Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability

Mikhail Makarov[1], Alma C. Sanchez Rocha[1], Robin Krystufek[2,3], Ivan Cherepashuk[1], Volha Dzmitruk[4], Tatsiana Charnavets[4], Anneliese M. Faustino,[5] Michal Lebl[3], Kosuke Fujishima[6,7], Stephen D. Fried[5,8], and Klara Hlouchova*[1,3]

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

[2]*Department of Physical Chemistry, 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]*Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, 25250 Vestec, Czech Republic*

[5]*Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21212, United States*

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

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

[8]*Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21212, United States*

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

Abstract

Whereas modern proteins rely on a quasi-universal repertoire of 20 canonical amino acids (AAs), numerous lines of evidence suggest that ancient proteins relied on a limited alphabet of 10 'early' AAs, and that the 10 'late' AAs were products of biosynthetic pathways. However, many non-proteinogenic AAs were also prebiotically available, which begs two fundamental questions: Why do we have the current modern amino acid alphabet, and Would proteins be able to fold into globular structures as well if different amino acids comprised the genetic code? Here, we experimentally evaluated the solubility and secondary structure propensities of several prebiotically relevant amino acids in the context of combinatorial synthetic 25-mer peptide libraries. The most prebiotically abundant linear aliphatic and basic residues were incorporated along with (or in replacement of) other early amino acids to explore these alternative sequence spaces. We show that foldability was a critical factor in the selection of the canonical alphabet. Unbranched aliphatic and short-chain basic amino acids were purged from the proteinogenic alphabet despite their high prebiotic abundance, because they generate polypeptides that are over-solubilized and have low packing efficiency. Surprisingly, we find that the inclusion of a short-chain basic amino acid diminishes polypeptides' secondary structure potential. Our results support the view that despite lacking basic residues, the early canonical alphabet was remarkably adaptive at supporting protein folding and explain why basic residues were only incorporated at a later stage of the alphabet evolution.

Introduction

The role of peptides and proteins in life's emergence has historically been sidelined as they have been, at times, perceived as only relevant after the evolution of nucleotide-based

polymers, the genetic code, and a translation apparatus. However, the prebiotic abundance of amino acids and their ease of condensation resulting in polymers capable of creating functional hubs along with various cofactors (such as metal ions and organic compounds) have been recently re-centering peptides' role during life's early evolution (1–4). While extant proteins are built from the sequence space spanned by the twenty canonical amino acids (cAAs) and rely on the specificity of the Central Dogma, early peptides (and peptide-like polymers) likely arose from a larger pool of prebiotically plausible monomers.

The modern protein alphabet was most likely selected during the first 10–15% of Earth history (~4.5 Ga–3.9 Ga) but the factors guiding its chemical and biological evolution remain unclear (5). Some of the cAAs are thought to be "late" additions to the genetic code, enabled only (or enriched significantly) by the evolution of their biosynthetic pathways. The "early" amino acids (Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr and Val) were apparently available through prebiotic synthesis along with many other non-canonical amino acids (ncAAs) and their alternatives (such as beta- and gamma- amino acids, hydroxy acids or dicarboxylic acids). Based on two independent meta-analyses and many parallel studies, the canonical alphabet was therefore selected first from the large pool of prebiotically plausible structures and then supplemented with late amino acids through metabolism (6,7). Why certain amino acids were selected over others (and when) has been repeatedly discussed in literature over the last few decades (8–12). This question can be further subdivided into two separate but related questions, namely: (i) Why were the specific early cAAs selected from the prebiotic environment, and (ii) What factors drove the selection of the additional residues in the following era? Has protein evolution been successful as a consequence of the selected cAAs, or could similar protein space be formed with alternative alphabets (8)?

While some of the choices for the early amino acids are quite expected (especially for the C2 and C3 amino acids), others have been repeatedly considered thought-provoking. Most strikingly, the early amino acid alphabet lacked positively charged residues, even though diaminopropionic and diaminobutyric acids (C3 and C4 amino acids, DAP and DAB, respectively) have been shown to be prebiotically plausible (13,14). Neither lysine nor arginine are considered prebiotically plausible, whilst ornithine is known to cyclize, promoting peptide chain scission (8). Moreover, the canonical alphabet does not include some of the most abundant linear aliphatic amino acids such as α-amino-n-butyric acid (ABA), norvaline (Nva) and norleucine (Nle) (C3, C4, and C5, respectively) while their branched (and less abundant) alternatives (such as Val, Leu, Ile) were selected (8) despite being available at much reduced abundances. It has been hypothesized that perhaps the earlier version of the amino acid alphabet included some ncAAs that were later replaced by late cAAs (2,15). This is an intriguing hypothesis especially in light of the fact that some of these ncAAs have been reported as promiscuous substrates of some aminoacyl-tRNA synthetases (16).

This study experimentally tests the properties that would have accompanied some of the most feasible ncAA candidates from the prebiotic pool. We incorporated the selected ncAAs into combinatorial peptide libraries along with (or replacing) other early amino acids to evaluate their effect on fundamental physicochemical properties such as solubility and secondary structure formation. The outcomes of our study show that an inclusion of some of the most feasible alternatives would ***not*** have supported the twin requirements of rich secondary structure potential and solubility. Hence, our study demonstrates that biophysical constraints on foldability provided a selective pressure shaping which building blocks entered the early

alphabet and therefore explain why certain early AAs (and not others) were selected to construct proteins.

Results

Design and synthesis of peptide libraries from alternative alphabets

25-mer combinatorial peptide libraries were synthesized using solid-support chemical synthesis with isokinetic mixtures of amino acids (see Supplementary Table S1).

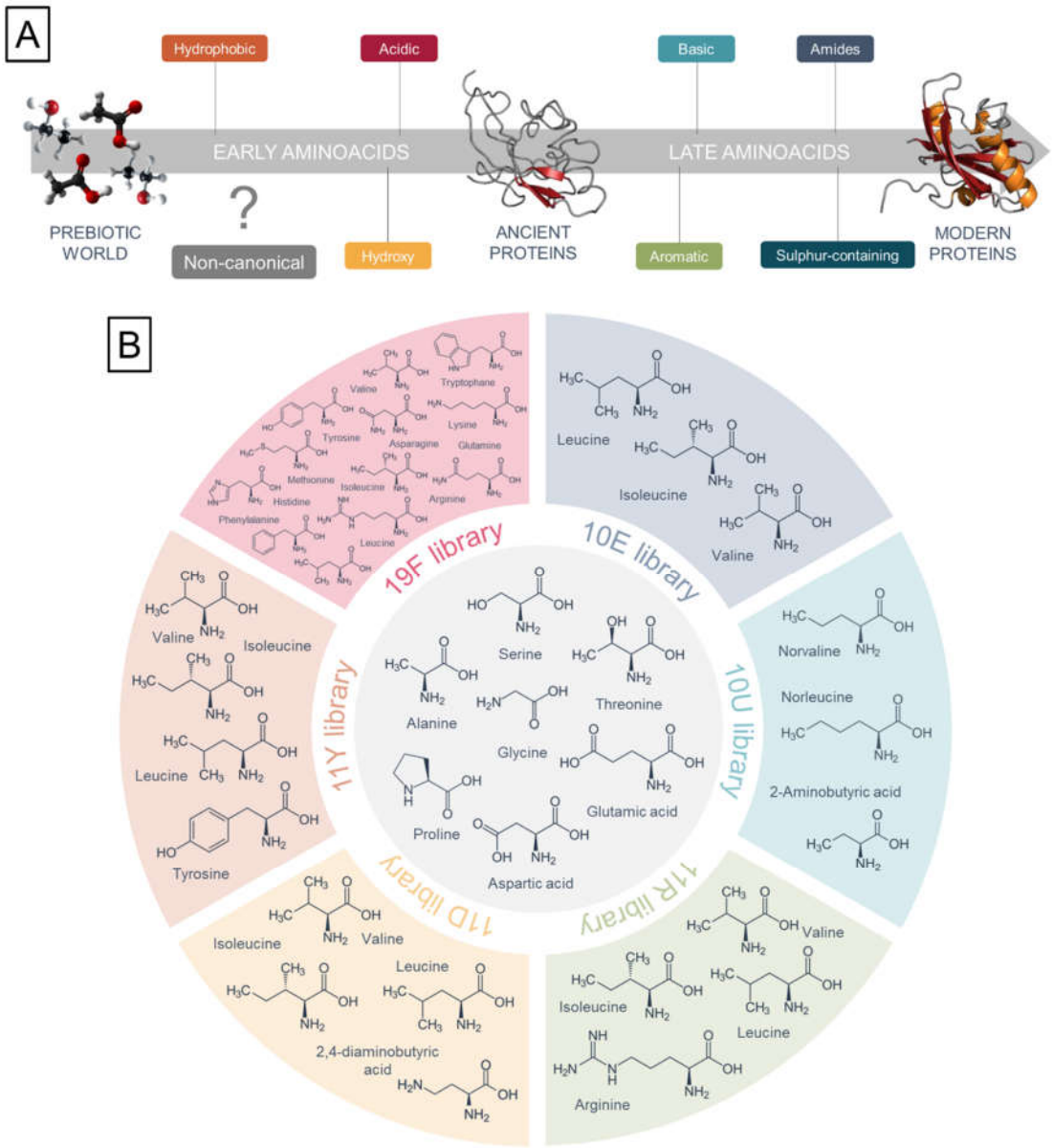


Figure 1. The assumed early and late stages of amino acid alphabet incorporation during protein evolution (A) and design of peptide libraries based on this order (B).

To mimic the sequence space available to different subsets of the amino acid alphabets, the libraries included the entire canonical alphabet without Cys (19F), its prebiotically available

subset of 10 (10E), an alternative of 10E where the branched aliphatic amino acids were substituted with their unbranched prebiotically-abundant alternatives (10U), the 10E library supplemented with: (i) Arg as a representative of a modern cationic cAA (11R); or DAB as a representative of a potentially early cationic AA (11D); or tyrosine, as a representative of an aromatic AA (11Y) (Fig. 1).

MALDI spectra of individual libraries confirmed the expected molecular weight range and distribution of the combinatorial libraries, reflecting their respective compositions (Supplementary Fig. S1). All the libraries passed the quality control and were lyophilized to establish Cl^- as the counterion for all downstream analyses.

Solubility and aggregation profiles

The ease of dissolution under different conditions could represent an important selection factor during the formation of the early protein alphabet, as prebiotically-relevant environments spanned alkaline hydrothermal vents (17) to acidic lakes (18). We measured the solubility profiles of these peptide libraries in the pH range 3–11, and also at low vs. high ionic strength, spectrophotometrically by absorption of peptide bonds at 215 nm and fluorometrically in case of library 19F due to its very poor solubility (Fig. 2). In general, ionic strength did not significantly modulate solubility, with one exception (of the 11Y library, see below).

At a fixed nominal peptide concentration (0.5 mg/ml), the 19F library is significantly less soluble than the libraries composed of fewer amino acids. This result was expected based on previous studies reporting that subsets of the canonical library that are enriched with the early amino acids are significantly more soluble than the full alphabet version (19–21). This phenomenon could be partly explained by the physicochemical nature of the full alphabet (such as the presence of aromatic amino acids) as well as the overall 'search complexity' of this larger alphabet; a significantly more rugged landscape would be expected when formed from 19 vs. 10 amino acid members causing an increased aggregation propensity (22).

The solubilities of the canonical library subsets (10E, 11R and 11Y) all increase with increasing pH (going from pH 3 to pH 11) though 11Y (which includes Tyr as a representative aromatic amino acid) is the least soluble of these three as expected. This result can be potentially explained by pointing out that all of these alphabets (as do modern proteomes (23)) tilt toward being acidic; hence, at alkaline pH the peptides will accrue larger net negative charge, and will repel one another electrostatically. This is broadly consistent with the practice of using mildly-alkaline buffers in protein refolding experiments, which tend to facilitate refolding by preventing aggregation (24).

The two libraries that include the non-canonical alternatives of aliphatic (10U) and basic (11D) amino acids are strikingly more soluble than 10E and 10R throughout the examined pH range (Fig. 2). Both of these results are unexpected for distinct reasons. Based on the physical chemistry of elementary hydrocarbons, one might expect that the 10U library would be less soluble because linear hydrocarbons are more hydrophobic than branched hydrocarbons of equal carbon-count because they create cavities with greater surface area (25). Moreover, the inclusion of a basic amino acid would raise the average isoelectric point of the library (relative to 10E) to be closer to neutral, which would decrease intermolecular repulsion, and thereby increase aggregation.

The soluble fractions of the libraries were additionally screened for the occurrence of soluble aggregates using gel filtration chromatography (Supplementary Fig. S2 and S3 in low and high ionic strength, respectively). No significant amount of such phenomenon was observed with the exception of 11Y and 10E libraries where a minor (up to ~10%) fraction of soluble aggregates was detected (Supplementary Fig. S2). Overall, the elution profiles correlated with the spectrophotometric solubility measurements in which the 11R library was the most soluble of the canonical subset libraries, and the 10U and 11D libraries were fully soluble across the entire pH range.

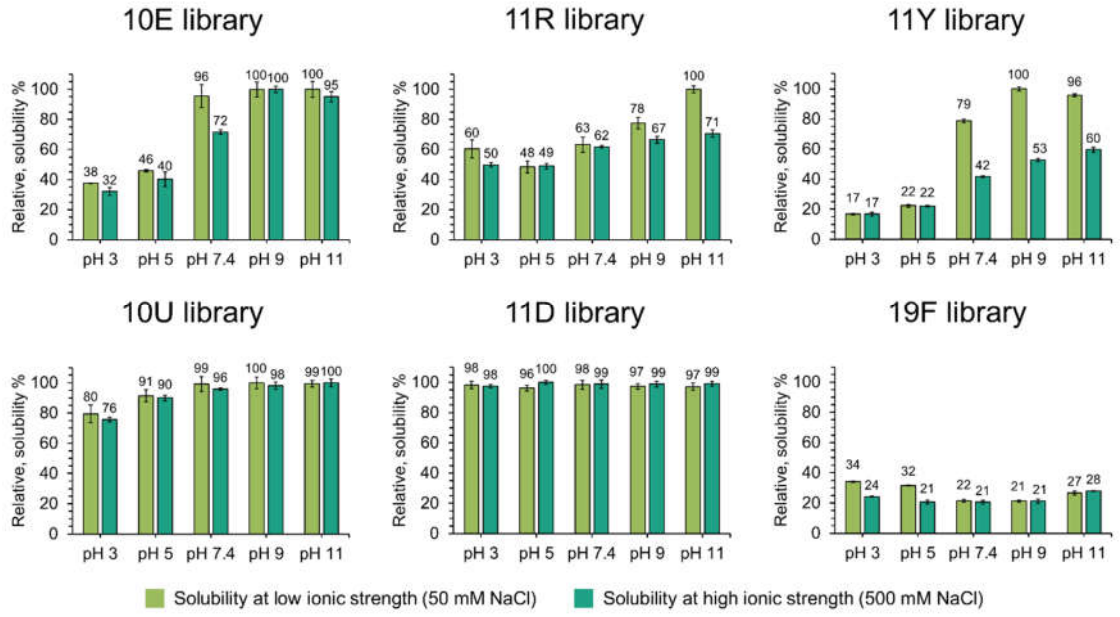


Figure 2. Solubility of 25-mer combinatorial peptide libraries at different pH and ionic strengths. Solubility was measured for 0.5 mg/ml nominal peptide library solutions in 20 mM ABP buffers (pH 3–11) at 50 and 500 mM NaCl. Solubility of 19F peptide library was measured by fluorescence spectroscopy, and solubility of the other peptide libraries was measured by UV-Vis absorption spectroscopy.

Secondary structure propensity

Natural non-folding sequences as well as random sequences with low secondary structure content have been previously reported to be highly soluble (21,26). The solubility profiles of the 11D and 10U libraries therefore suggest that the incorporation of the respective ncAAs may decrease the potential for secondary structure formation. CD spectroscopy was used to estimate average secondary structure content of the libraries over the studied pH range (Fig. 3) and upon its induction with 2,2,2-trifluoroethanol (TFE) in pH 7.4 (Fig. 4). Because no dominant secondary structure motif is expected in these highly diverse libraries and we are rather interested in average structural propensities, the secondary structure content was evaluated as the ratio of CD signal at 222 nm to the signal at 200 nm (shown by inlet graphs in Fig. 3 and 4), because both alpha helices and beta sheets absorb more at 222 nm than random coils do (27).

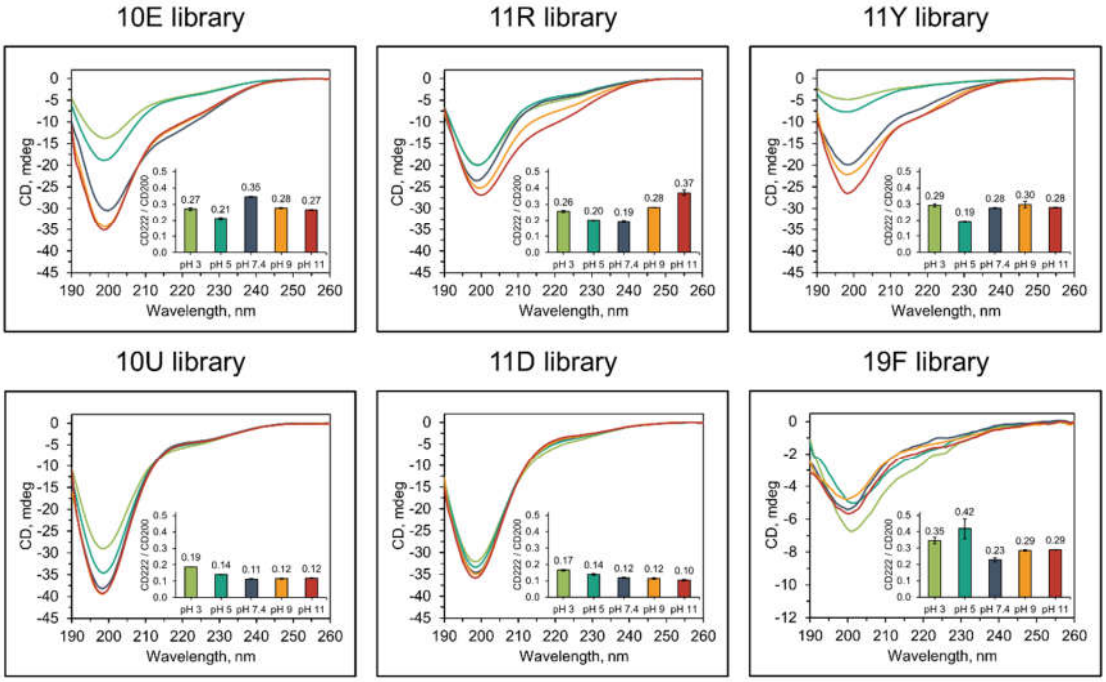


Figure 3. The effect of pH on the secondary structure of 25-mer combinatorial peptide libraries. CD spectra were measured for 0.2 mg/ml nominal peptide library solutions in a series of 10 mM ABP buffers (pH 3–11) at 25 mM NaCl. The inlet graph shows the ratios of CD signals at 222 nm to 200 nm.

The CD spectra of the 10U and 11D libraries showed significantly lower relative signals at 222 nm, indicating that these peptides possessed profoundly decreased secondary structural propensity when compared with the respective canonical libraries 10E and 11R (Fig. 3). The 10E library with branched aliphatic amino acids has a higher structural propensity than the 10U library with unbranched aliphatic alternatives across the whole pH range. The addition of DAB in the 11D library also significantly decreases secondary structure propensity while, the addition of canonical amino acids, Arg (in 11R) or Tyr (in 11Y) increases the secondary structure propensity mildly under some pH values. This trend is more profoundly visible upon TFE titration which is often used to induce helical structure (27–29). TFE addition potentiates secondary structure content robustly in the 10E library, an effect that is further amplified in alphabets that include Arg and Tyr (Fig. 4). Remarkably, the inclusion of DAB (in the 11D library) almost fully obviates the potentiating capacity of TFE, implying that DAB significantly decreases the secondary structure potential of 25-mer peptides. While the poor solubility of the 19F library is reflected in the overall low intensity of its CD spectra, its elevated structural propensity (when compared with all the other library subsets) becomes evident in the TFE titration (Fig. 4).

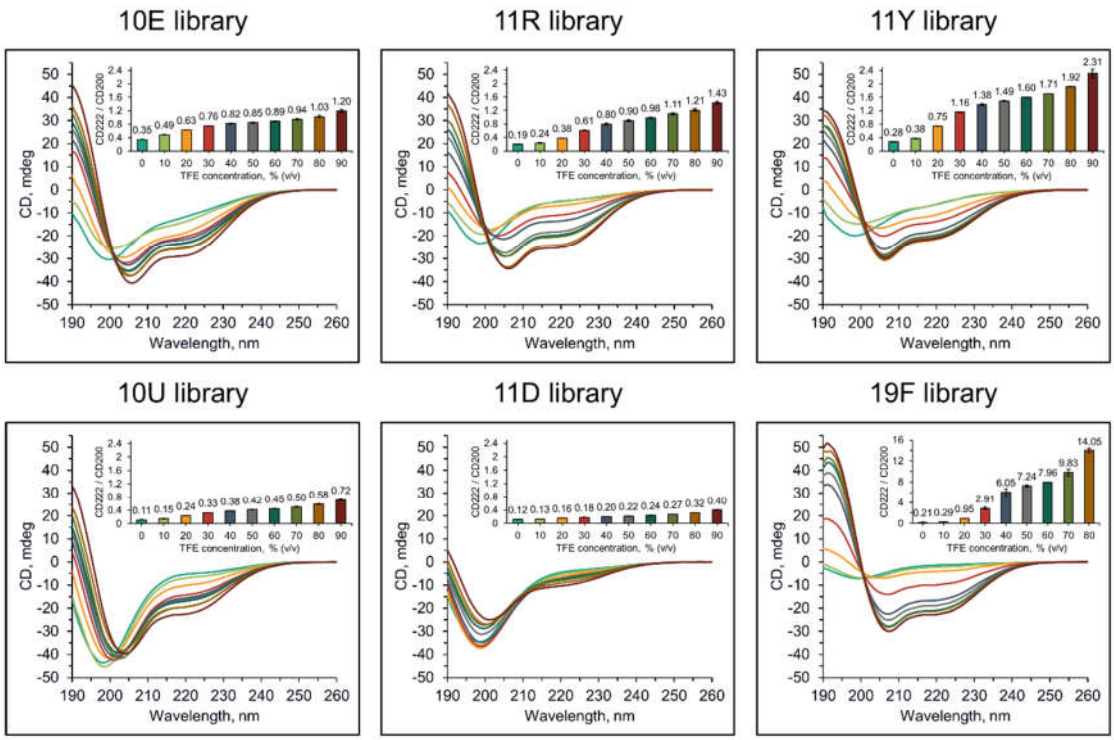


Figure 4. The effect of 2,2,2-trifluoroethanol on the secondary structure of 25-mer combinatorial peptide libraries. CD spectra were measured at 0.2 mg/ml nominal concentration in 10 mM ABP buffer (pH 7.4) with 0–90% (v/v) 2,2,2-trifluoroethanol. The inlet graph shows the ratio of CD signal at 222 to 200 nm.

Discussion

The absence of ABA, Nva and Nle in the proteinogenic alphabet has been considered striking given their high prebiotic abundance (30). Some workers have commented that this universal feature of biochemistry is a potential example of a “frozen accident” due to the absence of an obvious advantage of the canonical aliphatic residues over these ncAAs. At the same time, Weber and Miller hypothesized that inclusion of linear aliphatic amino acids would increase the side chain mobility and hence would not promote formation of ordered tertiary structure (8). Our experimental observations support this hypothesis. Moreover, structural prediction of 5,000 randomly chosen peptides by PEPstrMOD (using molecular dynamics based on force field libraries) confirms the same trend (31) (Fig. 5). In essence, even though unbranched amino acids are slightly more hydrophobic than their branched peers (though the difference is slight), they are expected to be much less adept at packing internal hydrophobic cores, a prerequisite to fold globular domains (32,33). This model neatly explains why the 10U library has low secondary structure-forming potential (note that for small peptides, the hydrogen bonding associated with secondary structure is normally isoergonic between unfolded and folded states, so the driving force for secondary structure formation arises from coupling to hydrophobic collapse). The lower packing efficiency of these side chains simultaneously explains the high solubility of the 10U library as well. Hence, the requirement of foldability most likely acted to purge ABA, Nva, and Nle from the pool of amino acids that ultimately became biotic.

The more prebiotically abundant diamino acids, DAP and DAB, are also strikingly absent in the modern amino acid alphabet. One possible scenario is that DAP and DAB were utilized in an ancient alphabet, and then were ultimately displaced with Arg, Lys, and His. Another school of thought holds that proteins lacked positively charged residues until biosynthetic pathways for the prebiotically-unavailable Arg, Lys and His evolved (7,34). While the second theory might seem unlikely at first glance, proteins lacking positively charged amino acids have recently been shown capable of folding and even binding with nucleic acids, assisted by metal ion cofactors (22,35–37). Our results concerning the 11D library further supports the second theory.

We find that when DAB is incorporated along with the 10E alphabet, it *diminishes* secondary structure potential. Hence, its inclusion would have been a step *backward* rather than a step *forward* toward the goal of assembling foldable polypeptides. The same trend is observed upon PEPstrMOD prediction (Fig. 5). The 11D library is also exceptionally soluble, so the combination of high solubility and low foldability implies that 11D polypeptides are mostly unfolded in aqueous solution.

Why does DAB harm protein structure propensity so sharply? It has been previously pointed out that these short-chain amines suffer from lactamization and acyl migration in peptides (8,38). This point is important concerning the long-term stability of peptide chains, but it probably cannot explain the results in our study given that lactamization is a pretty slow reaction, and we took care to perform experiments on fresh samples. A second possibility that we propose is that DAB disrupts structural potential because of a conflict associated with its simultaneous presence with Asp and Glu, i.e., negatively charged AAs that are also on short side-chains. Especially in these short 25-mer polypeptides, charged residues are likely to be on the surface. However, by having surface-anions and surface-cations in such close proximity, a polypeptide would be inclined to fold in such a way to favour formation of ion-pair salt bridges. The backbone angles required to form such salt bridges might be incommensurate with hydrophobic packing, hence the ion-pairs would stabilize the protein in an unpacked (and therefore unfolded) conformation. Naturally, there are two ways to avoid this folding problem: either place positive charges further away from the backbone (on a longer chain) so that they will primarily interact with water in a distinct solvation shell; *or* do not include positive charges at all. We hypothesize that early proteins avoided ion-pair-induced unfolding using the second strategy, whereas modern proteins adopted the first, though this strategy only became available once Arg and Lys could be biosynthesized. One possible experiment that would further test this theory is that it would predict that proteins with short-chain basic residues and long-chain acidic residues would *also* support secondary structure formation, though this is a counterfactual scenario in the context of the actual set of amino acids that were prebiotically abundant.

Finally, among the different library subsets, only the 10E and 11Y libraries are affected by addition of metal ions (Supplementary Fig. S4). This is consistent with the view that metal cations played an especially prominent role in supporting the structure and folding of early, highly acidic peptides, as exemplified by the 10E library.

In summary, our study on the solubility and secondary structure propensity of several prebiotically-relevant amino acid alphabets supports the assertion that foldability played a critical early role in governing which amino acids ultimately became part of the canonical alphabet. Unbranched amino acids and short-chain basic amino acids were excluded, despite

their prebiotic abundance, because they 'over-solubilized' polypeptides by stabilizing their unfolded conformations. More broadly, this study supports the view that the early canonical alphabet (10E), despite its deficiencies compared to the modern alphabet, was remarkably adaptive at supporting folding for the earliest proteins.

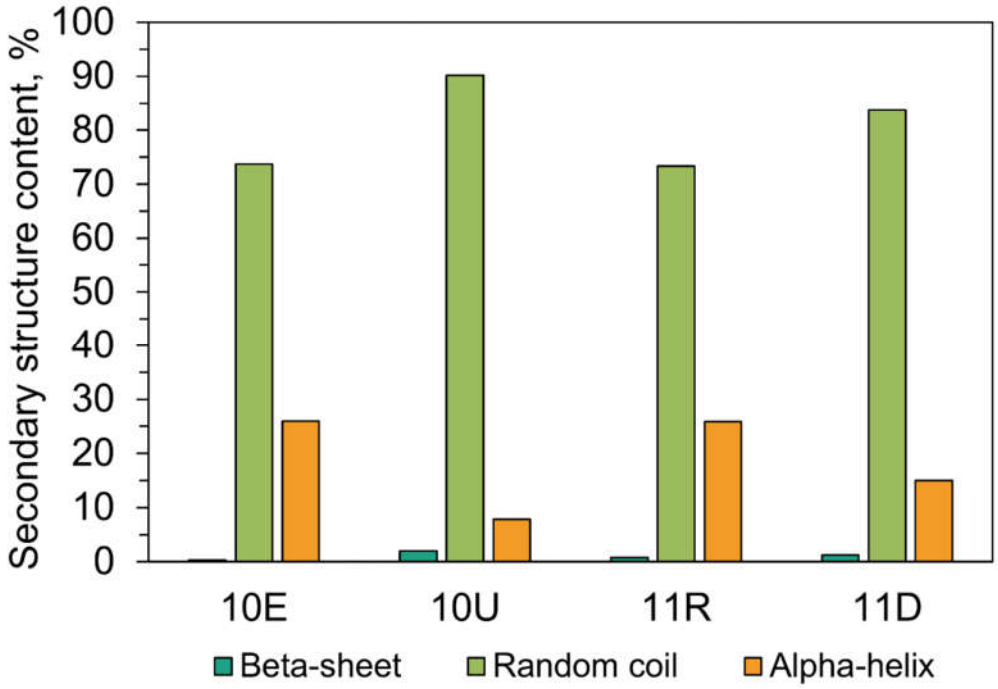


Figure 5. PEPstrMOD bioinformatic prediction of secondary structure of 5,000 sequences selected randomly from the 25-mer combinatorial peptide libraries.

Experimental Section

All chemicals if not stated otherwise were purchased from Sigma Aldrich.

Design and preparation of combinatorial peptide libraries

Split and mix synthesis of peptide libraries (39) has its limitations. Complete library of 25-mers composed of 10 amino acids would in theory require 10^{25} beads. One gram of solid phase support contains about 10^{6} beads – therefore it would cover only 10^{-21} percent of possible structures. To increase the probability of completeness of the possible structures it is necessary to conduct the synthesis considering individual molecules.

Average molecular weight of a 25-mer peptide composed of 10 amino acids is about 2,500 Daltons. Therefore 2.5 kg of these peptides would potentially contain 6×10^{23} individual peptides. In order to have a chance to synthesize a complete library, we would need 250 kg of peptides, where each peptide would be present only once. But the chance of finding any

given peptide in the mixture is about 70%. Therefore, studying a complete library is practically impossible.

We decided to synthesize the peptide mixture on 200 mg of the solid support and therefore create theoretically only 10^{-4}% of possible structures. The approach using individual molecules is still capable of production of 10^{17} times more possible structures than split and mix strategy. However, considering complete randomness of the synthetic process, we concluded that the sample is a representative collection of possible structures. One mg of synthetic peptide mixture contains about 2.4×10^{18} individual peptides.

Synthesis of the peptide mixture was accomplished by coupling mixtures of amino acids in which the ratio of individual amino acids was adjusted according to their reactivities. This approach was used by several authors (40) and various amino acid ratios were reported. We have used as the basis the ratios reported by Santi et al (41,42) and adjusted the ratios for unnatural amino acids based on their structure and our previous experience (Supplementary Table S1).

Synthesis was performed on the automatic peptide synthesizer Spyder Mark IV (43), using standard Fmoc peptide synthesis protocol. Fmoc amino acids were individually weighted (see Supplementary Table S1), mixed together and dissolved to create 0.3 M solution in 0.3 M N-hydroxybenzotriazole (HOBt) in dimethylformamide (DMF). Rink resin (200 mg, 0.42 mmol/g) was distributed into 10 ml syringes of the synthesizer and swelled in DMF for 10 minutes. The mixtures from Supplementary Table S1 were defined as extra amino acids 1 to 6 and placed into amino acid containers 21 to 26. Synthetic protocol was as follows: 2 × 1 min washing with DMF, 1 × 1 min and 1 × 20 min treatment with 20% piperidine in DMF, 4 × 1 min wash with DMF, 1 × 1 min wash with 0.3 M HOBt, 2 × 60 min coupling with amino acid mixture and 1 M diisopropylcarbodiimide in DMF. After 25 cycles of the synthesis, the resin was washed with DMF and dichloromethane and dried *in vacuo*. Dried resin was treated with 5 ml Mixture K (44) (trifluoroacetic acid-thioanisol-water-phenol-ethanedithiol, 82.5:5:5:5:2.5, v/v) for 2 hours. Resin was filtered off and peptide mixture was precipitated with diethylether three times and dried *in vacuo*. The pellet was then dissolved in 20% acetic acid and lyophilized. Prior to further experiments, all the peptide libraries were lyophilized three times with 1 mM HCl overnight in order to remove any residual trifluoroacetic acid.

Quality control of peptide libraries

The molecular weight distributions of combinatorial peptide libraries were confirmed by mass spectrometry using UltrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker, Germany) according to the standard procedure.

Solubility and aggregation propensity measurements

A series of ten Britton-Robinson buffers at pH 3.0, 5.0, 7.4, 9.0, and 11.0 was prepared by mixing 20 mM acetic acid, 20 mM boric acid and 20 mM phosphoric acid and adjusting pH to the desired value with 5 M NaOH. The ionic strength was adjusted to either 50 mM (low ionic strength) or 500 mM (high ionic strength) with NaCl, and all Britton-Robinson buffers were filter-sterilized using 0.22 µm PVDF membrane before use. Lyophilized peptide libraries were thoroughly resuspended in autoclaved MilliQ water to 5 mg/ml and subsequently diluted ten times into Britton-Robinson buffer to the final concentration of 0.5 mg/ml. The samples were

gently shaken at room temperature for 30 min and then centrifuged at maximum speed (21,300 × g) for 15 min at 4 ℃ in order to remove insoluble fraction. The supernatant (soluble fraction) was taken and used for solubility and aggregation propensity analyses.

The relative amount of peptides in soluble fractions was estimated spectrophotometrically by absorption of peptide bonds at 215 nm for all peptide libraries except the 19F peptide library. Due to the extremely low solubility, solubility of 19F peptide library was estimated by fluorescence spectroscopy using Pierce Quantitative Fluorometric Peptide Assay (Thermo Fisher Scientific, USA), and 0.5 mg/ml solution in DMSO was used as a standard. Spectrophotometric measurements were performed five times, whereas fluorometric measurements were performed in triplicates.

The relative amount of soluble aggregates was estimated by size-exclusion chromatography. The 100 µl aliquot of supernatant was loaded onto Superdex 75 Increase 10/300 GL column (Cytiva) that was pre-equilibrated with two bed volumes of the corresponding 20 mM Britton-Robinson buffer. The peptides were eluted from the column by one bed volume of the buffer at 0.5 ml/min flow rate at room temperature, and the eluted peptides were detected by absorption at 215 nm.

Circular Dichroism spectroscopy

The CD spectra were recorded using a Chirascan-plus spectrophotometer (Applied Photophysics, UK) over the wavelength range 190–260 nm in steps of 1 nm with an averaging time of 1 s per step. Cleared samples at 0.2 mg/ml nominal concentration in 1 mm path-length quartz cells were placed into a cell holder and spectra were recorded at room temperature. The CD signal was obtained as ellipticity in units of millidegrees and the resulting spectra were averaged from two scans and buffer-spectrum subtracted. All CD measurements were performed twice, and the resulting ratio of ellipticity at 222 nm to ellipticity at 200 nm was averaged and used to estimate the secondary structure content in peptide libraries.

To estimate the effect of pH on secondary structure, 5 mg/ml suspensions of peptide libraries in MilliQ water were diluted into 10 mM Britton-Robinson buffers at pH 3.0; 5.0; 7.4; 9.0 and 11.0 to the final concentration of 0.2 mg/ml, gently mixed at room temperature for 30 min and then centrifuged at maximum speed (21,300 × g) for 15 min at 4 ℃ in order to remove insoluble fraction.

To estimate the effect of 2,2,2-trifluoroethanol on secondary structure, 5 mg/ml suspensions of peptide libraries in MilliQ water were diluted in a series of 10 mM Britton-Robinson buffer (pH 7.4) containing 0–90% (v/v) 2,2,2-trifluoroethanol.

To estimate the effect of metal ions on secondary structure, 5 mg/ml suspensions of peptide libraries in MilliQ water were diluted in a series of 10 mM Tris buffer at pH 7.4 supplemented with 0; 10; 100; and 1,000 µM mixture of NaCl, KCl, $MgCl_2$, $MnCl_2$, and $ZnCl_2$.

Bioinformatic analysis of in silico peptide libraries

5,000 25-mer sequences were randomly generated *in silico* for the 10E, 10U, 11R and 11D libraries and their structure was predicted using PEPstrMOD (31). PEPstrMOD uses force field libraries Forcefield_NCAA and Forcefield_PTM and integrates them in Molecular Dynamics (MD) simulations using AMBER. Additionally, it implements the force field library

SwissSideChain with the MD package GROMACS. The parameters employed in the MD simulations were 50 picoseconds of simulation time and the selected peptide environment was set to "vacuum".

DSSP (45) was employed to determine the secondary structure elements from the peptide tertiary structures generated by PEPstrMOD. The obtained annotations of secondary structure were grouped into three larger classes: helix (H: α-helix, G: 310 helix and I: π-helix), sheet (B: isolated β-bridge and E: extended strand) and loop (T: hydrogen bonded turn, S: bend and "_" blank spaces).

Acknowledgements

We would like to thank Dr. Vyacheslav Tretyachenko and Dr. Valerio Guido Giacobelli for helpful discussions about this project. This work was supported by the Human Frontier Science Program grant HFSP-RGY0074/2019. M. M., A. C. S. R., and I. C. acknowledge support by the project the "Grant Schemes at CU" (reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935), project no. START/SCI/148. I.C. is supported by the Visegrad Fund Scholarship (no. 52110039). We acknowledge CMS CF Biophysic of CIISB, Instruct-CZ Centre, supported by MEYS CR (LM2018127)) and European Regional Development Fund-Project „UP CIISB" (No. CZ.02.1.01/0.0/0.0/18_046/0015974).

Competing Interest Statement

The authors declare no competing interests.

References

1. Milner-White EJ, Russell MJ. Functional capabilities of the earliest peptides and the emergence of life. Genes (Basel). 2011;2(4):671–88.

2. Fried SD, Fujishima K, Makarov M, Cherepashuk I, Hlouchova K. Peptides before and during the nucleotide world: An origins story emphasizing cooperation between proteins and nucleic acids. J R Soc Interface. 2022;19(187).

3. Frenkel-Pinter M, Frenkel-Pinter M, Samanta M, Ashkenasy G, Leman LJ, Leman LJ. Prebiotic Peptides: Molecular Hubs in the Origin of Life. Chem Rev. 2020;120(11):4707–65.

4. Freeland S. Undefining life's biochemistry: Implications for abiogenesis. J R Soc Interface. 2022;19(187).

5. Cleaves HJ. The origin of the biologically coded amino acids. J Theor Biol. 2010;263(4):490–8.

6. Trifonov EN. The triplet code from first principles. J Biomol Struct Dyn. 2004;22(1):1–11.

7. Higgs PG, Pudritz RE. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology. 2009;9(5):483–90.

8. Weber AL, Miller SL. Reasons for the occurrence of the twenty coded protein amino acids. J Mol Evol. 1981;17(5):273–84.

9. Cleaves HJ. The origin of the biologically coded amino acids. J Theor Biol [Internet]. 2010;263(4):490–8. Available from: http://dx.doi.org/10.1016/j.jtbi.2009.12.014

10. Freeland S. "Terrestrial" Amino Acids and their Evolution. Amin Acids, Pept Proteins Org Chem. 2010;1:43–75.

11. Vázquez-Salazar A, Lazcano A. Early Life: Embracing the RNA World. Curr Biol. 2018;28(5):R220–2.

12. Raggi L, Bada JL, Lazcano A. On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. Phys Chem Chem Phys. 2016;18(30):20028–32.

13. Bredehöft J, Thiemann W, Jessberger E, Carob G, Meierhenrich U. Identification of diamino acids in the Murichison meteorite. Proc Natl Acad Sci U S A. 2004;(25):9182–6.

14. Wolman Y, Haverland WJ, Miller SL. Nonprotein Amino Acids from Spark Discharges and Their Comparison with the Murchison Meteorite Amino Acids. Proc Natl Acad Sci. 1972;69(4):809–11.

15. Alvarez-Carreño C, Becerra A, Lazcano A. Norvaline and Norleucine May Have Been More Abundant Protein Components during Early Stages of Cell Evolution. Orig Life Evol Biosph. 2013;43(4–5):363–75.

16. Apostol I, Levine J, Lippincott J, Leach J, Hess E, Glascock CB, et al. Incorporation of norvaline at leucine positions in recombinant human hemoglobin expressed in Escherichia coli. J Biol Chem. 1997;272(46):28980–8.

17. Martin W, Russell MJ. On the origin of biochemistry at an alkaline hydrothermal vent. Philos Trans R Soc B Biol Sci. 2007;362(1486):1887–925.

18. Grotzinger JP, Kasting JF. New constraints on Precambrian ocean composition. J Geol. 1993;101(2):235–43.

19. Tanaka J, Doi N, Takashima H, Yanagawa H. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. Protein Sci. 2010;19(4):786–95.

20. Newton MS, Morrone DJ, Lee KH, Seelig B. Genetic Code Evolution Investigated through the Synthesis and Characterisation of Proteins from Reduced-Alphabet Libraries. ChemBioChem. 2019;20(6):846–56.

21. Tretyachenko V, Vymětal J, Bednárová L, Kopecký V, Hofbauerová K, Jindrová H, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Sci Rep. 2017;7(1):2–10.

22. Tretyachenko V, Neuwirthová T. Modern and prebiotic amino acids support distinct structural profiles in proteins. Open Biol. 2022;12(220040).

23. Kozlowski LP. Proteome-pI: Proteome isoelectric point database. Nucleic Acids Res. 2017;45(D1):D1112–6.

24. Wang Y, Van Oosterwijk N, Ali AM, Adawy A, Anindya AL, Dömling ASS, et al. A Systematic Protein Refolding Screen Method using the DGR Approach Reveals that Time and Secondary TSA are Essential Variables. Sci Rep. 2017;7(1).

25. Reynolds JA, Gilbert DB, Tanford C. Empirical Correlation Between Hydrophobic Free Energy and Aqueous Cavity Surface Area. Proc Natl Acad Sci. 1974;71(8):2925–7.

26. Vymětal J, Vondrášek J, Hlouchová K. Sequence versus composition: What prescribes IDP biophysical properties? Entropy. 2019;21(7).

27. Kelly SM, Jess TJ, Price NC. How to study proteins by circular dichroism. Biochim Biophys Acta - Proteins Proteomics. 2005;1751(2):119–39.

28. Buck M. Trifluoroethanol and colleagues: Cosolvents come of age. Recent studies with peptides and proteins. Q Rev Biophys. 1998;31(3):297–355.

29. Starzyk A, Barber-Armstrong W, Sridharan M, Decatur SM. Spectroscopic evidence for backbone desolvation of helical peptides by 2,2,2-trifluoroethanol: An isotope-edited FTIR study. Biochemistry. 2005;44(1):369–76.

30. Zaia DAM, Zaia CTBV, De Santana H. Which amino acids should be used in prebiotic chemistry studies? Orig Life Evol Biosph. 2008;38(6):469–88.

31. Singh S, Singh H, Tuknait A, Chaudhary K, Singh B, Kumaran S, et al. PEPstrMOD: Structure prediction of peptides containing natural, non-natural and modified residues. Biol Direct. 2015;10(1).

32. Tanford C. The hydrophobic effect and the organization of living matter. Science (80-). 1978;200(4345):1012–8.

33. Dill KA. Dominant forces in protein folding. Biochemistry. 1990;29:7133.

34. Trifonov EN. Consensus temporal order of amino acids and evolution of the triplet code. Gene. 2000;261(1):139–51.

35. Longo LM, Blaber M. Protein design at the interface of the pre-biotic and biotic worlds. Arch Biochem Biophys. 2012;526(1):16–21.

36. Despotović D, Longo LM, Aharon E, Kahana A, Scherf T, Gruic-Sovulj I, et al. Polyamines mediate folding of primordial hyperacidic helical proteins. Biochemistry. 2020;59(46):4456–62.

37. Giacobelli VG, Fujishima K, Lepšík M, Tretyachenko V, Kadavá T, Makarov M, et al. In Vitro Evolution Reveals Noncationic Protein-RNA Interaction Mediated by Metal Ions. Mol Biol Evol. 2022;39(3):1–11.

38. Poduska K, Katrukha GS, Silaev AB, Rudinger J. Amino Acids and Peptides. LII. Intramolecular Aminolysis of Amide Bonds in Derivatives of alpha,gamma-diaminobutyric acid, alpha,beta-diaminoproprionic acid, and ornithine. Collect Czechoslov Chem Commun. 1965;30:2410–33.

39. Lam KS, Lebl M, Krchňák V. The “one-bead-one-compound” combinatorial library method. Chem Rev. 1997;97(2):411–48.

40. Ostresh JM, Winkle JH, Hamashin VT, Houghten RA. Peptide libraries: Determination of relative reaction rates of protected amino acids in competitive couplings. Biopolymers. 1994;34(12):1681–9.

41. Rutter W, Santi D. Peptide mixtures [Internet]. US-5420246-A, 1995. Available from: https://pubchem.ncbi.nlm.nih.gov/patent/US-5420246-A

42. Ivanetich KM, Santi D V. Preparation of equimolar mixtures of peptides by adjustment of activated amino acid concentrations. Methods Enzymol. 1996;267:247–58.

43. Lebl M, Flegelova Z, Poncar P, Mudra P, Paces O, Knor M, et al. Device for parallel oligomer synthesis, method of parallel oligomer synthesis and use thereof. WO-2019115514-A1, 2017.

44. King DS, Fields CG, Fields GB. A cleavage method which minimizes side reactions following Fmoc solid phase peptide synthesis. Int J Pept Protein Res. 1990;36(3):255–66.

45. W K, C. S. Dictionary of Protein Secondary Structure: Pattern Recognition of

Hydrogen-Bonded and Geometrical Features. Biopolymers. 1983;22(12):637–2577.

Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability

Mikhail Makarov[1], Alma C. Sanchez Rocha[1], Robin Krystufek[2,3], Ivan Cherepashuk[1], Volha Dzmitruk[4], Tatsiana Charnavets[4], Anneliese M. Faustino,[5] Michal Lebl[3], Kosuke Fujishima[6,7], Stephen D. Fried[5,8], and Klara Hlouchova*[1,3]

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

[2]*Department of Physical Chemistry, 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]*Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, 25250 Vestec, Czech Republic*

[5]*Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21212, United States*

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

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

[8]*Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21212, United States*

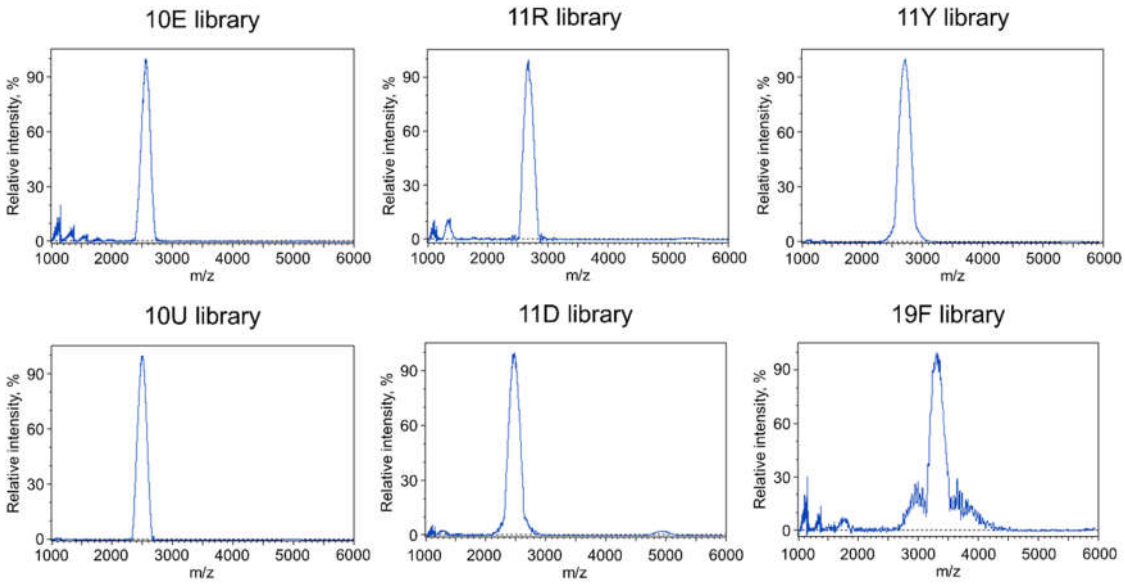
* To whom correspondence may be addressed: klara.hlouchova@natur.cuni.cz

Supplementary material

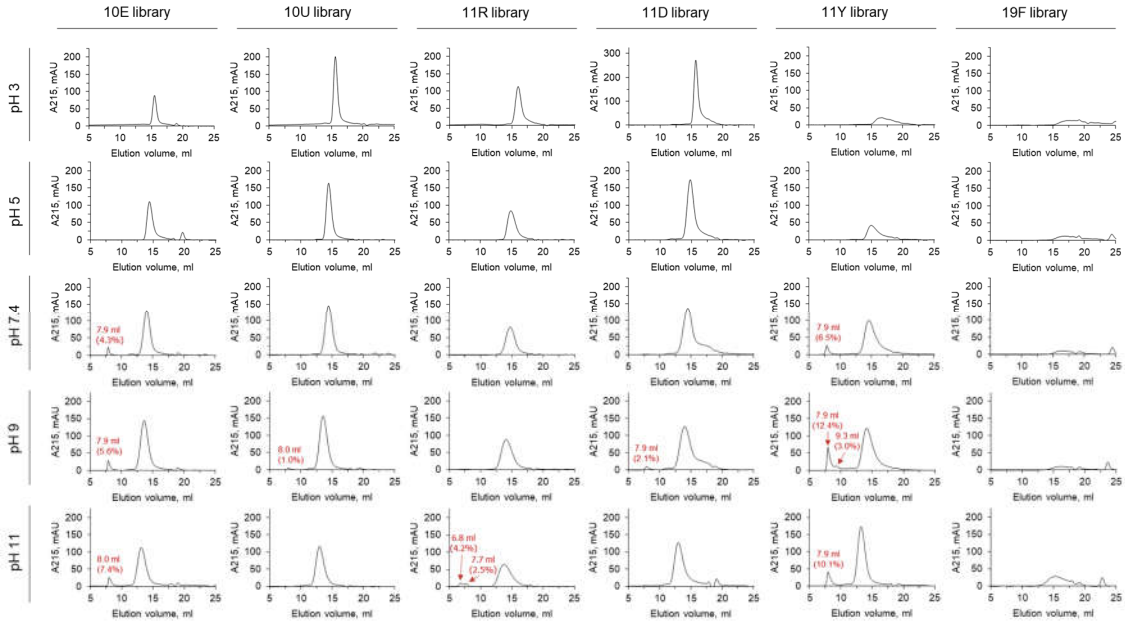
Supplementary Table S1: Weights (mg) of protected amino acids in individual mixtures used for synthesis

	19F	10E	11R	11Y	10U	11D
Fmoc-Ala	202	332	300	312	508	303
Fmoc-Asp(OtBu)	276	455	409	426	694	414
Fmoc-Glu(OtBu)	296	489	440	457	745	444
Fmoc-Phe	187					
Fmoc-Gly	164	270	243	253	412	246
Fmoc-His(Trt)	422					
Fmoc-Ile	1,171	1,934	1,740	1,809		1,759
Fmoc-Lys(Boc)	556					
Fmoc-Leu	335	553	497	517		503
Fmoc-Met	163					
Fmoc-Asn(Trt)	609					
Fmoc-Pro	278	460	414	430	702	418
Fmoc-Gln(Trt)	620					
Fmoc-Arg(Pmc)	807		1199			
Fmoc-Ser(But)	203	336	302	314	512	306
Fmoc-Thr(But)	363	599	539	560	914	545
Fmoc-Val	731	1,208	1,086	1,130		

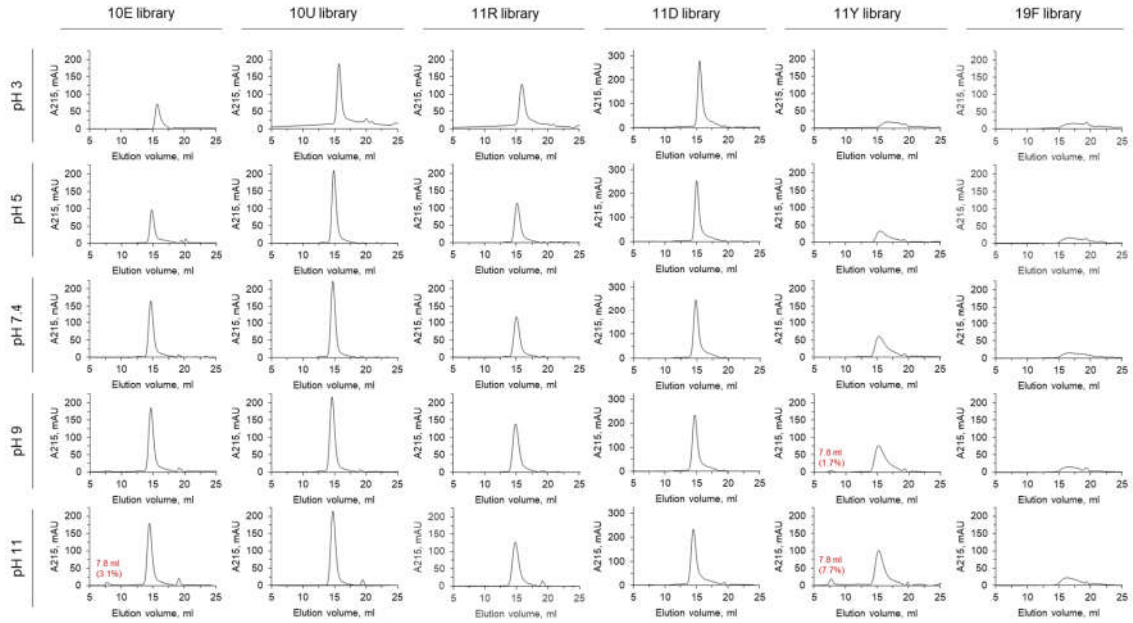
Fmoc-Trp(Boc)	254			
Fmoc-Tyr(But)	363	560		
Fmoc-Nva			752	
Fmoc-Nle			775	
Fmoc-Aba			648	
Fmoc-Dab(Boc)				734



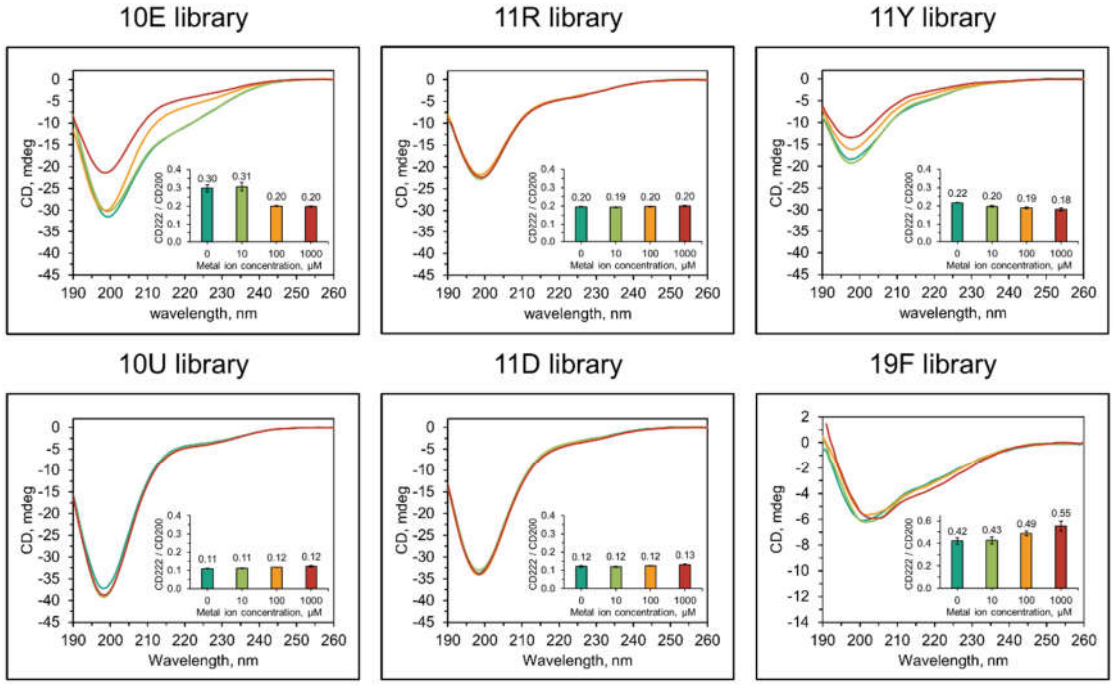
Supplementary Figure S1. Molecular weight distribution of 25-mer combinatorial peptide libraries estimated by MALDI-TOF-MS using sinapinic acid (SA) matrix.



Supplementary Figure S2. Aggregation propensity of 25-mer combinatorial peptide libraries at different pH and low ionic strength (50 mM NaCl). Aggregation propensity was measured for 0.5 mg/ml nominal peptide library solutions in a series of 20 mM ABP buffers (pH 3–11) by size-exclusion chromatography.



Supplementary Figure S3. Aggregation propensity of 25-mer combinatorial peptide libraries at different pH and high ionic strength (500 mM NaCl). Aggregation propensity was measured for 0.5 mg/ml nominal peptide library solutions in a series of 20 mM ABP buffers (pH 3–11) by size-exclusion chromatography.



Supplementary Figure S4. The effect of metal ions on the secondary structure of 25-mer combinatorial peptide libraries. CD spectra were measured for 0.2 mg/ml nominal peptide library solutions in a series of 10 mM Tris buffers at pH 7.4 supplemented with 0 (green line); 10 (yellow line); 100 (orange line); and 1,000 (red line) μM mixture of NaCl, KCl, $MgCl_2$, $MnCl_2$, and $ZnCl_2$. The inlet graph shows the ratios of CD signal at 222 to 200 nm.