean and Concentrator kit (Zymo Research) and the product was quantified by Nanodrop 2000c (Thermo Scientific). In the following transcription, 1 µg of DNA library was used as a template for mRNA synthesis by HiScribe T7 kit (NEB). The product was purified by NH4Ac precipitation and dissolved in RNAse-free water to a final concentration of 3 µg/ul.

The library DNA was analyzed by high throughput sequencing on Illumina MiSeq. The libraries for next generation sequencing (NGS) were prepared from 100 ng of DNA samples using the NEBNext Ultra II DNA Library Prep kit (New England Biolabs) with AMPure XP purification beads (Beckman Coulter). Length of the prepared library was determined by Agilent 2100 Bioanalyzer (Agilent Technologies) and quantified by Quantus Fluorometer (Promega). Sample was sequenced on MiSeq Illumina platform using the Miseq Reagent Kit v2 500-cycles (2x250) in a paired-end mode. Raw data was processed with the Galaxy platform and sequence analysis of assembled and filtered paired reads was performed with MatLab scripts developed at Heinis lab [55,56].

Protein library was expressed using PUREfrex 2.0 (GeneFrontier Corporation) recombinant in vitro translation system. Reaction was supplemented by 0.05 % (v/v) Triton X-100 and prepared according to manufacturer recommendations. Reaction was initiated by 3 µg of library mRNA. Expression followed for 2 hours at 25, 30 or 37 °C.

Affinity purification of protein libraries

Expressed protein libraries were diluted 10x in binding buffer (50mM Tris, 150 mM NaCl, 0.05% (v/v) Triton X-100, pH 7.5) and incubated for 2 hours at 25 °C with 3 µl / 20 µl reaction of TALON affinity purificaton matrix. Immobilized library was washed three times with binding buffer and eluted by addition of 20 µl / 20 µl reaction of elution buffer (50mM Tris, 150 mM NaCl, 10mM EDTA, 0.05% (v/v) Triton X-100, pH 7.5).

Solubility analysis of protein libraries

Cell free protein expression reactions were supplemented with 0.05 % Triton X-100 and protein libraries were expressed in different temperatures according to manufacturer recommendations. In order to analyze the quantity of total protein product, 10 µl of reactions were quenched by 40 µl addition of 300 µM puromycin in 50 mM Tris, 100 mM NaCl, 100 mM KCl, pH 7.5. Quenching proceeded for 30 minutes at 30 °C. Next, 5 µl of the quenched reaction mixture was taken for the following SDS-PAGE analysis of total library expression, the rest of the mixture was centrifuged for 30 minutes at 21 °C and 5 µl of supernatant was taken for SDS-PAGE analysis of soluble fraction of the library. Both fractions were analyzed by quantitative Western blotting (Sigma-Aldrich Monoclonal ANTI-FLAG® M2-Peroxidase (HRP) antibody, A8592) following the SDS-PAGE separation.

Lon proteolytic assay of protein libraries

Lon protease was expressed and purified according to the previously published protocol protocol (niwa2019). Cell free expression reactions were supplemented with 0.05 % Triton X-100 and reactions were prepared according to manufacturer recommendations. Libraries were expressed in presence or absence of chaperone DnaK (K+/K-) and in presence or absence of Lon protease (L+/L-). Chaperones were added to the final concentration of 5 µM DnaK, 1 µM DnaJ, 1 µM GrpE and Lon protease to 0.4 µM (hexamer)/reaction. The expression proceeded in 10 µ reaction volume for 2 hours at 30 °C and was quenched by 40 µl addition of 300 µM puromycin in 50 mM Tris, 100 mM NaCl, 100 mM KCl, pH 7.5. Quenching proceeded for 30 minutes at 30 °C. The sample preparation of total and soluble library fractions was identical to the solubility analysis experiment described above.

Thrombin proteolytic assay of protein libraries

Cell free expression reactions were supplemented with 0.05 % Triton X-100 and reactions were prepared according to manufacturer recommendations. Libraries were expressed in the presence or absence of chaperone DnaK (K+/K-). Chaperones were added to the final concentration of 5 µM DnaK, 1 µM DnaJ, 1 µM GrpE µM. Expression proceeded in 10 µl reaction volume for 2 hours at 30 °C and was quenched by 40 µl addition of 300 µM puromycin in 50 mM Tris, 100 mM NaCl, 100 mM KCl, pH 7.5. Quenching proceeded for 30 minutes at 30 °C. Post-translational thrombin proteolysis was prepared as follows - 5 µl of quenched reaction was diluted 4x by 15 µl of 50 mM Tris, 100 mM NaCl, 100 mM KCl, pH 7.5, 0.15 U of thrombin (SigmaAldrich, USA) was added, and total expressed library was digested for 2 hours at 30 °C. The soluble fraction of the library was prepared by centrifugation at 21 000 xg for 30 minutes at 21 °C and 5 µl of supernatant was thrombin digested according to the same protocol. Cleaved samples of the total expressed, and soluble library were analyzed by SDS-PAGE and Western blotting (Sigma-Aldrich Monoclonal ANTI-FLAG® M2-Peroxidase (HRP) antibody, A8592).

Thermostability assay

Libraries expressed in 10 µl volume were processed as described above in the Lon proteolytic assay protocol. The Lon absent libraries were further analyzed for their thermostability in chaperone present and absent conditions. Processed reactions were incubated at 42 °C for 15 minutes and immediately centrifuged at 21 000 xg for 30 minutes at 21 °C. The 5 µl supernatant fractions were subjected to thrombin proteolysis as described previously and analyzed by SDS-PAGE and quantitative western blotting.

Quality control of purified protein libraries

For mass spectrometry, the purified protein library sample was resuspended in water. The spectrum was collected after addition of 2,5-dihydroxybezoic acid matrix substance (Merck) using an UltrafleXtremeTM MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Germany) in linear mode.

Acknowledgements

We are grateful to Prof. Hideki Taguchi and Prof. Tatsuya Niwa for kindly providing us with the expression plasmid of the LON protease used in this study. We would also like to acknowledge Dan S. Tawfik† and Valerio Guido Giacobelli for helpful discussions about this manuscript. In addition, we would like to thank Kateřina Nováková for her technical help with collecting the MALDI spectra.This work was supported by the Czech Science Foundation (GAČR) grant number 17-10438Y and the Human Frontier Science Program grant HFSP-RGY0074/2019. K.F. is supported by ELSI - First Logic Astrobiology Donation Program.

References

1. Cleaves, H. J. The origin of the biologically coded amino acids. *J. Theor. Biol.* **263**, 490–498 (2010).

2. Zaia, D. A. M., Zaia, C. T. B. V. & De Santana, H. Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph.* **38**, 469–488 (2008).

3. Higgs, P. G. & Pudritz, R. E. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **9**, 483–490 (2009).

4. Trifonov, E. N. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139–151 (2000).

5. Weber, A. L. & Miller, S. L. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* **17**, 273–284 (1981).

6. Freeland, S. 'Terrestrial' Amino Acids and their Evolution. *Amin. Acids, Pept. Proteins Org. Chem.* **1**, 43–75 (2010).

7. Philip, G. K. & Freeland, S. J. Did evolution select a nonrandom 'alphabet' of amino acids? *Astrobiology* **11**, 235–240 (2011).

8. Ilardo, M. *et al.* Adaptive Properties of the Genetically Encoded Amino Acid Alphabet Are Inherited from Its Subsets. *Sci. Rep.* **9**, (2019).

9. Pace, N. R. The universal nature of biochemistry. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 805–808 (2001).

10. Holliday, G. L., Fischer, J. D., Mitchell, J. B. O. & Thornton, J. M. Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *FEBS J.* **278**, 3835–3845 (2011).

11. Di Mauro, E., Dunker, A. K. & Trifonov, E. N. Disorder to Order, Nonlife to Life: In the Beginning There Was a Mistake. 415–435 (2012) doi:10.1007/978-94-007-2941-4_23.

12. Newton, M. S., Morrone, D. J., Lee, K. H. & Seelig, B. Genetic Code Evolution Investigated through the Synthesis and Characterisation of Proteins from Reduced-Alphabet Libraries. *ChemBioChem* **20**, 846–856 (2019).

13. Tanaka, J., Doi, N., Takashima, H. & Yanagawa, H. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci.* **19**, 786–795 (2010).

14. Riddle, D. S. *et al.* Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805–809 (1997).

15. Longo, L. M., Lee, J. & Blaber, M. Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2135–2139 (2013).

16. Shibue, R. *et al.* Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* **8**, (2018).

17. Solis, A. D. Reduced alphabet of prebiotic amino acids optimally encodes the conformational space of diverse extant protein folds. *BMC Evol. Biol.* **19**, 1–19 (2019).

18. Giacobelli, V. *et al.* In vitro evolution reveals primordial RNA-protein interaction mediated by metal cations. *bioRxiv* (2021) doi:https://doi.org/10.1101/2021.08.01.454623.

19. Longo, L. M. *et al.* Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15731–15739 (2020).

20. Houben, B. *et al.* Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues. *EMBO J.* **39**, 1–22 (2020).

21. Chiarabelli, C., Vrijbloed, J. W., Thomas, R. M. & Luisi, P. L. Investigation of de novo Totally Random Biosequences. *Chem. Biodivers.* **3**, 827–839 (2006).

22. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).

23. Labean, T. H., Butt, T. R., Kauffman, S. A. & Schultes, E. A. Protein folding absent selection. *Genes (Basel).* **2**, 608–26 (2011).

24. Yu, J. F. *et al.* Natural protein sequences are more intrinsically disordered than random sequences. *Cell. Mol. Life Sci.* **73**, 2949–2957 (2016).

25. Tretyachenko, V. *et al.* Random protein sequences can form defined secondary

structures and are well-tolerated in vivo. *Sci. Rep.* **7**, (2017).

26. Tretyachenko, V., Voracek, V., Soucek, R., Fujishima, K. & Hlouchova, K. CoLide: Combinatorial library design tool for probing protein sequence space. *Bioinformatics* **37**, 482–489 (2021).

27. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

28. Vymětal, J., Vondrášek, J. & Hlouchová, K. Sequence versus composition: What prescribes IDP biophysical properties? *Entropy* **21**, 1–8 (2019).

29. Johannes, S. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

30. Melderen, L. Van & Aertsen, A. Regulation and quality control by Lon-dependent proteolysis. *Res. Microbiol.* **160**, 645–651 (2009).

31. Niwa, T., Uemura, E., Matsuno, Y. & Taguchi, H. Translation-coupled protein folding assay using a protease to monitor the folding status. *Protein Sci.* **28**, 1252–1261 (2019).

32. Bornberg-Bauer, E., Hlouchova, K. & Lange, A. Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* **68**, 175–183 (2021).

33. White, S. H. The evolution of proteins from random amino acid sequences: II. Evidence from the statistical distributions of the lengths of modern protein sequences. *J. Mol. Evol.* **38**, 383–394 (1994).

34. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 1–7 (2017).

35. Bornberg-Bauer, E., Hlouchova, K. & Lange, A. Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* **68**, 175–183 (2021).

36. Davidson, A. R. & Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 2146–2150 (1994).

37. Benner, S. A. Enzyme Kinetics and Molecular Evolution. *Chem. Rev.* **89**, 789–806 (1989).

38. Raggi, L., Bada, J. L. & Lazcano, A. On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. *Phys. Chem. Chem. Phys.* **18**, 20028–20032 (2016).

39. Brack, a & Orgel, L. E. Beta structures of alternating polypeptides and their possible prebiotic significance. *Nature* **256**, 383–387 (1975).

40. Lupas, A. N. & Alva, V. Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J. Struct. Biol.* **198**, 74–81 (2017).

41. Kovacs, N. A., Petrov, A. S., Lanier, K. A. & Williams, L. D. Frozen in Time: The History of Proteins. *Mol. Biol. Evol.* **34**, 1252–1260 (2017).

42. Makarov, M. *et al.* Enzyme catalysis prior to aromatic residues: Reverse engineering of a dephosphoCoA kinase. *bioRxiv* (2020).

43. Kimura, M. & Akanuma, S. Reconstruction and Characterization of Thermally Stable and Catalytically Active Proteins Comprising an Alphabet of ~ 13 Amino Acids. *J. Mol. Evol.* **88**, 372–381 (2020).

44. Longo, L. M., Tenorio, C. A., Kumru, O. S., Middaugh, C. R. & Blaber, M. A single aromatic core mutation converts a designed 'primitive' protein from halophile to mesophile folding. *Protein Sci.* **24**, 27–37 (2015).

45. Despotović, D. *et al.* Polyamines mediate folding of primordial hyperacidic helical proteins. *Biochemistry* **59**, 4456–4462 (2020).

46. Yadid, I., Kirshenbaum, N., Sharon, M., Dym, O. & Tawfik, D. S. Metamorphic proteins mediate evolutionary transitions of structure. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7287–7292 (2010).

47. Islas, S., Velasco, A. M., Becerra, A., Delaye, L. & Lazcano, A. Hyperthermophily and the origin and earliest evolution of life. *Int. Microbiol.* **6**, 87–94 (2003).

48. Heffernan, R. *et al.* Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem.* **39**, 2210–2216 (2018).

49. Jones, T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, (1999).

50. Frishman, D. & Argos, P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins Struct. Funct. Genet.* **27**, 329–335 (1997).

51. Cuff, J. a & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**, 502–511 (2000).

52. Levin, J. M. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.* **10**, 771–776 (1997).

53. Garnier, J., Gibrat, J. F. & Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540–553 (1996).

54. Fang, Y., Gao, S., Tai, D., Middaugh, C. R. & Fang, J. Identification of properties important to protein aggregation using feature selection. *BMC Bioinformatics* **14**, 314 (2013).

55. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).

56. Rebollo, I. R., Sabisz, M., Baeriswyl, V. & Heinis, C. Identification of target-binding peptide motifs by high-throughput sequencing of phage-selected peptides. *Nucleic Acids Res.* **42**, e169–e169 (2014).