ChemPubSoc Europe

DOI: 10.1002/cbic.201800668

CHEMBIOCHEM Full Papers

Genetic Code Evolution Investigated through the Synthesis and Characterisation of Proteins from Reduced-Alphabet Libraries

Matilda S. Newton[+],[a, b] Dana J. Morrone[+],[a, b] Kun-Hwa Lee,[a, b] and Burckhard Seelig*[a, b]

The universal genetic code of 20 amino acids is the product of evolution. It is believed that earlier versions of the code had fewer residues. Many theories for the order in which amino acids were integrated into the code have been proposed, considering factors ranging from prebiotic chemistry to codon capture. Several meta-analyses combined these theories to yield a feasible consensus chronology of the genetic code's evolution, but there is a dearth of experimental data to test the hypothesised order. We used combinatorial chemistry to synthesise libraries of random polypeptides that were based on different subsets of the 20 standard amino acids, thus representing different stages of a plausible history of the alphabet. Four libraries were comprised of the five, nine, and 16 most ancient amino acids, and all 20 extant residues for a direct side-by-side comparison. We characterised numerous variants from each library for their solubility and propensity to form secondary, tertiary or quaternary structures. Proteins from the two most ancient libraries were more likely to be soluble than those from the extant library. Several individual protein variants exhibited inducible protein folding and other traits typical of intrinsically disordered proteins. From these libraries, we can infer how primordial protein structure and function might have evolved with the genetic code.

Introduction

The genetic code of 20 amino acids we know today is not static. Although supposedly "universal",[1] the code has been found to contain a 21st and a 22nd amino acid (selenocysteine and pyrrolysine) in many extant organisms.[2,3] The addition of these residues indicates that the code is not "frozen",[4] but can expand. The 64 codons could, in principle, evolve to encode as many as 63 chemically distinct amino acids. As the genetic code shows signs of expansion, it is reasonable to assume that, at earlier points in time, it coded for fewer amino acids. Likewise, it is inconceivable that the code would have arisen fully formed. Many lines of evidence and reasoning indicate that earliest life—before the last universal common ancestor (LUCA)—used a genetic code consisting of fewer amino acids.[5–8] Further amino acids were gradually added to result in the chemical complexity we find in the code today.

Numerous hypotheses have been proposed that attempt to define a plausible chronological order for the emergence of amino acids in the genetic code. Arguments include the famous Miller–Urey spark experiment; the hypothesis on the importance of complementarity; the thermostability of the triplet code; and the codon-capture theory.[9–12] Other analyses consider the physicochemical properties of amino acids like size, charge and hydrophobicity; the code's co-evolution; biosynthetic theories; or the detection of amino acids on meteorites, to name a few.[13–16] A common trend among the hypotheses is that the earliest amino acids were comparably small; this is supported by the observation that prebiotic synthesis seems to favour less complex amino acids.[17,18] It is widely agreed that the most recent additions to the genetic code include the aromatic amino acids phenylalanine, tyrosine and tryptophan.[17,18] Unbiased, comprehensive meta-analyses of the proffered hypotheses and experimental evidence suggest a consensus chronological order for the incorporation of amino acids into the genetic code[19,20] (Figure 1 A).

The evolution of the genetic code resulted in a shifting proteome composition over time, thereby altering the chemical properties of primordial proteins. As a result, these proteins probably also changed in their biophysical and functional properties, which is the particular focus of this work. Although numerous theories have been proposed describing the nature of these early amino acid alphabets and the gradual addition of "later" amino acids to give today's genetic code, there are still only a few experimental data to link these hypotheses to the biochemical characteristics of the corresponding proteins.

There is a precedent for studying proteins with a reduced amino acid composition. Several groups have presented examples of extant proteins that still fold and function correctly when their amino acid complexity is artificially reduced;[2–23]

[a] Dr. M. S. Newton,[+] Prof. D. J. Morrone,[+] K.-H. Lee, Prof. B. Seelig
Department of Biochemistry, Molecular Biology and Biophysics
University of Minnesota
Minneapolis, MN, 55455 (USA)

[b] Dr. M. S. Newton,[+] Prof. D. J. Morrone,[+] K.-H. Lee, Prof. B. Seelig
BioTechnology Institute, University of Minnesota
1479 Gortner Avenue, 140 Gortner Laboratory
St. Paul, MN, 55108-6106 (USA)
E-mail: seelig@umn.edu

[+] These authors contributed equally to this work.

Supporting information and the ORCID identification numbers for the authors of this article can be found under https://doi.org/10.1002/cbic.201800668.

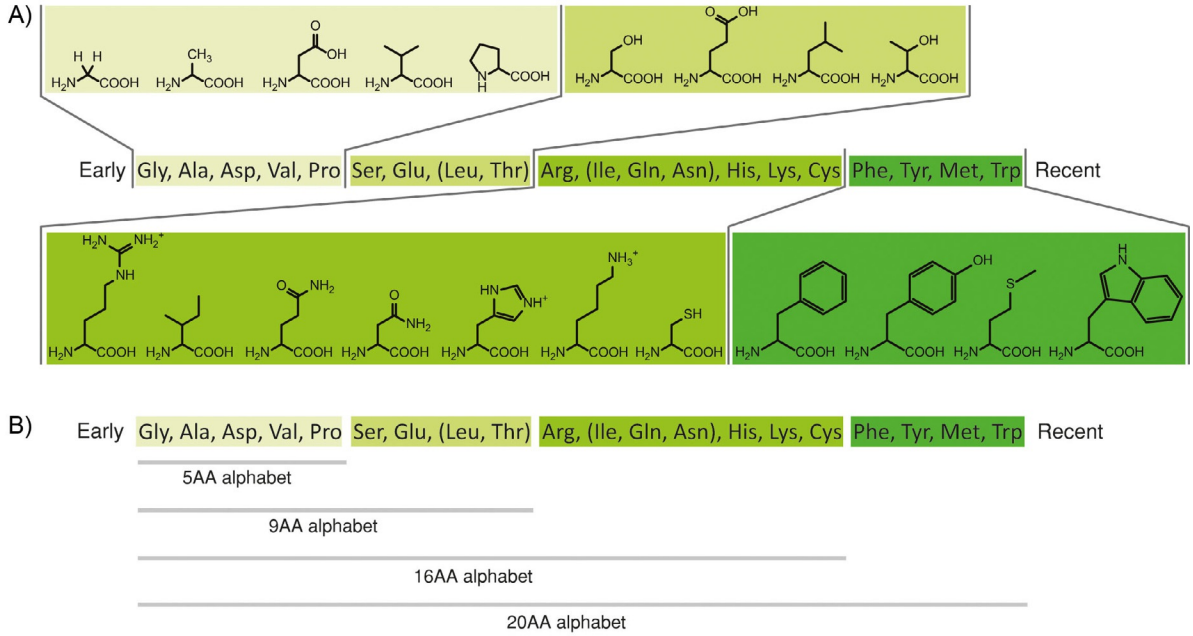

Figure 1. Plausible history of amino acid additions to the genetic code and reduced amino acid alphabets from different stages of code evolution. A) The consensus chronological order of incorporation of amino acids into the genetic code over time.[19] Amino acids of approximately the same proposed age are depicted in parentheses. B) The four different amino acid alphabets that were used to synthesise the four polypeptide libraries.

still other projects used rational design techniques to generate de novo, folded proteins with as few as ten different amino acids.[24] Although this work confirmed that a limited contingent of amino acids can support functional proteins, the amino acid alphabets used in those example studies[21,24–28] were chosen to reduce the chemical redundancy of the extant code for engineering purposes, rather than to represent a likely early genetic code.

We experimentally probed the properties of reduced-alphabet proteins that represent snapshots along the most likely chronology of the genetic code's evolution. Using the consensus temporal order of amino acid additions to the genetic code,[19] we defined three primordial protein alphabets that comprise the likely earliest five, nine and 16 amino acids (Figure 1 B). From each alphabet, we synthesised protein libraries of random sequence—each 83 amino acids in length. Proteins of comparable size have previously been found to have biological activity[29,30] and it is reasonable to presume that a pre-LUCA organism would have only been able to support a limited genome, thus favouring short genes.[31,32] For comparison, we also synthesised a fourth library comprised of the canonical 20 amino acids.

The four libraries of five, nine, 16 and 20 were generated through trimer chemical synthesis. Although it has been standard practice to generate such protein libraries through the use of degenerate codons,[33,34] no degenerate codons exist that would code for the reduced alphabets described. Therefore, we synthesised our libraries through the sequential coupling of mixtures of trinucleotide phosphoramidites,[35] with each trimer representing a desired codon. This method uniquely allowed us to generate libraries comprised of only the desired amino acids in the precise ratios we desired, something that usually cannot be achieved with degenerate codon synthesis.

This study aimed to characterise the impact of amino acid composition on the biophysical properties of the reduced-alphabet protein variants as a first step to investigating the effect that an evolving, increasingly complex genetic code would have upon protein structure and function. Previous studies have attempted to estimate the function of ancient hypothetical genetic codes,[36–40] but a major strength of our approach is in the quality of our libraries. Specifically, we designed the libraries to strictly adhere to the reduced alphabets within the randomised regions, with no compromise for non-library amino acids. Furthermore, the fully random nature of the sequences—rather than construction from a pre-existing protein scaffold—allowed an unbiased exploration of the folding behaviours of the alphabets. Finally, the approach of generating multiple libraries for different points in the code's evolution allowed nuanced comparison and insight into the changing chemistries of primordial proteins. To our knowledge, this work therefore contains the most accurate assembly of reduced alphabets to date and importantly links experimental data to hypotheses on early alphabets.

To determine the propensity of the four alphabets to yield soluble, stable structures, we expressed about two dozen protein variants from each library in *Escherichia coli* and analysed their secondary-structure content, hydrophobic packing and oligomerisation state. The reduced-library proteins demonstrated a surprising propensity for solubility when expressed—with the five- and nine-amino-acid alphabets showing the highest proportion of soluble protein, probably due in part to their

overabundance of negatively charged residues. These analyses also showed that the proteins lack significant secondary structure and could be described as intrinsically disordered proteins.

Results

Library construction and compositions

We synthesised DNA libraries to encode random polypeptide libraries of different amino acid alphabets to represent "snapshots" in the likely evolutionary development of the genetic code. Accordingly, each library was constructed to encode an increasingly complex set of amino acids (Figures 1B and 2), conforming to the consensus order reported by Trifonov.[19] We acknowledge that some finer details of this consensus order might still be a matter of debate, but the general order of very early amino acids and late amino acid additions to the genetic code is widely accepted. Our most reduced library was set to have just the five most ancient amino acids (5AA: Gly, Ala, Asp, Val, Pro), as a previous study showed that a small β-sheet protein could be engineered to contain only five different amino acids but still fold properly.[46] For our next point in the evolution of the genetic code, we chose a library of nine amino acids, which are all believed to easily form prebiotically[20] (9AA: Gly, Ala, Asp, Val, Pro, Ser, Glu, Leu, Thr). These nine contained the five amino acids from the first library and also added residues that significantly increase the polarity and charge of the proteins they encode. Next, we constructed a library from 16 amino acids (16AA: Gly, Ala, Asp, Val, Pro, Ser, Glu, Leu, Thr, Arg, Ile, Gln, Asn, His, Lys, Cys) that included all extant residues except the four amino acids that are generally thought be the most recent additions to the genetic code. We chose this particular composition because this library lacks only the aromatic amino acids and methionine, the most common start codon (although we were constrained to maintain a single ATG start codon). Finally, we made a library (20AA) composed of the 20 canonical extant amino acids for comparison.

Proteins from each library had a randomised region of 80 amino acids that conformed to its given alphabet and was constructed from chemically synthesised 20-codon gene cassettes that were subsequently ligated.[47] A length of 20 trinucleotides (60 nt) was chosen due to the limitations of accurate chemical synthesis and had the benefit of shuffling fragments to increase sequence diversity further. Trimer phosphoramidites were used in chemical synthesis—as opposed to classic monomer-based sequential synthesis—which allowed precise control over codon usage. The relative amino acid abundance in each library was based on residue usage in today's proteome (as defined by the ExPASy server,[48] see Table S1 in the Supporting Information). This amino acid usage was confirmed by Sanger sequencing (Tables S2 and S3, Figure S1).

The stretch of 20 trinucleotides was flanked at the 5′ and 3′ termini by conserved sequences of 20 nt to be used as primer binding sites and to introduce restriction endonuclease recognition sites. Type IIS restriction sites were used, as these generate cohesive ends outside their recognition sequences and thus avoid undesired cloning artefacts that would code for amino acids not included in some alphabets. Following PCR amplification and endonuclease digestion, cassettes were then ligated together to build up an open reading frame of 40 codons and then, following a second cycle of digestion and ligation, 80. This cloning strategy enabled us to have cloning scars of only a single glycine at each ligation junction. Conveniently, glycine is part of all four libraries.

We performed the ligation reactions for the libraries at a scale that enabled the net production of more than 10^{12} unique sequence variants in each of the four libraries. This vast diversity was desired to take full advantage of the high-throughput capacity of in vitro selection methods, such as the mRNA display technology, which we plan to use in subsequent projects. Following the construction of the libraries, variants were cloned into expression plasmids and submitted for sequence verification. Sanger sequencing of 206 library variants

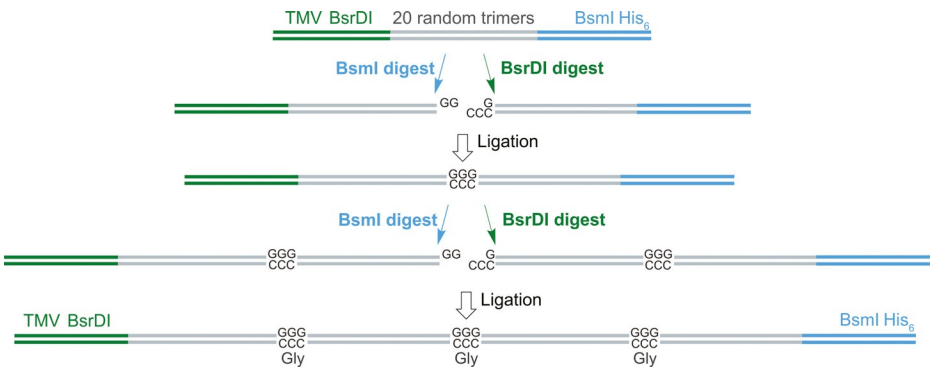

Figure 2. Construction of the DNA library encoding the protein library. A chemically synthesised cassette of 20 random nucleotide trimers was concatemerised to yield a DNA library encoding 80 random amino acid positions. The starting cassette was flanked by constant termini. Two aliquots of the library were separately digested with type IIS restriction enzymes BsrDI and BsmI. These digested fragments were ligated at the resulting complementary overhangs to produce a GGG scar (codon for Gly) and double the length of the random trimer region. Digestion and ligation was repeated to provide a final library of 80 random codons with three glycine scars at defined positions and flanked by the constant termini. Each library was constructed separately with four individual cassettes, each encoding the appropriate alphabet.

revealed that 60% of the sequences exactly conformed to their design and alphabet. The other 40% comprised sequences with indels, truncations or missense substitutions (Table S4).

Characterisation of expression and solubility of protein library variants

Sequence-confirmed variants ($\geq$20) from each library were expressed as C-terminally His_6-tagged proteins for purification and detection purposes. Initial test expressions determined that induction with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) yielded the most soluble protein. Following the induction of variants, the soluble and insoluble (solubilised with 8 M urea) fractions were analysed by SDS-PAGE. Although SDS-PAGE and Coomassie staining were sufficient to identify most over-expressed variants in the 16 AA and 20 AA libraries, the 5 AA and 9 AA libraries showed lower-abundance proteins that were easier to detect by the more sensitive western blot analysis with a His_6-tag-binding antibody.

We categorised protein variants as "no expression" (no protein detectable in the soluble or insoluble fraction), "highly soluble" (when $\geq$80% of the total detected protein was in the soluble lane), "partially soluble" (21–79% in the soluble lane), or "no/low soluble" ($\leq$20% in the soluble lane; Table 1 and Figures 3 and S2). Despite using sensitive western blot analysis, which allowed the detection of low-abundance proteins without obfuscation by endogenous protein, no overexpression was detected for 33–65% of protein variants from the different libraries (5 AA: 13 out of 20 tested; 9 AA: 13 out of 24 tested; 16 AA: 8 out of 23 tested; 20 AA: 8 out of 24 tested; Figure S2). Of the expressible proteins, variants from the 20 AA library showed the lowest solubility with 11 of 16 variants having had $\leq$20% soluble protein. All three reduced-alphabet libraries showed higher solubility than the library from the extant alphabet, from which only one variant (20-5) showed high ($\geq$80%) solubility. For each of the other three libraries, we found at least five variants with a solubility as high as that of variant 20-5. For several insoluble variants from different libraries, refolding procedures by denaturing/renaturing were attempted to obtain soluble protein, but these methods were not successful. Select proteins were expressed at a larger 1 L scale and purified by His_6 affinity chromatography for subsequent biophysical characterisation.

Table 1. Distribution of protein solubility for all variants tested from the four libraries, as determined by SDS-PAGE and western blot analysis.

	Alphabet			
	5 AA	9 AA	16 AA	20 AA
expressed in *E. coli*	7	11	15	16
high solubility ($\geq$80%)	6	8	10	1
partial solubility (21–79%)	0	3	3	4
no/low solubility ($\leq$20%)	1	0	2	11
no expression in *E. coli*	13	13	8	8
total variants analysed	20	24	23	24

Biophysical characterisation of soluble protein variants

Determining secondary structure by circular dichroism: The CD spectra of select soluble variants from each library indicated that the proteins existed primarily in extended random-coil form, as demonstrated by the absorbance minima between 198–202 nm (Figures 4 and S3). However, several variants displayed inducible increases in α-helicity. None of the variants, when analysed under buffer-only conditions, showed significant CD spectral characteristics of α-helical or β-strand-containing regions. The addition of 2,2,2-trifluoroethanol (TFE) to a protein solution induces or stabilises secondary structure, particularly α-helical regions.[49,50] TFE's effects on protein structure might be explained either by its promoting intramolecular hydrogen bonding, stabilising hydrophobic regions, or by disrupting protein–solvent interactions.[50,51] Twelve out of 13 library variants tested (5-11, 5-18, 5-20, 9-2, 9-4, 9-6, 9-21, 16-1, 16-13, 16-18, 16-23, 20-5) showed altered spectra in the presence of 25 or 50% TFE. The single minimum characteristic of random coil shifted right to 202–206 nm, and a second negative peak at ≈220 nm was formed, which together indicated increased α-helical character (an all α-helical protein typically shows CD spectrum minima at 205 and 222 nm). The *H* values, a measure of helix content,[45] also increased, thus reflecting this spectral shift (Table S5). Variants 16-18, 9-6, 20-5, 5-11 and 16-23 showed the greatest helical content (>60%) in the presence of 25 or 50% TFE, thus indicating that the effect is not alphabet-specific.

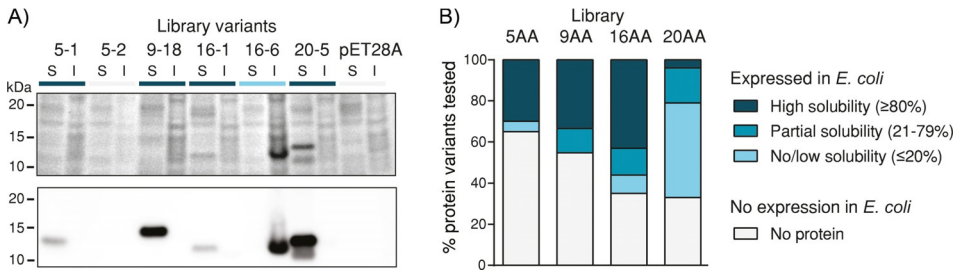

Figure 3. Solubility analysis of library variants by SDS-PAGE. A) Library variants were expressed in *E. coli*, and the soluble and insoluble fractions were separated by SDS-PAGE. The upper panel shows a small representative subset of variants tested on a gel stained with Coomassie Brilliant Blue. The lower panel shows the same gel area but with the library variants visualised by anti-His_6 western blot. The library variant names are indicated at the top; pET28A refers to a negative control of cells carrying an empty expression vector under identical expression conditions. B) Summary of the solubility analysis of all 91 protein variants tested.

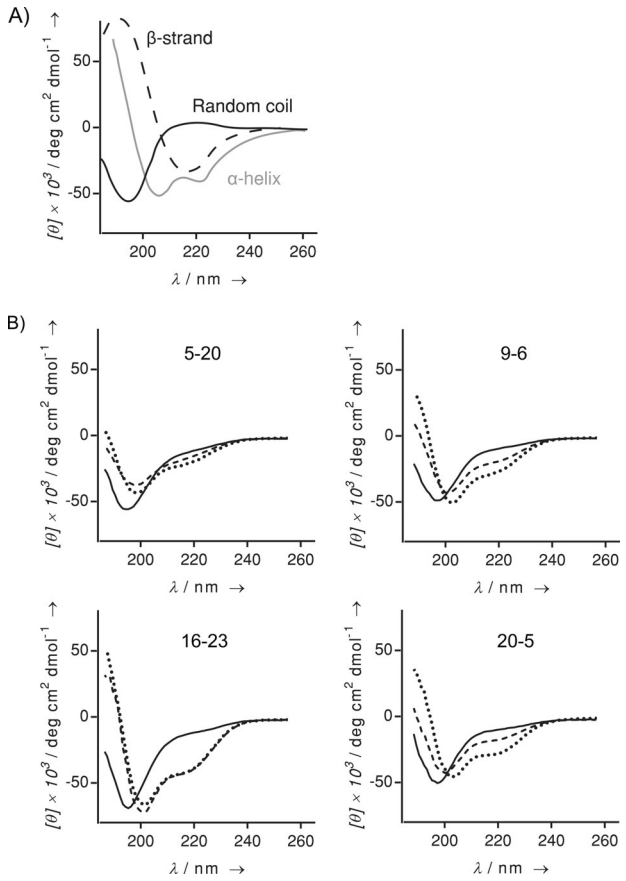

Figure 4. Secondary-structure content of soluble variants from the four libraries measured by CD spectroscopy. A) Reference CD spectra are shown for all random-coil (——), α-helix (——) and β-strand (-----) proteins. B) One typical Ni-NTA-purified variant protein from each library. The spectra are presented as molar ellipticities ($\theta \times 10^3$ deg cm^2 dmol^{-1}) for protein ——: in buffer, -----: in the presence of 25% trifluoroethanol (TFE) and ····: in 50% TFE. Additional CD spectra are shown in Figure S3.

Assessing protein folding by ANS fluorescence: The tertiary structure content of proteins can be assessed with the help of the amphiphilic dye 1-anilinonaphthalene-8-sulfonate (ANS), which increases its fluorescence when shielded from water, such as when it binds hydrophobic areas of proteins. Well-folded proteins with highly packed hydrophobic cores display ANS fluorescence. Molten globules—partially folded proteins with distinct secondary but little or no tertiary structure—exhibit even greater ANS fluorescence, as hydrophobic regions that would be buried in well-folded protein are more easily accessible to the dye. Library variants were subjected to ANS binding, followed by excitation and fluorescence detection (Figure 5). For comparison, we included two well-folded proteins: ribonuclease A and BSA, as well as our artificial RNA ligase,[29,52] which is a small 87-residue protein that has an unusual three-dimensional structure with high conformational dynamics.[53] Of the 13 library variants examined, only variant 16-1 exhibited fluorescence similar to that of the well-folded control proteins (Figure 5). This variant also exhibited the highest inducible helical structure detectable by CD. Although the variant displayed several fold less fluorescence than ribonuclease A and BSA, it had substantially more ANS fluorescence than all other variants tested, the spectra for which were all similar to that of the atypical ligase 10C protein.

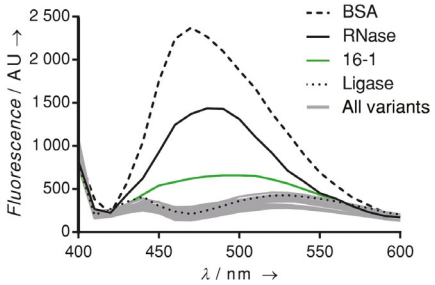

Figure 5. ANS fluorescence emission spectra for library variants. Proteins (10 μM) were mixed with 500 μM ANS and excited at 385 nm. The fluorescence emission spectra for BSA (-----), ribonuclease A (RNase; ——), the RNA ligase 10C (····), variant 16-1 (——) are shown; all remaining tested soluble variants (——) gave similar spectra.

Characterising the oligomeric state by size-exclusion chromatography: Size-exclusion chromatography was used to determine the oligomeric state of protein variant 20-1, which was the single highly soluble protein from the 20 AA library. Soluble variants from the three reduced-alphabet libraries were not amenable to this method because detection by UV absorption requires the presence of aromatic residues, which were only present in 20 AA library variants. Using a Superdex 75 column, which can separate proteins ranging in MW from 3000 to 70 000 Da, we found that the solution of the 20-1 variant (predicted mass of 10 kDa) yielded multiple peaks with retention times between $t_R = 10.4$ min and $t_R = 16.9$ min, thus indicating that the protein exists in multiple quaternary states (Figure 6). For comparison, we also analysed two standards of albumin and ribonuclease A of 65 and 13 kDa, respectively.

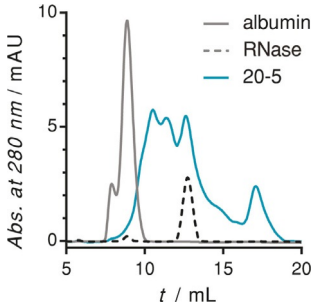

Figure 6. Assessment of the oligomeric state of the purified soluble variant 20-5 (——) by size-exclusion chromatography. Monomeric standards of a 13 kDa ribonuclease (-----) and a 65 kDa albumin (——) are shown for comparison. The calculated molecular weight of variant 20-5 is 10 kDa.

Discussion

To experimentally test a hypothesised succession of amino acid inclusion during the evolution of the genetic code, we generated four randomised protein libraries that represent different stages in the code's history. The libraries of 5 AA, 9 AA,

16 AA, and the extant 20 AA contained 80-residue randomised regions that conformed strictly to their defined amino acid alphabet. Randomly chosen variants from each library were characterised for solubility by recombinant expression and further analysed for their propensity for secondary, tertiary and quaternary structures by CD spectroscopy, ANS fluorescence and size-exclusion chromatography, respectively.

The libraries we constructed represent a vigorous experimental exploration of the evolution of the genetic code, thereby complementing and advancing prior studies. Although other groups have generated reduced amino acid libraries before, our study has two major strengths over previous work: firstly, our library design conforms strictly to the defined alphabets and secondly, the construction of four libraries increases the power of valuable comparison. Trinucleotide synthesis, used in the current work, allows precise control of the amino acid alphabet and relative abundance. We further ensured that the cloning protocol used (Figure 2) did not introduce non-library residues at the cloning scars from cassette ligations. Sanger sequencing of 206 individual variants showed that 60% of our library variants conformed exactly to their intended alphabet, without truncations, frameshifts or non-alphabet missense mutations (Table S4). For comparison, Kumachi et al. selected a 30-residue tRNA-binding peptide from a four-amino-acid code library (using the degenerate codon GNC: Gly, Ala, Asp, Val) but found that only 13% of all variants (6/46) conformed to their alphabet as determined by sequencing after the selection.[40] Some degree of mutation that leads to undesirable non-alphabet codons is inevitable during library synthesis and propagation. However, techniques such as controlling translation conditions through cell-free expression systems[54,55] could further improve the adherrance of a library to its alphabet in future applications.

Library design is just as important for quality libraries as accurate synthesis. Notable previous work by Yanagawa and colleagues similarly generated multiple libraries (using codons GNN, RNN, and NNN for the early five, 12, and all 20 amino acids, respectively),[39] but used a library construction strategy that resulted in regularly interspersed non-alphabet amino acids: cloning scars introduced nine additional tryptophan residues within every library variant, corresponding to every 16th residue. Unlike the invariant glycine cloning scar in our libraries, tryptophan is considered to have been the last addition to the genetic code,[19] and a high tryptophan content would be likely to have a great influence upon the biophysical properties of proteins, especially in the case of the libraries from the reduced alphabets of early amino acids. Furthermore, although degenerate codons allow easy, low-cost library generation, they do not always allow the researcher to sample the exact cohort of desired amino acids or the ratios in which they appear. For example, GNN and RNN libraries do not strictly follow the consensus chronology[19] explored here, due to the arrangement of the code. In contrast, the use of trinucleotide phosphoramidites allows one to precisely manipulate both the choice of amino acids and the relative abundance of each. The amino acid distributions of our four libraries were designed to mimic the extant proteome (determined through ExPASy[48]). Sanger sequencing demonstrated that the amino acid distribution among the correct sequences in each library showed 70–80% agreement with the intended distribution (Figure S1 and Table S3). Finally, the number of our libraries, which represent four different time points in the code's history, allows valuable comparison of biophysical characteristics and potential for cellular function. Although other groups have probed the function of ancient alphabets—in addition to earlier examples, cofactor binding was examined for VNM codon libraries encoding the oldest 15 residues[36,37]—none, to our knowledge, has compared multiple libraries strictly conforming to a consensus order[19] for the genetic code's evolution.

All four libraries described in our study were biophysically characterised independently under identical conditions to allow comparison of the different alphabets. Heterologous expression of individual library variants in *E. coli* BL21(DE3) cells indicated that the extant (20 AA) and most recent (16 AA) alphabets are best-suited to expression (independent of solubility) under this model system. These libraries had the highest fraction of expressed proteins detectable by western blot to total variants tested (16 AA: 15/23; 20 AA: 16/24; Table 1 and Figures 3 B and S2), followed by the 9 AA library (11/24). The 5 AA library was least amenable to heterologous expression (7/20). The proteins detected for the 9 AA and 5 AA libraries were also, on average, expressed at lower abundance than proteins from the more modern alphabets. Proteins from these libraries appeared as fainter bands in the western blots and required a longer detection time (Figure S2). These two observations probably indicate lower protein abundance in the cell, but decreased affinity of proteins from different alphabets to the polyvinylidene difluoride (PVDF) membrane used in western blots could have some effect too. Furthermore, our detection methods did not distinguish between low protein abundance due to low expression (for which the limited availability of charged tRNA for our reduced alphabets could be a contributing factor) or high degradation rates. Tanaka et al. reported an opposite expression trend: whereas 60% of the proteins from their most diverse NNN library could be expressed (compared to 77% for our 20 AA library), as high as 87 and 86% of proteins from their reduced-alphabet RNN and GNN libraries could be expressed, respectively.[39] This drastically different trend might be due in part to the high abundance of the modern amino acid tryptophan in all of their libraries. In another study of random sequence proteins of a similar length to those described here, Tretyachenko et al. reported that 53% of random 20 AA sequences with similar amino acid usage to modern proteins could be expressed in *E. coli* (8/15).[56] Finally, a 20 AA random library of proteins 95 amino acids in length created by Urabe and colleagues contained only 20% solubly expressed variants (as detected by western blot).[57] The observation that our libraries become progressively more amenable to heterologous expression in *E. coli* as they become more similar to the modern proteome could indicate that the modern expression system has been optimised for modern proteins. We acknowledge the limitations of using an extant organism with its translation, chaperone and degradation systems to express "primordial" proteins, but there is currently no satisfactory alternative.

The *E. coli* expression system is simply a technically feasible tool for protein synthesis that allows our different libraries to be compared with one another, and with previous similar reduced-alphabet studies.

The two most ancient alphabets in our study displayed a greater propensity for solubility among their expressed variants than the more modern libraries. In spite of low expression levels in *E. coli*, the majority of variants that did express from the 5 AA and 9 AA libraries showed high solubility (5 AA: 6/7 expressed were $\geq$80% soluble; 8/11 for 9 AA; Table 1, Figure 3 B). In contrast, the 20 AA library was dominated by proteins with low or no solubility. Eleven out of 16 expressible variants had $\leq$20% soluble protein (Table 1, Figure 3 B). The 16 AA library appeared to be the "sweet-spot" for expression and solubility: lacking the bulky, aromatic residues, tryptophan, tyrosine and phenylalanine, it displayed comparable expression levels to the 20 AA library but a higher propensity for soluble protein. Thirteen of the 15 expressible variants had 21–100% soluble protein, corresponding to 56% of all proteins tested from that library. An analysis of the amino acid distribution of different subsets of variants from each library (i.e., expressed vs. no expression; high/partial solubility vs. low/no solubility) uncovered several correlations between the biophysical properties of the proteins and their amino acid distribution, thereby indicating that selection is occurring for solubly expressing proteins (Table S6). For instance, 20 AA variants with low or no solubility contained higher proportions of bulky, aromatic tyrosine and tryptophan residues, which are likely to lead to aggregation. These initial solubility assay results—notably the trend of early alphabets displaying high solubility when expression was observed—support the idea that our four libraries are amenable to further experimentation, such as future functional studies.

Based on the variants tested here, the 16 AA library yielded the proteins with the highest secondary and tertiary structural content. We performed CD analysis on four solubly expressing variants from each of the reduced libraries, and the single highly expressed 20-5 variant from the extant library (Figures 4 and S3). Under buffer-only conditions, all the spectra had similar profiles, with a minimum at $\approx$198 nm, which is typical of a random coil. This indicates that the variants do not contain substantial α-helical or β-strand secondary structural components. A quantitative measurement of helicity (a function of absorbance at 222 nm), however, gave helicity values ranging from 12–27% (Table S4), indicating a slight α-helical character. The addition of organic solvent TFE increased helicity in the majority of variants, with variant 16-23 exhibiting 99% helicity in the presence of 25% TFE (Table S5). The CD spectrum of 16-23 still differs from the reference graph for an all-α-helical protein in Figure 4 B, in that the magnitude of the minimum at 208 nm suggests additional random-coil content.[58] This variant displayed the greatest TFE-induced shift in helicity from 27 to 99% (72% change). Although variants from all libraries demonstrated TFE-inducible increases in helicity, the average shift for each library (5 AA: 20%; 9 AA: 33%; 16 AA: 42%; 20-5 variant: 45%) followed a trend of a greater shift being correlated with greater alphabet complexity.

The 16 AA library was also the only alphabet to yield a variant that exhibited potential tertiary structure—hydrophobic packing—as detected by ANS fluorescence (variant 16-1; Figure 5). The lower propensity for secondary structure in the 5 AA library can be explained by the abundance of the rigid proline and flexible glycine residues, which are known to disfavour secondary structure.[59] These two amino acids are about four to five times more abundant in the 5 AA library than in the extant proteome (Figure S1). β-Branched amino acids like valine (present in 5 AA library) might destabilise α-helices.[60] Therefore, it appears that the composition of the 5 AA library is less likely to promote canonical secondary structure than are the later alphabets. The libraries representing intermediate codes (9 AA and 16 AA) have a greater diversity of residues than the 5 AA library; this has the dual effect of reducing the abundance of secondary-structure-disrupting side chains and also introducing amino acids that can have stabilising effects. It has been demonstrated that the dipole moment of α-helices can be stabilised by basic residues (absent from 5 AA library) at their C termini and acidic residues at the N termini.[61] The greater diversity of chemistries introduced in the 9 AA (Glu added) and 16 AA (Arg, His, Lys added) libraries could provide such an effect. The further increase in chemical complexity in the 16 AA library could explain why 16-1 was the only variant to display hydrophobic packing. Although it is difficult to make inferences about tertiary structure from the primary sequence, the increase in polar residues in the 16 AA library (Arg, Gln, Asn, His, Cys are added to the 9 AA alphabet) likely enables stronger interactions at the protein/solvent interface. This effect, in turn, could facilitate the formation of a hydrophobic core, even without the bulky aromatic residues tyrosine, phenylalanine or tryptophan.

The biophysical properties of randomly chosen variants from the four libraries showed a trend that followed the increasing chemical diversity of side chains. Variants from the 5 AA library exhibited the lowest propensity for expression in *E. coli*, and yet 85% of expressed variants were highly soluble. The 5 AA variants had the least TFE-induced increase in secondary structure. Variants from the 9 AA library were more likely to be heterologously expressed (46% detected by western blot) and displayed similar solubility, with a greater average response to TFE in CD analyses. The 16 AA variants continued this trend of an increased propensity for expression and inducible secondary structure, including the variants with the greatest helical content (16-23), and the single variant with demonstrable hydrophobic packing (16-1). Several other research groups have taken diverse approaches to investigating intermediate alphabets such as our 9 AA and 16 AA: several computational studies have suggested that about ten different amino acids would be required for an alphabet to ensure a foldable proteins.[62,63] In an experimental approach, a 108-residue folded four-helix bundle protein made from just seven different amino acids was designed and structurally solved.[64] In a further remarkable example, a chorismate mutase was successively modified to contain only nine different amino acids (chemically redundant, not primordial) and was still found to function inside a cell.[21,25,65] In addition to showing the highest propensity for

soluble heterologous expression, inducible secondary structure and tertiary packing, the 16 AA library also contains many side chains important for catalysis[66] that are lacking from the 5 AA and 9 AA libraries. These results suggest that an alphabet of the 16 earliest amino acids could probably support many extant protein functions. The incorporation of the four likely latest amino acids (Met, Tyr, Phe and Trp)[19] into the genetic code appears to represent a selective trade-off. Although the fraction of expressed variants from the 20 AA library was comparable to that from the 16 AA library, the majority of these displayed no or low solubility. This poor solubility was likely to be due to the 20 AA library's inclusion of the three aromatic residues (Tyr, Phe, Trp). Indeed, tyrosine and tryptophan were enriched in the variants with low to no solubility, compared to other 20 AA library sequences. It is interesting to note that, in the soluble 20 AA variants, tyrosine was depleted whereas phenylalanine, which differs by just the hydroxy group, was not. In contrast, the different constitutional isomers isoleucine and leucine showed comparable abundances. These results emphasise the sensitivity of protein properties to the chemistries of amino acids and that, despite apparent similarities, each amino acid added to the code has been subject to selective pressure.[13] Aromatic residues, which were only part of our 20 AA library, expanded the chemical toolbox of the code to increase functionality, such as coordinating aromatic ligands or participating in catalysis,[66] and also enhance protein folding and stability by their potential to increase core hydrophobicity. For the genetic code to have evolved to the current 20-amino-acid alphabet,[67] these benefits must have outweighed the accompanying decrease in solubility for partially folded or unfolded regions. However, it is important to appreciate the evolution of the genetic code as a gradual process, during which newly encoded amino acids became integrated into pre-existing functional proteins under selective pressure. This process is highly unlikely to continue to yield new functional proteins spontaneously from naïve random sequences, like in the earliest days of life. This caveat does not preclude, however, the artificial selection of functional proteins from fully randomised polypeptide libraries, such as a de novo ATP-binding protein.[30,68]

Although only a few of our protein variants exhibited clearly defined secondary structure when analysed by CD, it is reasonable to postulate that these proteins could behave like intrinsically disordered proteins (IDPs). IDPs typically do not have a defined tertiary structure and regularly contain mostly random coil. Members of this diverse family of proteins display dynamic conformations and can form stable structures upon a change in chemical or physical environment,[69] such as an interaction with a ligand or a solvent like TFE. IDPs exhibit strong CD minima at $\approx$200 nm[70] and do not exhibit significant ANS signal,[69] as was observed for most of our library variants. Building on the TFE-inducible α-helical character we observed for 12 of the 13 tested variants, Lu et al. used CD to detect a population-level change in the secondary structure of their 15-amino-acid alphabet library in the presence of ATP and NADP.[37] These data support the idea that a protein does not have to display discrete secondary or compact tertiary structure to function. Furthermore, in earlier work, we generated an artificial RNA ligase enzyme[29,52,71] that had a CD spectrum similar to all but one variant reported here and, yet, this protein had a stably folded three-dimensional structure[53] and a melting temperature of 72 °C. Importantly, the example of this unnatural enzyme highlights that, in contrast to common belief, the absence of canonical secondary structure elements, namely α-helices or β-strands, or the lack of a large compact hydrophobic core does not preclude a stable three-dimensional structure and enzymatic activity. Indeed there are also examples of peptides that are not globular, yet possess activity.[72]

IDPs have been suggested to evolve more rapidly than globular proteins due to the lack of strict structural constraints.[69] Structural flexibility predisposes IDPs to multifunctionality;[69] a trait that both increases the functional adaptability of a cell,[73] and is thought to be essential for the evolution of catalytic function.[32] Furthermore, the earliest cells would be unlikely to be able to maintain a large genome due to the poorer fidelity and processivity of early polymerases; therefore, it would have been more economical to express proteins able to perform multiple cellular tasks.[31,32,73,74] Although this still needs to be investigated in future studies, the proteins described in this work might possess some of these characteristics, as demonstrated by their biophysical properties. This characteristic of IDPs can serve as a reminder that the earliest proteins might have had biological activity without adhering to the still-common notion that proteins must be the globular, well-folded macromolecules.

Conclusions

This work describes the analysis of the biophysical properties of several dozen protein variants from four different amino acid alphabet libraries and represents a foundational study for further experimental exploration of the genetic code's history. We have shown that reduced-alphabet proteins are surprisingly soluble compared to random extant-alphabet proteins, and that they exhibit secondary structure under certain conditions. These variants represent only a tiny fraction of the $>10^{12}$ variants from the libraries synthesised; future work will probe the functional capacity of reduced-alphabet proteins, employing in vitro selection techniques such as mRNA display.[75,76] mRNA display can not only probe the libraries' diversities fully, but will also allow us to manipulate the selection conditions with great precision to reflect predicted early life. It will be valuable to compare the characteristics of functionally selected library variants to the variants from unselected libraries described here. The biophysical properties found in this study by testing a small number of random variants from the reduced amino acid libraries demonstrated that these primordial alphabets could have played a key role in early life, whether as autonomous proteins or as accessory proteins in an RNA-dominated world.[77] We believe that our biophysical characterisation of proteins representing early points in the genetic code's evolution lends support to the validity of hypothesised reduced alphabets, and is an important first step in investigating the early evolution of today's protein-dominated world.

Experimental Section

Library construction, sequencing and cloning: The libraries were constructed from cassettes of 20 randomised codons (XXX) that represented a specific mixture of codons according to each of the four alphabets (Figure 2 and Table S1). The cassettes were ligated together to form the final libraries with a randomised region of 83 amino acids flanked by the following constant regions: 5′-TTACA ATTAC AGCAA TGGGG $[(XXX)_{20}$ $GGG]_4$ ATTCC TCACC ATCAC CA-3′. The 5′ region contained a tobacco mosaic virus translational enhancer; the 3′ region contained a His_6 purification tag. The randomised region contained three invariant glycine residues (GGG codon) as a result of cloning scars between cassettes, but glycine was part of all alphabets.

The trimer phosphoramidites used for cassette chemical synthesis were a product of Glen Research (VA), and 60-mer oligonucleotides (corresponding to 60 nucleotides from 20 codons) were assembled by the Keck Biotechnology Resources Center (Yale University, New Haven, CT) flanked by constant regions 5′-TTACA ATTAC AGCAA TGGGG-3′ and 5′-GGCAT TCCTC ACCAT CACCA-3′ at the 5′- and 3′-ends, respectively. These constant regions contained the recognition sites for the type IIS restriction enzymes BsrDI and BsmI. Accordingly, each variant's coding region contained a constant N-terminal Met-Gly sequence, and a Gly-Ile-Pro sequence C-terminal to the randomised region of the library. The trimer codons used in the synthesis and the desired ratios of residues in each library's random region are listed in Table S1. All primers and library cassette oligonucleotides were purified by urea-PAGE gel electrophoresis. The gel bands were excised and extracted overnight into TE buffer (10 mM Tris, 1 mM EDTA, pH 7.4) by using the crush-and-soak method. From these purified single-stranded DNA cassettes, which encoded 20 random codon positions, an aliquot corresponding to at least 10^{12} unique sequences was amplified by PCR. Forward primer 5′-GGACA ATTAC TATTT AAATT ACAGC AATGG G-3′, reverse primer 5′-TGGTG ATGGT GATGG TGAGG AATGC C-3′ and Phusion polymerase (New England Biolabs) were used for seven PCR cycles of 95 °C for 30 s, 60 °C for 20 s and 72 °C for 15 s. Primer concentrations were 0.5 μM, dNTPs were 200 μM, and template concentration was 5 nM. PCR products were purified by phenol/chloroform extraction, concentrated with butanol, and precipitated with ethanol. PCR cleanup kits (Qiagen) were used to de-salt samples prior to restriction digest. Samples were quantified by determining the absorbance at 260 nm, and their molecular weight was verified by agarose gel electrophoresis. The purified PCR product was then divided into two equal pools, and each pool was digested with either BsrDI or BsmI for 4 h at 65 °C. Following digestion and ethanol precipitation, products were separated on a 3% agarose gel (120 V, 60 min) to separate the desired digested library fragments from the terminal constant primer-binding region. The appropriate band was excised, purified (Qiagen gel extraction kit) and subjected to ligation with T4 DNA ligase in the commercial reaction buffer (New England Biolabs). The ligation was carried out for 30 min at room temperature. Ligation concentrations were 0.2 μM DNA, 1 mM ATP and 20 $U\,\mu L^{-1}$ T4 DNA ligase. Following ligation, the products were concentrated in a vacuum concentrator at the medium setting and subsequently resolved on a 3% agarose gel, excised and purified (Qiagen gel extraction kit). The ligated product, consisting of two joined cassettes of 20 codons each, was then subjected to a second round of PCR amplification, digestion and ligation to build up a final library containing 80 trinucleotides (four cassettes of 20 codons each). Identical PCR, purification, digestion and ligation protocols were followed for this second round. Each completed library represented greater than 2×10^{12} unique sequences, as determined by spectrophotometric quantification at 260 nm. To confirm the sequences, the final DNA libraries were digested with BsrDI and BsmI and ligated into a pER13 plasmid that had been modified to introduce these restriction sites into the multiple cloning site.[41,42] Where necessary, further sequencing was performed on variants cloned into pET28a by using the NcoI and BamHI restriction sites. Individual colonies were picked, and plasmids were subsequently isolated and sequenced (Beckman Coulter Genomics, Danvers, MA).

Assaying solubility by *E. coli* expression and western blot: Sequence-confirmed mutants were expressed from pET28a that contained a T7 promotor under control of the *lac* operon. Following transformation into *E. coli* BL21(DE3) cells and plating onto lysogeny broth agar with kanamycin, single colonies of each sequenced variant were selected and grown overnight in LB containing kanamycin (36 $\mu g\,mL^{-1}$), and then used to inoculate LB-kanamycin cultures (10 mL) and grown at 37 °C. Cultures were induced with isopropyl-β-D-thiogalactopyranoside (IPTG; 1 mM) at an OD_{600} of 0.6. Expression continued for 3 h at 37 °C. Cells were harvested by centrifugation at 3000 *g* and resuspended in BugBuster HT (Novagen) at 5 $mL\,g^{-1}$ wet cell pellet with SIGMA*FAST* protease inhibitor (Sigma–Aldrich, 1 tablet per 500 mL). Cell cultures were lysed for 30 min at room temperature with shaking, then the soluble fraction was separated by centrifugation (21 000 *g*). Equal volumes of soluble fraction and solubilised pellet (solubilised to original volume in 1× PBS, 250 μM NaCl, 8 M urea: representing the insoluble proteins) were analysed by SDS-PAGE (4–12% Bis·Tris, NuPage gels, Life Technologies) in MES buffer (Life Technologies), thereby enabling the direct comparison of protein amounts for both soluble fraction and pellet on the electrophoresis gel.

For increased detection sensitivity, all electrophoresis gels were detected by western blotting with an antibody specific to the C-terminal His_6 tag present in all proteins. Following SDS-PAGE, the separated proteins were transferred onto PVDF membranes (Roche) in transfer buffer (25 mM Tris, 192 mM glycine, 20% *v/v* methanol, 0.005% *w/v* SDS) with a Pierce Power System, according to the manufacturer's instructions. The membranes were washed in blotting buffer (25 mM Tris, pH 7.4, 150 mM NaCl, 0.1% *v/v* Tween-20) and blocked in blotting buffer with 5% *w/v* milk powder. Proteins were detected by using a horseradish peroxidase (HRP)-conjugated mouse monoclonal anti-His_6 tag antibody (Thermoscientific; 1:500 dilution in blotting buffer with 2% *w/v* milk powder) and the detection reagents from the Pierce fast western blot kit, according to the manufacturer's instructions. Membranes were analysed with a LICOR Odyssey Fc imaging system by using exposure times of 2 or 10 min. ImageJ software[43] was used to quantify band intensities in cases where variants showed partial solubility.

Large-scale expression and purification of selected soluble variants: Protein variants that showed soluble over-expression in the small-scale solubility assays (above) were also expressed in 1 L cultures to obtain sufficient amounts of protein for biophysical characterisation. The conditions for expression, harvest and lysis were otherwise identical to those for the small-scale screen. The soluble fractions were diluted 1:1 into purification buffer (10 mM sodium phosphate, 150 mM NaCl, 5 mM imidazole, 0.5 mM β-mercaptoethanol, pH 7.4) and mixed with pre-washed and equilibrated Ni-NTA resin (300 μL, Life Technologies). Following 30 min of rotated incubation with the resin at 4 °C, the supernatant and resin were transferred to a 10 mL chromatography column (Bio-Rad). Column washes were performed at 4 °C with 2×3 column volumes of purification buffer. Elution was conducted in a step-wise fashion with CD buffer (0.5 mL, 10 mM sodium phosphate, 150 mM NaF, pH 7.4)

increasing from 20 to 50, 100, 150 and 250 mM imidazole. Fractions were assessed for the presence of protein by SDS-PAGE. Fractions containing protein were pooled and dialysed twice for 12 h against CD buffer (2 L) to remove imidazole. Following dialysis, protein samples were spin concentrated (Amicon Ultra, 3000 MW), and the concentration was determined by bicinchoninic acid (BCA) assay.[44] Protein samples were stored in CD buffer at 4 °C.

Characterisation of the biophysical properties of protein variants

Determination of secondary structure content by circular dichroism: CD spectra were taken on a JASCO J-185 CD spectropolarimeter. Proteins were scanned from 260 to 190 nm at a rate of 50 nm s^{-1} for ten accumulations and four scans. All protein samples (10 μM) were read with a 1 mm path-length cuvette at 23 °C. All spectra were baseline corrected against the CD spectrum of the buffer only. TFE (Sigma–Aldrich) was added to samples to final concentrations of 25 and 50%. In the presence of TFE, samples were allowed to equilibrate at 23 °C for 15 min.

The measured ellipticities ($[\theta]_{obs}$ [mdeg]) were converted to mean residue ellipticities ($[\theta]$ [deg cm^2 $dmol^{-1}$]) according to Equation (1):

$$[\theta] = [\theta]_{obs} \times \left(\frac{MRW}{10\,lc}\right) \quad (1)$$

MRW is the mean residue molecular weight of the peptide (calculated by dividing the molecular weight of the peptide by the number of residues), l is the optical path length of the cell [cm], and c is the concentration of the peptide [mg mL^{-1}]. The degree of helical structure (% α-helix) was calculated according to Equation (2):

$$\%\ \alpha\text{-helix} = \frac{[\theta]_{222} \times 100}{-40\,000(1 - 2.5/n)} \quad (2)$$

Here $[\theta]_{222}$ is the mean residue ellipticity at 222 nm, and n the number of the residues in the peptide.[45]

Assessment of protein folding by ANS fluorescence: ANS fluorescence was measured in a Molecular Devices SpectraMax M5 spectrofluorometer. With the exception of the size standard bovine serum albumin (BSA), samples were mixed in 200 μL volumes containing 10 μM protein and 500 μM ANS dissolved in CD buffer. Due to its greater mass and presumed greater surface area, BSA 1.5 μM, was mixed in 200 μL with 500 μM ANS. Baseline spectra were recorded with CD buffer only. After incubation at room temperature for 30 min, samples were excited at 385 nm, and fluorescence was recorded from 400 to 600 nm. Six readings were taken per sample with the scan speed set to normal.

Determining oligomeric state by size-exclusion chromatography: Gel filtration chromatography was performed by using a Superdex 75 resin (GE Healthcare) in a 10 mm × 300 mm column (Tricorn) on the ÄKTA FPLC system (GE Healthcare). Protein samples in CD buffer were injected by using a 10 μL loop with an isocratic elution at a flow rate of 1 mL min^{-1}.

Acknowledgements

We thank Dr. Maureen Quin for assistance with the western blots, and Prof. Romas Kazlauskas and Fredarla Miller for their comments on the manuscript. This work was funded in part by grants from the US National Aeronautics and Space Administration (NASA) Agreement (NNX14AK29G), the Simons Foundation (340762), the Minnesota Medical Foundation (4036–9663-10), the University of Minnesota Biocatalysis Initiative, and the Office of the VP of Research at the University of Minnesota (Grant-in-Aid).

Conflict of Interest

The authors declare no conflict of interest.

Keywords: genetic code · origin of proteins · primordial peptides · protein libraries

[1] D. Voet, J. G. Voet, *Biochemistry*, 3rd ed. **2004**, Wiley, Hoboken, pp. 80–126.
[2] J. E. Cone, R. M. Del Rio, J. O. E. N. Davis, T. C. Stadtman, *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 2659–2663.
[3] M. A. Gaston, L. Zhang, K. B. Green-Church, J. A. Krzycki, *Nature* **2011**, *471*, 647–650.
[4] F. H. C. Crick, *J. Mol. Biol.* **1968**, *38*, 367–379.
[5] E. V. Koonin, A. S. Novozhilov, *IUBMB Life* **2009**, *61*, 99–111.
[6] J. Wong, S.-K. Ng, W.-K. Mat, T. Hu, H. Xue, *Life* **2016**, *6*, 12.
[7] M. Di Giulio, *J. Mol. Evol.* **2016**, *83*, 93–96.
[8] M. Eigen, P. Schuster, *Naturwissenschaften* **1978**, *65*, 341–369.
[9] S. L. Miller, *Science* **1953**, *117*, 528–529.
[10] A. P. Johnson, H. J. Cleaves, J. P. Dworkin, D. P. Glavin, A. Lazcano, J. L. Bada, *Science* **2008**, *322*, 404–404.
[11] M. Eigen, P. Schuster, *Naturwissenschaften* **1977**, *64*, 541–565.
[12] S. Osawa, T. H. Jukes, K. Watanabe, A. Muto, *Microbiol. Rev.* **1992**, *56*, 229–264.
[13] G. K. Philip, S. J. Freeland, *Astrobiology* **2011**, *11*, 235–240.
[14] T. A. Ronneberg, L. F. Landweber, S. J. Freeland, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 13690–13695.
[15] S. A. Sandford, J. Aléon, C. M. O'D. Alexander, T. Araki, S. Bajt, G. A. Baratta, J. Borg, J. P. Bradley, D. E. Brownlee, J. R. Brucato, et al., *Science* **2006**, *314*, 1720–1724.
[16] C. D. K. Herd, A. Blinova, D. N. Simkus, Y. Huang, R. Tarozo, C. M. O'D. Alexander, F. Gyngard, L. R. Nittler, G. D. Cody, M. L. Fogel, et al., *Science* **2011**, *332*, 1304–1307.
[17] H. J. Cleaves II, *J. Theor. Biol.* **2010**, *263*, 490–498.
[18] H. J. Cleaves, S. L. Miller, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 7260–7263.
[19] E. N. Trifonov, *J. Biomol. Struct. Dyn.* **2004**, *22*, 1–11.
[20] P. G. Higgs, R. E. Pudritz, *Astrobiology* **2009**, *9*, 483–490.
[21] S. V. Taylor, K. U. Walter, P. Kast, D. Hilvert, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10596–10601.
[22] S. Akanuma, T. Kigawa, S. Yokoyama, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 13549–13553.
[23] R. Shibue, T. Sasamoto, M. Shimada, B. Zhang, A. Yamagishi, S. Akanuma, *Sci. Rep.* **2018**, *8*, 1227.
[24] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *Science* **1993**, *262*, 1680–1685.
[25] K. U. Walter, K. Vamvaca, D. Hilvert, *J. Biol. Chem.* **2005**, *280*, 37742–37746.
[26] M. T. Reetz, S. Wu, *Chem. Commun.* **2008**, 5499–5501.
[27] Z. Sun, R. Lonsdale, X. D. Kong, J. H. Xu, J. Zhou, M. T. Reetz, *Angew. Chem. Int. Ed.* **2015**, *54*, 12410–12415; *Angew. Chem.* **2015**, *127*, 12587–12592.
[28] E. L. Peterson, J. Kondev, J. A. Theriot, R. Phillips, *Bioinformatics* **2009**, *25*, 1356–1362.
[29] B. Seelig, J. W. Szostak, *Nature* **2007**, *448*, 828–831.
[30] A. D. Keefe, J. W. Szostak, *Nature* **2001**, *410*, 715–718.
[31] A. Szilágyi, Á. Kun, E. Szathmáry, *Biol. Direct* **2012**, *7*, 38.
[32] R. Jensen, *Annu. Rev. Microbiol.* **1976**, *30*, 409–425.
[33] M. V. Golynskiy, B. Seelig, *Trends Biotechnol.* **2010**, *28*, 340–345.
[34] C. Neylon, *Nucleic Acids Res.* **2004**, *32*, 1448–1459.
[35] A. L. Kayushin, M. D. Korosteleva, A. I. Miroshnikov, W. Kosch, D. Zubov, N. Piel, *Nucleic Acids Res.* **1996**, *24*, 3748–3755.

[36] S. Kang, B. Chen, T. Tian, X. Jia, X. Chu, R. Liu, P. Dong, Q. Yang, H. Zhang, *Biochem. Biophys. Res. Commun.* **2015**, *466*, 400 – 405.
[37] M.-F. Lu, Y. Xie, Y.-J. Zhang, X.-Y. Xing, *Protein Pept. Lett.* **2015**, *22*, 579 – 585.
[38] N. Doi, K. Kakukawa, Y. Oishi, H. Yanagawa, *Protein Eng. Des. Sel.* **2005**, *18*, 279 – 284.
[39] J. Tanaka, N. Doi, H. Takashima, H. Yanagawa, *Protein Sci.* **2010**, *19*, 786 – 795.
[40] S. Kumachi, Y. Husimi, N. Nemoto, *ACS Omega* **2016**, *1*, 52 – 57.
[41] T. Seitz, R. Thoma, G. A. Schoch, M. Stihle, J. Benz, B. D'Arcy, A. Wiget, A. Ruf, M. Hennig, R. Sterner, *J. Mol. Biol.* **2010**, *403*, 562 – 577.
[42] M. V. Golynskiy, J. C. Haugner, B. Seelig, *ChemBioChem* **2013**, *14*, 1553 – 1563.
[43] C. A. Schneider, W. S. Rasband, K. W. Eliceiri, *Nat. Methods* **2012**, *9*, 671 – 675.
[44] P. K. Smith, R. I. Khrohn, G. T. Hermanson, A. K. Mallia, F. H. Gatner, M. D. Provenzano, E. K. Fujimoto, N. M. Goeke, B. J. Olson, D. C. Klenk, *Anal. Biochem.* **1985**, *150*, 76 – 85.
[45] Y. H. Chen, J. T. Yang, K. H. Chau, *Biochemistry* **1974**, *13*, 3350 – 3359.
[46] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, D. Baker, *Nat. Struct. Biol.* **1997**, *4*, 805 – 809.
[47] G. Cho, A. D. Keefe, R. Liu, D. S. Wilson, J. W. Szostak, *J. Mol. Biol.* **2000**, *297*, 309 – 319.
[48] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch in *Proteomics Protocols Handbook* (Ed.: J. M. Walker), Humana, Totowa, **2005**, pp. 571 – 607.
[49] F. D. Sönnichsen, J. E. Van Eyk, R. S. Hodges, B. D. Sykes, *Biochemistry* **1992**, *31*, 8790 – 8798.
[50] K. Gast, D. Zirwer, M. Müller-Frohne, G. Damaschun, *Protein Sci.* **1999**, *8*, 625 – 634.
[51] H. Reiersen, A. R. Rees, *Protein Eng.* **2000**, *13*, 739 – 743.
[52] A. Morelli, J. Haugner, B. Seelig, *PLoS One* **2014**, *9*, e112028.
[53] F.-A. Chao, A. Morelli, J. C. Haugner, L. Churchfield, L. N. Hagmann, L. Shi, L. R. Masterson, R. Sarangi, G. Veglia, B. Seelig, *Nat. Chem. Biol.* **2013**, *9*, 81 – 83.
[54] Y. Shimizu, A. Inoue, Y. Tomari, T. Suzuki, T. Yokogawa, K. Nishikawa, T. Ueda, *Nat. Biotechnol.* **2001**, *19*, 751 – 755.
[55] K. Amikura, Y. Sakai, S. Asami, D. Kiga, *ACS Synth. Biol.* **2014**, *3*, 140 – 144.
[56] V. Tretyachenko, J. Vymětal, L. Bednárová, V. Kopecký, Jr., K. Hofbauerová, H. Jindrová, M. Hubálek, R. Souček, J. Konvalinka, J. Vondrášek, K. Hlouchová, *Sci. Rep.* **2017**, *7*, 15449.
[57] I. D. Prijambada, T. Yomo, F. Tanaka, T. Kawama, K. Yamamoto, A. Hasegawa, Y. Shima, S. Negoro, I. Urabe, *FEBS Lett.* **1996**, *382*, 21 – 25.
[58] N. J. Greenfield, *Nat. Protoc.* **2006**, *1*, 2876 – 2890.
[59] C. N. Pace, J. M. Scholtz, *Biophys. J.* **1998**, *75*, 422 – 427.
[60] A. M. Facchiano, G. Colonna, R. Ragone, *Protein Eng.* **1998**, *11*, 753 – 760.
[61] J. Richardson, D. Richardson, *Science* **1988**, *240*, 1648 – 1652.
[62] K. Fan, W. Wang, *J. Mol. Biol.* **2003**, *328*, 921 – 926.
[63] L. R. Murphy, A. Wallqvist, R. M. Levy, *Protein Eng.* **2000**, *13*, 149 – 152.
[64] C. E. Schafmeister, S. L. LaPorte, L. J. W. Miercke, R. M. Stroud, *Nat. Struct. Biol.* **1997**, *4*, 1039 – 1046.
[65] M. M. Müller, J. R. Allison, N. Hongdilokkul, L. Gaillon, P. Kast, W. F. van Gunsteren, P. Marlière, D. Hilvert, *PLoS Genet.* **2013**, *9*, e1003187.
[66] G. L. Holliday, J. B. O. Mitchell, J. M. Thornton, *J. Mol. Biol.* **2009**, *390*, 560 – 577.
[67] M. Ilardo, M. Meringer, S. Freeland, B. Rasulev, H. J. Cleaves, H. J. Cleaves II, *Sci. Rep.* **2015**, *5*, 9414.
[68] P. Lo Surdo, M. A. Walsh, M. Sollazzo, *Nat. Struct. Mol. Biol.* **2004**, *11*, 382 – 383.
[69] J. D. Forman-Kay, T. Mittag, *Structure* **2013**, *21*, 1492 – 1499.
[70] L. B. Chemes, L. G. Alonso, M. G. Noval, G. de Prat-Gay, *Methods Mol. Biol.* **2012**, *895*, 387 – 404.
[71] B. Seelig, *Nat. Protoc.* **2011**, *6*, 540 – 552.
[72] C. M. Rufo, Y. S. Moroz, O. V. Moroz, J. Stöhr, T. A. Smith, X. Hu, W. F. DeGrado, I. V. Korendovych, *Nat. Chem.* **2014**, *6*, 303 – 309.
[73] W. M. Patrick, E. M. Quandt, D. B. Swartzlander, I. Matsumura, *Mol. Biol. Evol.* **2007**, *24*, 2716 – 2722.
[74] M. P. Ferla, J. L. Brewster, K. R. Hall, G. B. Evans, W. M. Patrick, *Mol. Microbiol.* **2017**, *105*, 508 – 524.
[75] R. W. Roberts, J. W. Szostak, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12297 – 12302.
[76] N. Nemoto, E. Miyamoto-Sato, Y. Husimi, H. Yanagawa, *FEBS Lett.* **1997**, *414*, 405 – 408.
[77] H. S. Bernhardt, *Biol. Direct* **2012**, *7*, 23.

Manuscript received: November 1, 2018
Accepted manuscript online: December 3, 2018
Version of record online: February 15, 2019