OPEN BIOLOGY

royalsocietypublishing.org/journal/rsob

Research

Cite this article: Tretyachenko V, Vymětal J, Neuwirthová T, Vondrášek J, Fujishima K, Hlouchová K. 2022 Modern and prebiotic amino acids support distinct structural profiles in proteins. *Open Biol.* **12**: 220040. https://doi.org/10.1098/rsob.220040

Received: 8 February 2022
Accepted: 26 May 2022

Subject Area:
biochemistry/biophysics/synthetic biology

Keywords:
protein sequence space, protein structure, amino acid alphabet, genetic code evolution, random proteins

Author for correspondence:
Klára Hlouchová
e-mail: klara.hlouchova@natur.cuni.cz

†Present address: R&D Informatics Solutions, MSD Czech Republic s.r.o., Prague, Czech Republic.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6032490.

THE ROYAL SOCIETY PUBLISHING

Modern and prebiotic amino acids support distinct structural profiles in proteins

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, and [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 2520882 Japan

The earliest proteins had to rely on amino acids available on early Earth before the biosynthetic pathways for more complex amino acids evolved. In extant proteins, a significant fraction of the 'late' amino acids (such as Arg, Lys, His, Cys, Trp and Tyr) belong to essential catalytic and structure-stabilizing residues. How (or if) early proteins could sustain an early biosphere has been a major puzzle. Here, we analysed two combinatorial protein libraries representing proxies of the available sequence space at two different evolutionary stages. The first is composed of the entire alphabet of 20 amino acids while the second one consists of only 10 residues (ASDGLIPTEV) representing a consensus view of plausibly available amino acids through prebiotic chemistry. We show that compact conformations resistant to proteolysis are surprisingly similarly abundant in both libraries. In addition, the early alphabet proteins are inherently more soluble and refoldable, independent of the general Hsp70 chaperone activity. By contrast, chaperones significantly increase the otherwise poor solubility of the modern alphabet proteins suggesting their coevolution with the amino acid repertoire. Our work indicates that while both early and modern amino acids are predisposed to supporting protein structure, they do so with different biophysical properties and via different mechanisms.

1. 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 later evolution to the full, standard alphabet [3,4].

The factors that drove the selection of 20 coded amino acids remain puzzling. Solubility, ease of biosynthesis, (un)reactivity with tRNA and potential peptide product stability seem to explain some selective '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

speculation that similar amino acid selection would be expected on other Earth-like planets [5,8,9].

In extant proteins, a significant fraction 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 three-dimensional structure [11]. However, sparse 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 an 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]. Most of these analyses relied on random sequences as proxies of unevolved proteins. Besides reporting on amino acids alphabet intrinsic properties, such sequence libraries help us understand the nature of the 'dark protein space' or 'never born proteins' (i.e. the sequence space that is not used by nature [22]). 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 [22,24,25]. However, more systematic and high-throughput screening is still necessary to confirm these observations, which are key for both understanding the phenomenon of protein evolution as well as protein design initiatives. 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 whether 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. Our approach takes advantage of combinatorial samples to address the statistically inaccessible characterization of random sequence space by low-throughput single-protein studies. As such, we not only report on the full and early amino acid alphabet structural propensities but also perform a search of the vast sequence space that can be created using these alphabets. 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.

2. Results

2.1. Library expression and quality control

The combinatorial protein libraries studied in this work consisted of 105-amino-acid-long proteins with 84-amino-acid-long variable parts, FLAG/HIS tag sequences on N′/C′ ends, and a thrombin cleavage site in the middle of the protein construct (electronic supplementary material, figure 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 the prebiotically plausible (A,S,D,G,L,I,P,T,E,V; early alphabet, 10E) amino acid distributions (electronic supplementary material, table S1) [26]. The CoLiDe algorithm was chosen on the basis of its suitability for construction of vast combinatorial libraries. In comparison to alternative degenerate codon design tools it was specifically optimized for long variable protein libraries design rather than libraries suitable for site-specific mutagenesis investigations of protein variants. The design method consists in a selection of such degenerate codons which upon their combination in a degenerate DNA template produce a protein-coding library with the desired mean amino acid distribution. Although characteristics of different degenerate codons may produce a sequence-biased sample (each degenerate position will yield only a subset of the designed amino acid alphabet), this study aims to investigate effects of amino acid composition effects on random protein behaviour rather than sequence determinants of protein folding. 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 (electronic supplementary material, figure 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 approximately 0.06 in both libraries 20F and 10E (electronic supplementary material, table S2, figure S3). The variability analysis of the sequenced library showed that 96% of sequences were unique; no significant sequence enrichment was observed (figure 1, electronic supplementary material, table S3). Due to synthesis errors, STOP codons were introduced into 12% of the library sequences. The rates of misincorporation of undesired amino acids into library 10E did not exceed 1% and maximum deviation on single amino acid occurrence was 30% from the designed frequency (electronic supplementary material, table S2). 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 (electronic supplementary material, figure S4).

2.2. Secondary structure, aggregation and solubility predictions

Sequences of both 20F and 10E libraries acquired by high-throughput sequencing were analysed by a consensus protein secondary structure prediction [29]. 200 000 sequences were

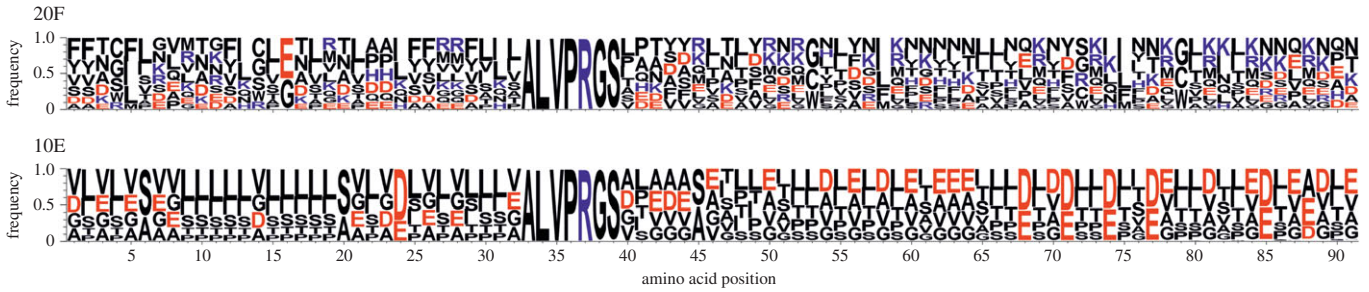

Figure 1. Sequence logo representation of full (top) and early (bottom) alphabet libraries variability constructed from the corresponding sequenced DNA templates. Sizes of the letters represent frequencies of specific amino acids per each position in the set of translated sequenced templates. Proteins coded by degenerate DNA templates can be represented by a linear combination of all residues with each amino acid occurring with its distinct frequency. Sequence logo created by WebLogo 3 [28].

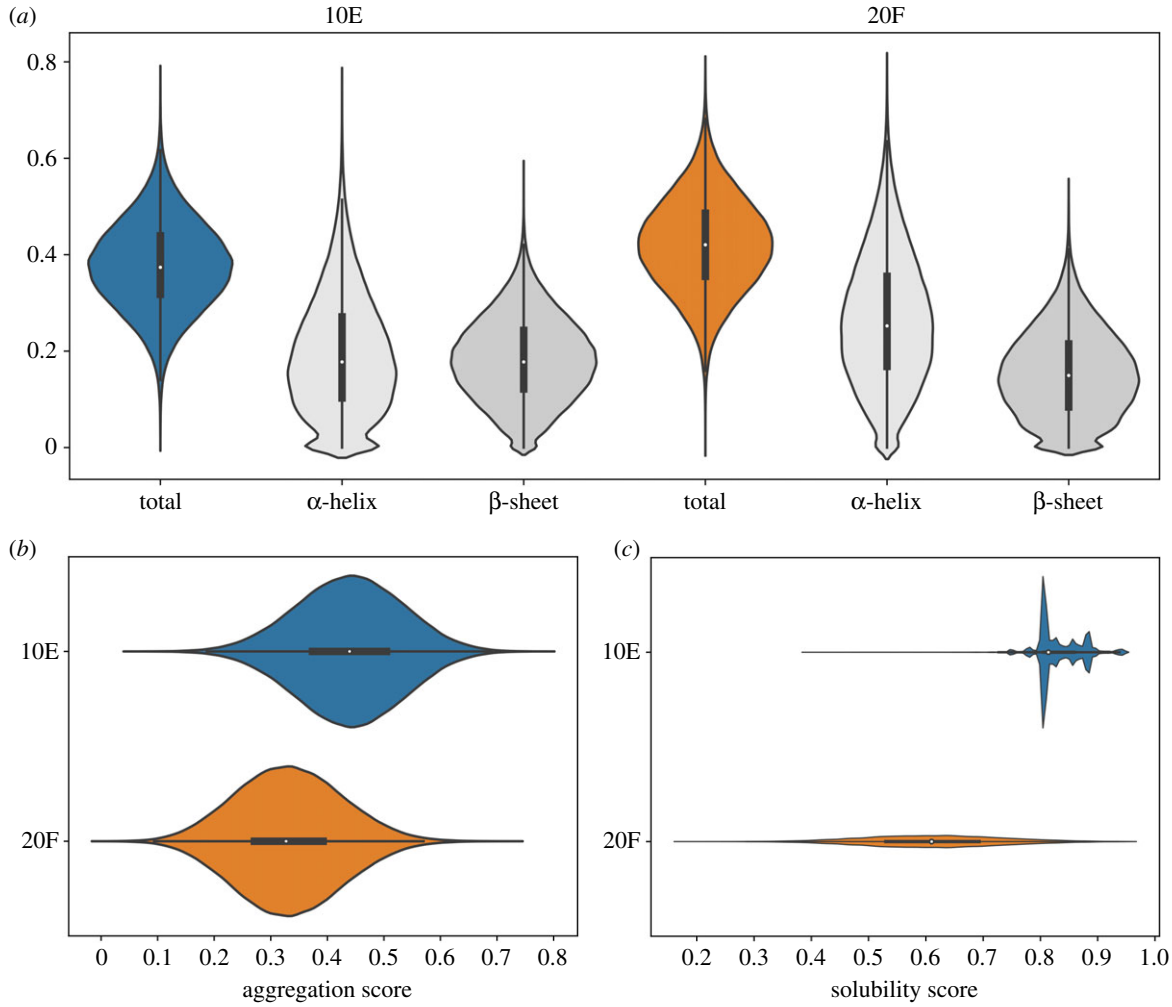

Figure 2. Bioinformatic prediction of (*a*) α-helix and β-sheet content, (*b*) aggregation propensity and (*c*) solubility analysis of a sample of 200 000 sequences acquired by high-throughput sequencing of the early (blue) and full (orange) alphabet library DNA templates. Secondary structure was assigned using consensus prediction of spider3, psipred, predator, jnet, simpa and GOR IV secondary structure predictors [30–35]. Aggregation prediction was performed by the ProA algorithm in a protein prediction mode [36]. Aggregation score is defined as the ratio of predicted aggregation-prone residues per sequence. Solubility was predicted with Protein-Sol predictor and scaled solubilities were plotted [37]. The box extends from first quartile to third quartile with a point in the middle representing the median. The whiskers extend from the box by 1.5× the inter-quartile range. Kernel density estimation surrounding the boxplots represent the distribution of data.

analysed from each library. Interestingly, despite the different amino acid distributions, comparable α-helix and β-sheet forming tendencies were reported in both libraries with only a slight increase in α-helix content in the 20F library (33% versus 30% in 10E) (figure 2*a*). The overall α-helix and β-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 a significantly higher aggregation tendency of 10E library proteins in comparison to 20F library proteins (figure 2*b*). On the other hand, higher predicted solubility of 10E proteins reflects its lower average pI values (average pI of 4.06) in

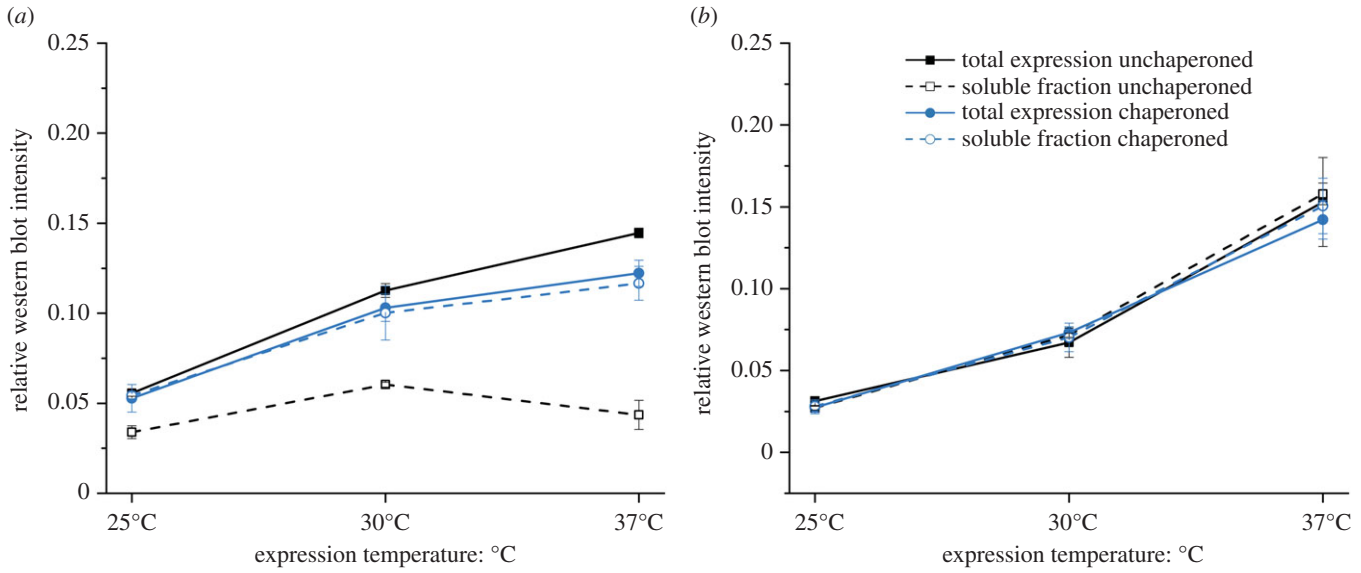

Figure 3. A summary of expression and solubility analysis of the (*a*) full 20F and (*b*) early 10E alphabet libraries at three different temperatures. Total expression (solid line) and soluble fraction (dashed line) were compared in chaperoned (blue line) and unchaperoned (black line) conditions. For original data see electronic supplementary material, figures S5 and S6, and table S4.

comparison to the broad solubility distribution of 20F proteins (figure 2*c*). Nevertheless, proteins of both libraries fall into the soluble category (solubility score greater than 0.45) as defined by Hebditch *et al*. [37] based on experimental validation of 3173 *E. coli* proteins expressed in a cell-free translation system [37,38].

2.3. Expression and solubility analysis in the absence and presence of the 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°C, 30°C and 37°C) and with/without DnaK/DnaJ/GrpE chaperone system supplementation (further referred as to DnaK). The analysis was carried out in triplicate, and western blot signals of both total expression and soluble fractions were quantified with ImageJ [39]. For both 20F and 10E libraries, the expression yields improved with increasing temperature, with the overall yield being mildly lower in the chaperone supplemented reactions at 37°C (figure 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 20F chaperone supplemented reaction the soluble fraction improved with expression temperature proportionally with the total expression, in the 20F chaperone absent condition, the soluble fraction yields did not significantly change with the transition from 30°C to 37°C (figure 3*a*). On the other hand, chaperone supplementation did not have a significant effect on the 10E library expression or solubility (figure 3*b*).

2.4. Assessment of proteolytic resistance

The structural potential of random protein libraries was assessed by proteolysis. The digestion assessment was performed in triplicate by Lon and thrombin proteases in co-translational and post-translational conditions, respectively (figure 4). The Lon protease is a part of the *E. coli* protein misfolding system and is known to specifically digest unfolded proteins in exposed hydrophobic regions [40]. Here we adapted a previously published protocol on single protein structure assessment for combinatorial library characterization [41]. 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*., wherein 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 [22]. The unstructured proteins are expected to be quickly degraded on the exposed cleavage site. While co-translational Lon protease assay represents real-time analysis of protein folding kinetics, thrombin protease digestion aims for an indirect final folding assessment via proteolysis on accessible or buried cleavage sites. Both assays target different stages of protein folding pathways and bring distinct insights into the overall random protein folding behaviour.

According to the 20F library analysis, the soluble/undegradable structured proteins represent approximately 30–35% of the total product (figure 5*a*). Upon addition of the DnaK chaperone system, most of the library solubilizes, but the protease resistant content does not increase significantly and occupies approximately 40–50% of the total product. In comparison, chaperone addition does not have an impact on the solubility or protease resistance of the 10E library (figure 5*b*). Interestingly, the protease resistant content (soluble undegradable) in the 10E library is similar to in the 20F library after the addition of chaperones.

2.5. Protein heat shock refoldability characterization

Following expression, solubility, and protease resistance assessment, we analysed the temperature sensitivity of the 20F and 10E proteins. The libraries expressed with and without chaperone supplementation were subjected to 15 min/42°C heat shock. The aggregated fraction was removed by

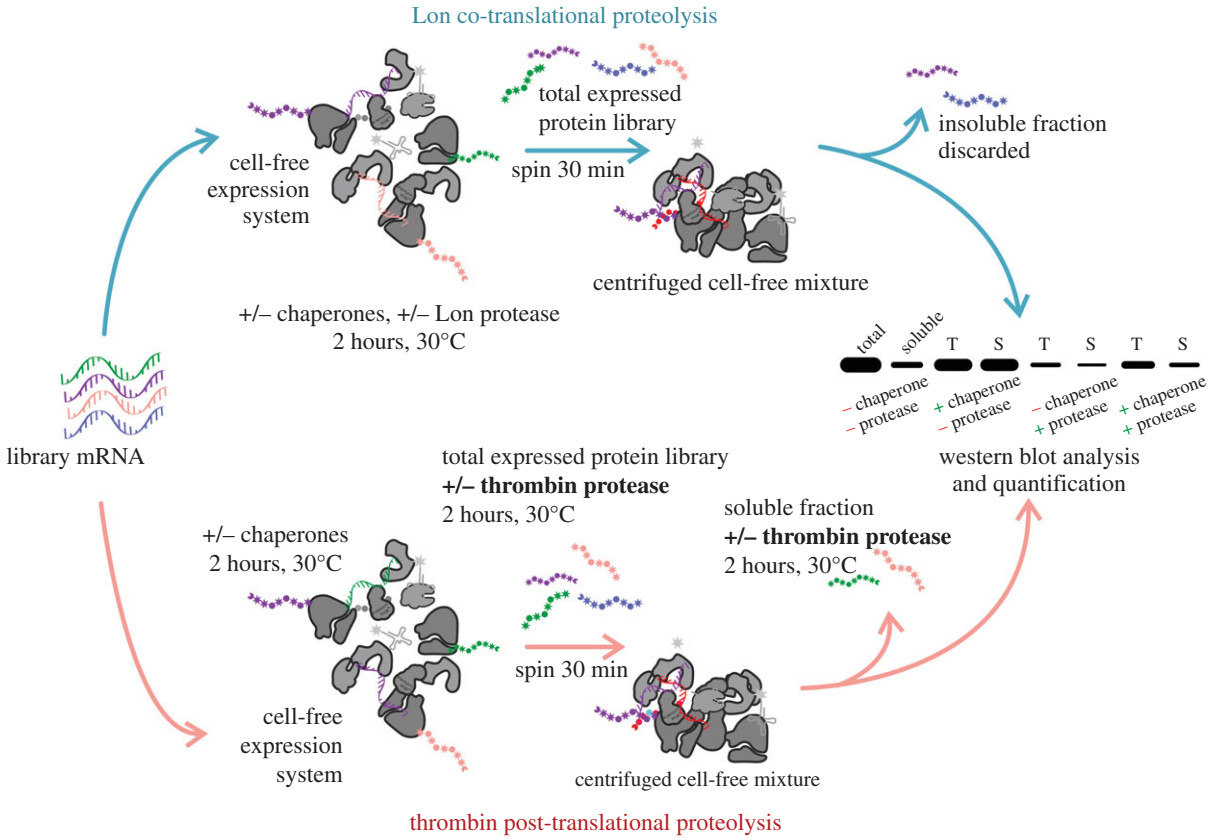

Figure 4. Scheme of the proteolytic resistance experimental pipeline. In the co-translational proteolytic assay (top) the Lon protease is present during the cell-free expression; in the post-translational proteolytic assay (bottom) thrombin protease is added to the separated total and soluble fractions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

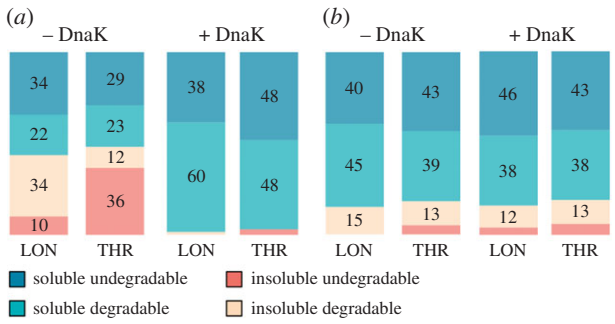

Figure 5. An integrated solubility/proteolysis resistance analysis of the (*a*) full 20F and (*b*) early 10E alphabet libraries. Libraries were expressed either in the absence (left double column) or presence (right double column) of the DnaK chaperone system. Proteolysis was performed by protease Lon (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 electronic supplementary material, figures S7–S10, and tables S5 and S6.

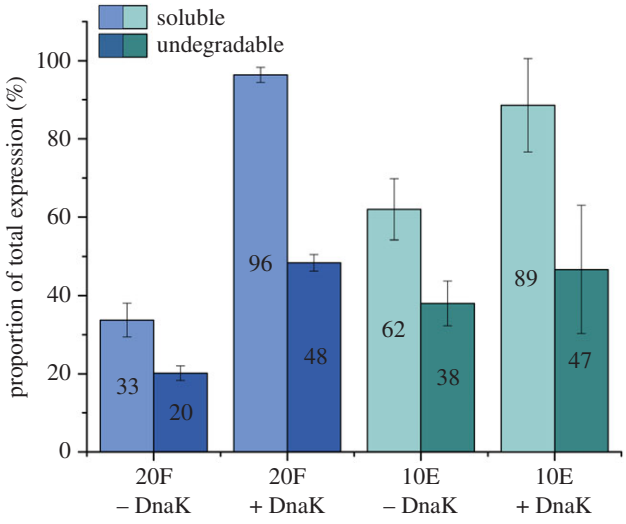

Figure 6. Refoldability analysis showing soluble proportions (light blue and green) of the total expression of the full and early alphabet libraries after a heat shock (42°C/15 min) treatment and their respective thrombin resistant proportions (dark blue and green) of the total expression in unchaperoned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

centrifugation, and the soluble fraction was compared with and without thrombin treatment (figure 6).

The 10E library is intrinsically more soluble than 20F (approx. 60% versus 30% of the libraries remain soluble after heat shock, respectively) while the DnaK chaperone system induces higher post-heat shock solubility in both libraries. The protease-resistant fraction of the soluble part of the libraries remains the same (approx. 40%) as before heat shock treatment with the exception of the unchaperoned 20F library, which demonstrates a decrease in both the soluble and degradation resistant fractions (figure 6).

3. Discussion

In this study, a high-throughput systematic approach was used to experimentally analyse 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 [42,43]. 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 [24,25,44]. Here, two combinatorial protein libraries encompassing up to 10^{12} individual sequences from two distinct alphabets (representing hypothetical stages of genetic code evolution) have been characterized.

3.1. 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 representing the properties of today's proteomes. It has previously been 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. No significant aggregation was observed upon the library expression at 25°C. While previous studies of similar construct size evaluated the solubility of individual proteins that were overexpressed (many of them with partial solubility) in 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 a 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 used to supplement the CFPS system 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 structural potential without introducing potential bias, we used an *in situ* double proteolysis experiment adapting two previously reported approaches [22,41]. The experiment combined co-translational proteolysis by disorder-specific Lon protease and a post-translational cleavage by thrombin designed to cut the potentially exposed cleavage site engineered in the center of random proteins. Besides the increased robustness of the structure content estimation, such a combined approach provides unique insight into the library translation dynamics.

The double proteolysis experiment revealed that approximately 30–35% of library 20F proteins are protease resistant. Upon the addition of chaperones (which solubilizes the library as described above), the ratio of protease resistant species rose only mildly to approximately 40–50%. The more prevalent protease sensitive nature of the full alphabet library echoes the reported nature of *de novo* proteins, i.e. proteins that emerge in current biology from previously non-coding DNA (summarized in [43]).

Overall, these results show that while inherent protein solubility is limited in random sequence space made of the full amino acid alphabet, it can be induced significantly by the activity of molecular chaperones. At the same time, the DnaK chaperone system has only a minor effect on the level of protease resistance, suggesting that the majority of the potentially solubilized sequences are devoid of higher structure arrangements. In comparison, the same folding assessment of 76 randomly selected *E. coli* proteins by Niwa *et al*. were Lon-resistant in their soluble fraction, suggesting a high level of folding optimization of biological versus random proteins [41]. These results are in agreement with earlier studies of the Hecht group which pointed out that even though structured arrangements are achievable within random sequence space, most of its representants appear to be in relaxed molten globule states [45,46]. Nevertheless, the approximately 40% natural abundance of soluble and yet protease-resistant sequences in unevolved sequence space may be surprising in light of earlier hypotheses and even exceeds the estimates of folding frequency reported by previous coarse-grained studies [22,47]. However, major differences in the experimental set-ups (cell-free versus cell-based expression, low-level versus overexpression, high- versus low-throughput methodology, overall amino acid composition 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.

3.2. Protease resistance is comparable in proteins from the full canonical alphabet and its early subset, unaffected by chaperones

A 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 [4]. We emphasize that with this study, we do not try to establish that there was necessarily a time in life's evolution during which 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 used a 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 [1,48,49].

Although the overall secondary structure propensity of the early alphabet is comparable to the full alphabet, according to the bioinformatic prediction, the occurrence of α-helix is slightly (approx. 3%) lower. While these differences are statistically borderline, they may have interesting implications for the evolution of protein structural properties. Brack and Orgel proposed that beta-sheet structures were prebiotically significant, and the later significance of α-helices in protein folds was also recently implied by the structural analysis of ribosomal protein content, showing that the most ancient protein-protein fragments of this molecular fossil are mostly disordered and of

β-sheet formation [50–52]. Despite the similar secondary structure propensities of the full and early alphabets, the 10E library proteins are significantly more soluble (approx. 90%) upon expression. They retain similar solubilities in chaperoned/unchaperoned conditions unlike the 20F library proteins. This observation supports the previously stated hypothesis of chaperone coevolution with the incorporation of the first positively charged amino acids into the early amino acid alphabets [20]. That way proteins composed of the full alphabet would be kept in solution despite their lower inherent solubility.

The significantly higher solubility of the 10E library proteins (and similar protein compositions) is in agreement with previous studies [12,13]. This phenomenon could be related to their highly acidic nature as well as the lower complexity of 10E library proteins resulting from the limited amino acid alphabet. While 20F proteins represent a highly variable sample of protein folding space with many opportunities for aggregation initiation, the 10E proteins display a narrower subspace with much more uniform sequence and physico-chemical characteristic distributions. That would make the 10E search landscape significantly less complex. In addition, the overall negative charge and absence of positively charged/aromatic amino acids of the 10E alphabet are factors which were previously shown to suppress both non-specific aggregations as well as independent protein folding formation [20,53]. At the same time though, the 10E alphabet contains a significant proportion of hydrophobic amino acids. Using the combination of ProA and Protein-Sol bioinformatic predictors of protein aggregation and solubility, the 10E library would be expected to be highly soluble despite its tendency to form higher oligomeric conformation.

Accordingly, the 10E library indeed displays a significant protease-resistant behaviour. 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 level of protease resistance within the 10E library can be speculated to be caused by structure formation via oligomerization. However, several independent folders have been recently identified from the same or similar protein composition in experiments reducing extant protein compositions [15,16,18,54,55]. Where characterized in more detail, assistance of salts, metal ions or cofactor binding were found to explain the folding properties [15,18,55,56]. In addition, Despotović *et al*. recently confirmed that folded conformations of a highly acidic 60-residue protein can be induced by positively charged counterions, in case of Mg^{2+} the reported concentration corresponding roughly to its concentration in the CFPS reaction (approx. 10 mM) [57]. In parallel, the Hecht group reported that binding of metal ions (with high nanomolar to low micromolar affinity) is a surprisingly frequent property of unevolved sequences and therefore does not require much sequence optimization [58]. These studies allow us to speculate that the high protease resistance of the 10E alphabet could result from the folding assistance in cation/cofactor-rich environment (besides metal ions and organic cofactors the CFPS reaction contains species such as the polyamine spermidine), 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 compositions) [59]. Early protein evolution by oligomerization of shorter parts could have provide a quick solution to protein structure without complex evolutionary optimization of independent protein folding. Our study presented here cannot unambiguously differentiate between these two possible scenarios or their combination 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.

3.3. Early alphabet proteins are inherently more temperature resistant in a cell-like milieu

One of the notable assumed characteristics of the early prebiotic Earth is the elevated temperature of the environment [60]. The temperature-induced aggregation propensity of random protein libraries was investigated by their exposure to a mild 15 min heat shock at 42°C. Interestingly, the quantity of soluble proteins in reactions without chaperones were approximately two times greater in the early alphabet library (approx. 30% versus approximately 60% for 20F and 10E libraries, respectively) which might indicate a natural tendency to withstand elevated temperature. On the other hand, addition of chaperones decreases aggregation tendencies of both 20F and 10E libraries up to almost full solubility upon heat shock treatment. This observation confirms our previous conclusions about the strong dependence of the canonical amino acid alphabet proteins on chaperone activity and extends it to aggregation prevention of the early amino acid alphabet proteins. Additionally, the fraction of protease resistant proteins remains unchanged (approx. 40%) upon heat shock for both libraries, suggesting that the proteins destabilized by elevated temperature belong to the unstructured category.

While most of the above-referenced studies reducing the composition of extant proteins toward the early set of amino acids did not observe an increase in their temperature resistance [15,16,18,55,56], we are here concerned with a comparison of unevolved sequences from the two amino acid repertoires and their inherent properties.

3.4. 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 sparse sampling, our study maps protease resistance, solubility, and temperature resistance in random sequences composed of the natural versus the early evolutionary canonical alphabets. Along with the advantages of the high-throughput approach to directly compare the two protein alphabets, the methodology applied in our study is inevitably limited by the nature of the library samples. Although the sample sizes in our experimental study are still minuscule in comparison to the vast potential random sequence space, this work presents a qualitative view on structure-forming potential within the unevolved protein domain. Protease resistance serves as a relatively low-

resolution technique to study the overall structural propensities rather than specific tertiary structure arrangements. Future studies would be needed to address detailed structural and functional properties of purified proteins that can be selected from the diverse libraries.

The analyses presented here were performed in a cell-like environment (rich in salts and cofactors) that may better represent protein formation conditions during both the origins of life and in extant biology. Under such conditions, the early alphabet sequences (i) are inherently more soluble and (ii) remain in solution when unfolded. These properties are partially achievable to full amino acid alphabet proteins through interaction with molecular chaperones which suggests a compelling argument for protein chaperone activity evolution. Interestingly, our study reports that both alphabets frequently give rise to proteolysis-resistant soluble structures, occupying up to approximately 40% of all sequences. Based on our computational investigation as well as on previous reports, we hypothesize that structure formation within this library is enabled partly by the cell-like milieu, assisted by salts, metal cations and cofactors, as well as increased tendency to form soluble oligomeric structures. Follow-up studies are suggested to further explore these findings as our initial proteolytic structure assessment does not allow for differentiation of various flavours of protein structure such as homo/hetero-oligomeric assemblies, molten globular or stable hydrophobic globular arrangements.

The properties of the random sequence libraries presented in this study have direct implications for the evolution of proteins in extant biology, as well as in the earlies pre-LUCA period. However, the results presented here as well as the suggested follow-up studies are also of prime relevance to comprehending dark protein space and to evolving novel strategies of protein design principles [61–63].

4. Methods

4.1. 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] and the amino acid ratios listed in electronic supplementary material, table S1. The randomized part of the libraries consisted of 84 amino acids; the remainder is attributed to the FLAG affinity purification site on the N-end of the construct, the hexahistidine tag on the C-end, and the and thrombin protease recognition site (ALV**PRG**S) in the middle of the construct (electronic supplementary material, figure S1).

4.2. Bioinformatic analysis

All bioinformatic analyses were performed on a sample of 200 000 sequences obtained from high-throughput sequencing of experimental DNA templates. Prediction of secondary structure potential of the studied libraries was performed by a consensus predictor as described previously [29]. It combines outputs of the spider3, psipred, predator, jnet, simpa, and GOR IV secondary structure predictors [30–35]. None of the predictors were 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 [36]. ProA algorithm is a support vector machine classifier trained on known aggregation prone and soluble sequences. The predictor combines 16 physico-chemical features which were shown to correlate with protein aggregation propensity. The output of the algorithm is per-residue binary classification based on empirically derived prediction score. We define the aggregation score of the protein as the sum of all aggregation-prone residues divided by the length of the protein sequence. Solubility of the protein libraries was predicted by Protein-Sol package and scaled solubility output was used for reporting [37]. The protein sequences and prediction results are available at the OSF platform website (https://osf.io/4e9s2/).

4.3. Preparation of experimental libraries

20F and 10E DNA libraries were synthesized commercially as two overlapping degenerate oligonucleotides (see electronic supplementary material 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 (electronic supplementary material, table S1). The overlapping oligonucleotides were annealed and extended by Klenow fragment to form double-stranded DNA (dsDNA). Annealing was performed by heating the complementary oligonucleotide mixture (48 µl total reaction volume, 2 µM final concentration of each) in NEB2 buffer provided with 200 µM dNTPs to 90° C for 2 min and cooling down to 32°C with a 1°C min^{-1} temperature gradient. The Klenow extension was performed by Klenow polymerase (NEB): 10 U of Klenow polymerase was added to annealed oligonucleotides, incubated for 5 min at 25°C, 37°C for 1 hour (polymerization step), and 50°C for 15 min (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 NH_4Ac precipitation and dissolved in RNAse-free water to a final concentration of 3 µg ul^{-1}.

The library DNA was analysed by high-throughput sequencing on Illumina MiSeq. The libraries for next generation sequencing (NGS) were prepared from 100 ng DNA samples using the NEBNext Ultra II DNA Library Prep kit (New England Biolabs) with AMPure XP purification beads (Beckman Coulter). The length of the prepared library was determined by Agilent 2100 Bioanalyzer (Agilent Technologies) and quantified by Quantus Fluorometer (Promega). The sample was sequenced on a MiSeq Illumina platform using the Miseq Reagent Kit v2 500-cycles (2 × 250) in a paired-end mode. Raw data were processed with the Galaxy platform, and sequence analysis of assembled and filtered paired reads was performed with MatLab scripts developed at Heinis laboratory [64,65]. The raw sequencing data are available at OSF platform website (https://osf.io/4e9s2/).

The protein library was expressed using the PURExpress 2.0 (GeneFrontier Corporation) recombinant *in vitro* translation system. The reaction was supplemented by 0.05% (v/v)

Triton X-100 and prepared according to manufacturer recommendations. The reaction was initiated by 3 µg of library mRNA. Expression followed for 2 h at 25°C, 30°C or 37°C.

4.4. Affinity purification of protein libraries

Expressed protein libraries were diluted 10x in binding buffer (50 mM Tris, 150 mM NaCl, 0.05% (v/v) Triton X-100, pH 7.5) and incubated for 2 h at 25°C with 3 µl 20 μl^{-1} reaction of TALON affinity purification matrix. The immobilized library was washed three times with binding buffer and eluted by addition of 20 µl 20 μl^{-1} reaction of elution buffer (50 mM Tris, 150 mM NaCl, 10 mM EDTA, 0.05% (v/v) Triton X-100, pH 7.5).

4.5. 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 analyse the quantity of total protein product, 10 µl of each reaction was quenched by addition of 40 µl of 300 µM puromycin in 50 mM Tris, 100 mM NaCl, 100 mM KCl, pH 7.5. Quenching proceeded for 30 min 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 min at 21°C, and 5 µl of supernatant was taken for SDS-PAGE analysis of the soluble fraction of the library. Both fractions were analysed by quantitative western blotting (Sigma-Aldrich Monoclonal ANTI-FLAG M2-Peroxidase (HRP) antibody, A8592) following the SDS-PAGE separation.

4.6. Lon proteolytic assay of protein libraries

Lon protease was expressed and purified according to the previously published protocol [41]. Cell free expression reactions were supplemented with 0.05% Triton X-100; reactions were prepared according to manufacturer recommendations. Libraries were expressed in the presence or absence of the DnaK chaperone (K+/K−) and in the 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. Expression proceeded in 10 µl reaction volume for 2 h 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 min at 30°C. The sample preparation of total and soluble library fractions was identical to the solubility analysis experiment described above.

4.7. Thrombin proteolytic assay of protein libraries

Cell free expression reactions were supplemented with 0.05% Triton X-100; reactions were prepared according to manufacturer recommendations. Libraries were expressed in the presence or absence of the 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 h 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 min 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 (Sigma Aldrich, USA) was added, and the total expressed library was digested for 2 h at 30°C. The soluble fraction of the library was prepared by centrifugation at 21 000*g* for 30 min at 21°C, and 5 µl of supernatant was thrombin digested according to the same protocol. Cleaved samples of the total expressed and soluble libraries were analysed by SDS-PAGE and western blotting (Sigma-Aldrich Monoclonal ANTI-FLAG M2-Peroxidase (HRP) antibody, A8592). The final quantification was performed on undigested fractions of the library proteins rather than on formation of cleavage fragments due to experimental errors in small fragments transfer and secondary cleavage of formed fragments of library 20F.

4.8. Temperature resistance assay

Libraries expressed in 10 µl volume were processed as described above in the Lon proteolytic assay protocol. The Lon absent libraries were further analysed for their temperature resistance in the presence and absence of chaperone. Processed reactions were incubated at 42°C for 15 min and immediately centrifuged at 21 000*g* for 30 min at 21°C. The 5 µl supernatant fractions were subjected to thrombin proteolysis as described previously and analysed by SDS-PAGE and quantitative western blotting.

4.9. 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.

Data accessibility. The raw library sequencing data are available on the OSF platform (https://osf.io/4e9s2/).

The library DNA template sequences, all the original western blots and statistical analyses have been uploaded as electronic supplementary material [66].

Authors' contributions. V.T.: conceptualization, data curation, formal analysis, investigation, methodology, writing—original draft; J.Vy.: data curation, formal analysis; T.N.: investigation; J.Vo.: conceptualization, formal analysis; K.F.: formal analysis, methodology, writing—review and editing; K.H.: conceptualization, formal analysis, funding acquisition, investigation, project administration, resources, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. The authors declare no competing interests.

Funding. 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.

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 regarding this manuscript. In addition, we would like to thank Kateřina Nováková for her technical help with collecting MALDI spectra.

References

1. Cleaves HJ. 2010 The origin of the biologically coded amino acids. *J. Theor. Biol.* **263**, 490–498. (doi:10.1016/j.jtbi.2009.12.014)
2. Zaia DAM, Zaia CTBV, De Santana H. 2008 Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph.* **38**, 469–488. (doi:10.1007/s11084-008-9150-5)
3. Trifonov EN. 2000 Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139–151. (doi:10.1016/S0378-1119(00)00476-5)
4. Higgs PG, Pudritz RE. 2009 A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **9**, 483–490. (doi:10.1089/ast.2008.0280)
5. Weber AL, Miller SL. 1981 Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* **17**, 273–284. (doi:10.1007/BF01795749)
6. Freeland S. 2010 'Terrestrial' amino acids and their evolution. *Amin. Acids, Pept. Proteins Org. Chem.* **1**, 43–75.
7. Philip GK, Freeland SJ. 2011 Did evolution select a nonrandom 'alphabet' of amino acids? *Astrobiology* **11**, 235–240. (doi:10.1089/ast.2010.0567)
8. Ilardo M *et al.* 2019 Adaptive properties of the genetically encoded amino acid alphabet are inherited from its subsets. *Sci. Rep.* **9**, 1–9. (doi:10.1038/s41598-019-47574-x)
9. Pace NR. 2001 The universal nature of biochemistry. *Proc. Natl Acad. Sci. USA* **98**, 805–808. (doi:10.1073/pnas.98.3.805)
10. Holliday GL, Fischer JD, Mitchell JBO, Thornton JM. 2011 Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *FEBS J.* **278**, 3835–3845. (doi:10.1111/j.1742-4658.2011.08190.x)
11. Di Mauro E, Dunker AK, Trifonov EN. 2012 Disorder to order, nonlife to life: in the beginning there was a mistake. In *Genesis—In the beginning: precursors of life, chemical models and early biological evolution* (ed. J Seckbach), pp. 415–435. Berlin, Germany: Springer.
12. Tanaka J, Doi N, Takashima H, Yanagawa H. 2010 Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci.* **19**, 786–795. (doi:10.1002/pro.358)
13. Newton MS, Morrone DJ, Lee KH, Seelig B. 2019 Genetic code evolution investigated through the synthesis and characterisation of proteins from reduced-alphabet libraries. *ChemBioChem* **20**, 846–856. (doi:10.1002/cbic.201800668)
14. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. 1997 Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805–809. (doi:10.1038/nsb1097-805)
15. Longo LM, Lee J, Blaber M. 2013 Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc. Natl Acad. Sci. USA* **110**, 2135–2139. (doi:10.1073/pnas.1219530110)
16. Shibue R, Sasamoto T, Shimada M, Zhang B, Yamagishi A, Akanuma S. 2018 Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* **8**, 1–8. (doi:10.1038/s41598-018-19561-1)
17. Solis AD. 2019 Reduced alphabet of prebiotic amino acids optimally encodes the conformational space of diverse extant protein folds. *BMC Evol. Biol.* **19**, 1–19. (doi:10.1186/s12862-019-1464-6)
18. Giacobelli V, Fujishima K, Lepšík M, Tretyachenko V, Kadavá T, Bednárová L, Novák P, Hlouchová K. 2022 In vitro evolution reveals noncationic protein-RNA interaction mediated by metal ions. *Mol. Biol. Evol.* **39**, msac032. (doi:10.1093/molbev/msac032)
19. Longo LM, Despotović D, Weil-Ktorza O, Walker MJ, Jabłońska J, Fridmann-Sirkis Y, Varani G, Metanis N, Tawfik DS. 2020 Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl Acad. Sci. USA* **117**, 15 731–15 739. (doi:10.1073/pnas.2001989117)
20. Houben B *et al.* 2020 Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues. *EMBO J.* **39**, 1–22. (doi:10.15252/embj.2019102864)
21. Keefe AD, Szostak JW. 2001 Functional proteins from a random-sequence library. *Nature* **410**, 715–718. (doi:10.1038/35070613)
22. Chiarabelli C, Vrijbloed JW, Thomas RM, Luisi PL. 2006 Investigation of de novo totally random biosequences. *Chem. Biodivers.* **3**, 827–839. (doi:10.1002/cbdv.200690087)
23. Labean TH, Butt TR, Kauffman SA, Schultes EA. 2011 Protein folding absent selection. *Genes (Basel).* **2**, 608–626. (doi:10.3390/genes2030608)
24. Yu JF, Cao Z, Yang Y, Wang CL, Su ZD, Zhao YW, Wang JH, Zhou Y. 2016 Natural protein sequences are more intrinsically disordered than random sequences. *Cell. Mol. Life Sci.* **73**, 2949–2957. (doi:10.1007/s00018-016-2138-9)
25. Tretyachenko V *et al.* 2017 Random protein sequences can form defined secondary structures and are well-tolerated *in vivo*. *Sci. Rep.* **7**, 1–9. (doi:10.1038/s41598-017-15635-8)
26. Tretyachenko V, Voracek V, Soucek R, Fujishima K, Hlouchova K. 2021 CoLide: combinatorial library design tool for probing protein sequence space. *Bioinformatics* **37**, 482–489. (doi:10.1093/bioinformatics/btaa804)
27. Bateman A *et al.* 2021 UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489. (doi:10.1093/nar/gkaa1100)
28. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004 WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190. (doi:10.1101/gr.849004)
29. Vymětal J, Vondrášek J, Hlouchová K. 2019 Sequence versus composition: what prescribes IDP biophysical properties? *Entropy* **21**, 1–8. (doi:10.3390/e21070654)
30. Garnier J, Gibrat JF, Robson B. 1996 GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540–553. (doi:10.1016/S0076-6879(96)66034-0)
31. Frishman D, Argos P. 1997 Seventy-five percent accuracy in protein secondary structure prediction. *Proteins Struct. Funct. Genet.* **27**, 329–335. (doi:10.1002/(SICI)1097-0134(199703)27:3<329::AID-PROT1>3.0.CO;2-8)
32. Levin JM. 1997 Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.* **10**, 771–776. (doi:10.1093/protein/10.7.771)
33. Jones T. 1999 Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202. (doi:10.1006/jmbi.1999.3091)
34. Cuff JA, Barton GJ. 2000 Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**, 502–511. (doi:10.1002/1097-0134(20000815)40:3<502::AID-PROT170>3.0.CO;2-Q)
35. Heffernan R, Paliwal K, Lyons J, Singh J, Yang Y, Zhou Y. 2018 Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem.* **39**, 2210–2216. (doi:10.1002/jcc.25534)
36. Fang Y, Gao S, Tai D, Middaugh CR, Fang J. 2013 Identification of properties important to protein aggregation using feature selection. *BMC Bioinform.* **14**, 314. (doi:10.1186/1471-2105-14-314)
37. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. 2017 Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098–3100. (doi:10.1093/bioinformatics/btx345)
38. Niwa T, Ying BW, Saito K, Jin W, Takada S, Ueda T, Taguchi H. 2009 Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl Acad. Sci. USA* **106**, 4201–4206. (doi:10.1073/pnas.0811922106)
39. Johannes S *et al.* 2012 Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682. (doi:10.1038/nmeth.2019)
40. Melderen LV, Aertsen A. 2009 Regulation and quality control by Lon-dependent proteolysis. *Res. Microbiol.* **160**, 645–651. (doi:10.1016/j.resmic.2009.08.021)
41. Niwa T, Uemura E, Matsuno Y, Taguchi H. 2019 Translation-coupled protein folding assay using a protease to monitor the folding status. *Protein Sci.* **28**, 1252–1261.
42. White SH. 1994 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. (doi:10.1007/BF00163155)

43. Bornberg-Bauer E, Hlouchova K, Lange A. 2021 Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* **68**, 175–183. (doi:10.1016/j.sbi.2020.11.010)
44. Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017 Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 1–7. (doi:10.1038/s41559-017-0127)
45. Kamtekar S, Hecht MH. 1995 Protein motifs. 7. The four-helix bundle: what determines a fold? *FASEB J.* **9**, 1013–1022. (doi:10.1096/fasebj.9.11.7649401)
46. Rojas NR, Kamtekar S, Simons CT, McLean JE, Vogel KM, Spiro TG, Farid RS, Hecht MH. 1997 De novo heme proteins from designed combinatorial libraries. *Protein Sci.* **6**, 2512–2524.
47. Davidson AR, Sauer RT. 1994 Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA* **91**, 2146–2150. (doi:10.1073/pnas.91.6.2146)
48. Benner SA. 1989 Enzyme kinetics and molecular evolution. *Chem. Rev.* **89**, 789–806. (doi:10.1021/cr00094a004)
49. Raggi L, Bada JL, Lazcano A. 2016 On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. *Phys. Chem. Chem. Phys.* **18**, 20 028–20 032. (doi:10.1039/C6CP00793G)
50. Brack A, Orgel LE. 1975 Beta structures of alternating polypeptides and their possible prebiotic significance. *Nature* **256**, 383–387. (doi:10.1038/256383a0)
51. Kovacs NA, Petrov AS, Lanier KA, Williams LD. 2017 Frozen in time: the history of proteins. *Mol. Biol. Evol.* **34**, 1252–1260. (doi:10.1093/molbev/msx086)
52. Lupas AN, Alva V. 2017 Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J. Struct. Biol.* **198**, 74–81. (doi:10.1016/j.jsb.2017.04.007)
53. Lawrence MS, Phillips KJ, Liu DR. 2007 Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* **129**, 10 110–10 112. (doi:10.1021/ja071641y)
54. Kimura M, Akanuma S. 2020 Reconstruction and characterization of thermally stable and catalytically active proteins comprising an alphabet of ∼13 amino acids. *J. Mol. Evol.* **88**, 372–381. (doi:10.1007/s00239-020-09938-0)
55. Makarov M *et al.* 2021 Enzyme catalysis prior to aromatic residues: reverse engineering of a dephospho-CoA kinase. *Protein Sci.* **30**, 1022–1034. (doi:10.1002/pro.4068)
56. Longo LM, Tenorio CA, Kumru OS, Middaugh CR, Blaber M. 2015 A single aromatic core mutation converts a designed 'primitive' protein from halophile to mesophile folding. *Protein Sci.* **24**, 27–37. (doi:10.1002/pro.2580)
57. Despotović D, Longo LM, Aharon E, Kahana A, Scherf T, Gruic-Sovulj I, Tawfik DS. 2020 Polyamines mediate folding of primordial hyperacidic helical proteins. *Biochemistry* **59**, 4456–4462. (doi:10.1021/acs.biochem.0c00800)
58. Wang MS, Hoegler KJ, Hecht MH. 2019 Unevolved de novo proteins have innate tendencies to bind transition metals. *Life* **9**, 8.
59. Yadid I, Kirshenbaum N, Sharon M, Dym O, Tawfik DS. 2010 Metamorphic proteins mediate evolutionary transitions of structure. *Proc. Natl Acad. Sci. USA* **107**, 7287–7292. (doi:10.1073/pnas.0912616107)
60. Islas S, Velasco AM, Becerra A, Delaye L, Lazcano A. 2003 Hyperthermophily and the origin and earliest evolution of life. *Int. Microbiol.* **6**, 87–94. (doi:10.1007/s10123-003-0113-4)
61. Woolfson DN. 2021 A brief history of de novo protein design: minimal, rational, and computational. *J. Mol. Biol.* **433**, 167160. (doi:10.1016/j.jmb.2021.167160)
62. Hecht MH, Zarzhitsky S, Karas C, Chari S. 2018 Are natural proteins special? Can we do that? *Curr. Opin. Struct. Biol.* **48**, 124–132. (doi:10.1016/j.sbi.2017.11.009)
63. Tong CL, Lee KH, Seelig B. 2021 De novo proteins from random sequences through in vitro evolution. *Current Opinion in Structural Biology* **68**, 129–134. (doi:10.1016/j.sbi.2020.12.014)
64. Rebollo IR, Sabisz M, Baeriswyl V, Heinis C. 2014 Identification of target-binding peptide motifs by high-throughput sequencing of phage-selected peptides. *Nucleic Acids Res.* **42**, e169. (doi:10.1093/nar/gku940)
65. Afgan E *et al.* 2018 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544. (doi:10.1093/nar/gky379)
66. Tretyachenko V, Vymětal J, Neuwirthová T, Vondrášek J, Fujishima K, Hlouchová K. 2022 Modern and prebiotic amino acids support distinct structural profiles in proteins. Figshare. (doi:10.6084/m9.figshare.c.6032490)