Charles University

Faculty of Science

Study programme: Biochemistry

Branch of study: 4XBIOCH

Mgr. Vjačeslav Treťjačenko

The effect of amino acid repertoire on protein structure and function

Vliv repertoáru aminokyselin na strukturu a funkci bílkovin

Doctoral thesis

Supervisor: Mgr. Klára Hlouchová Ph.D.

Prague, 2021

DECLARATION

I declare that I have worked on this thesis under the guidance of my supervisor and that all sources of the previous knowledge are properly cited. No part of this work was used and will not be used for obtaining any other academic degree than Ph.D. from Charles University.

Prague ……………….

……………..…………………

Mgr. Vjačeslav Treťjačenko

DECLARATION OF AUTHORSHIP

I declare that Mgr. Vjačeslav Treťjačenko contributed significantly to the experiments and to all 3 scientific publications contained in this Ph.D. thesis. He performed most of the experiments, substantially contributed to their planning, and took a significant part in the primary data interpretation and their preparation for publication.

Prague ……………….

…...…...…………………

Mgr. Klára Hlouchová, Ph.D.

To my grandfather Vladimir.

ACKNOWLEDGEMENTS

Here I would like to thank in the first place to Klára Hlouchová. We shared many scientific ups and downs over the extraordinary 8 years. I had an honor to participate in the first discussions which started in a two membered club of astrobiology enthusiasts and expanded up to the internationally recognized scientific group with numerous collaborations all around the globe. Thank you for the endless supply of tea, being a good friend and showing us all, that people come first in every aspect of Science.

Also, I would like to thank to all members of the group, I enjoyed much of your company inside and outside of the lab. I am grateful to be around and for the opportunities to talk about your life, science, and everything. Special thanks are reserved for Dr. Jan Klouda for his unbiased review of this work.

Last but not least, I would like to thank my family for being supportive, encouraging me to find my own place in the world and being always by my side.

ABSTRACT (in english)

To understand protein structure emergence is to comprehend the evolutionary transition from messy chemistry to the first heritable molecular systems. Early proteins were probably flexible in structure, promiscuous in activity and ambiguous in sequence. Moreover, first sequences were presumably composed of prebiotically plausible amino acids from endogenous and exogenous sources which form only a subset of the extant protein alphabet. Here we investigate the effect of most recent additions to the amino acid alphabet on protein structure/function relationship and the properties of random proteins as the evolutionary point-zero for the earliest sequences as well as for proteins emerging *de novo* from the non-coding parts of the genome. Random or never born proteins are of a special interest for the contemporary biology as they unveil the unexposed side of the protein sequence space. We constructed an *in silico* library of random proteins with the natural amino acid alphabet, analyzed its structure/disorder/aggregation content and selected 45 sequences for subsequent experimental preparation and biophysical characterization. We observed that structure content in random sequence space does not differ significantly from the natural proteins. However, the analyses of the aggregation propensity showed a significant level of optimization in natural protein space. Experimental characterization led to the surprising discovery of random disordered proteins being the most tolerated sequences upon the *in vivo* expression. Next, we designed a high throughput pipeline for experimental library preparation with proteins composed either of canonical 20 amino acids as well as of prebiotically plausible set of 10 amino acids. In order to implement this design experimentally we built CoLiDe – COmbinatorial Library Design tool based on degenerate codon composition optimization. We designed the libraries using CoLiDE, prepared them in a cell free expression system, and tested their properties by means of chaperone interaction analysis and selective proteolysis. Preliminary results suggest structure formation in prebiotic amino acid library and higher disorder content in canonical amino acid library of random proteins. Subsequently, as a case study we analyzed structure and function of contemporary protein dephospho coenzyme A kinase upon substitution of its aromatic amino acids by their prebiotically plausible counterparts. This analysis showed that protein function can be maintained in the absence of aromatic amino acids although structure is inevitably destabilized. Moreover, we observe significant structural changes upon ligand binding in aromatic-less mutants foreshadowing the essential effects of ancient cofactors on early protein stabilization.

Overall, this thesis represents one of the first windows into properties of evolutionary early proteins, with respect to prebiotically plausible amino acids. Its results imply that even proteins composed of prebiotically early amino acids have structural and functional propensities and could play an important role in the early biosphere.

ABSTRAKT (in czech)

Porozumění původu prvotních proteinů je pochopením přechodu komplexních chemických směsí k prvním biologickým systémům. Prvotní proteiny byly pravěpodobně strukturně flexibilní, s promiskuitní aktivitou a se sekvencemi představujícími spíše fyzikálně chemické vlastnosti než definované sekvenční motivy. Rané proteiny byly rovněž pravděpodobně složeny pouze z prebioticky dostupných aminokyselin z endogenních a exogenních zdrojů. V této práci jsme se zaměřili jak na studium vlivu nejpozdějších přírůstků aminokyselinového repertoáru na strukturu a funkci proteinů tak na charakterizaci nahodných sekvencí jakožto prekurzorů pro vznik nejranějších tak i současných proteinů generovaných z původně transkripčně/translačně neaktivních oblasti genomu. Výzkum náhodných proteinů je obzvlášt zajimavý z pohledu neprobádáné strany světa proteinových sekvencí. V této práci jsme charakterizovali *in silico* soubor náhodných proteinových sekvencí s přirozenými výskyty aminokyselin pomocí predikce sekundárních struktur/proteinové nesupořádánosti/agregace a rovněž jsme vybrali 45 sekvencí pro následující *in vitro* charakterizaci. Pomocí analýzy *in silico* knihovny jsme mohli konstatovat, že výskyt sekundárních struktur v náhodném sekvenčním prostoru není výrazně odlišný od toho v přírodních proteinech. Na druhou stranu, evoluční optimizace se nejvíce projevovala v antiagregačních vlastnostech přirozených proteinových sekvencí. Experimentální charakterizace vedla k překvapivému odhalení, že neuspořádáné sekvence jsou nejvíce tolerovanými náhodnými proteiny *in vivo.* Následně jsme připravili experimentální strategii pro charakterizaci proteinových knihoven složených z 20 a z prebioticky dostupných 10 aminokyselin. Za účelem experimentální charakterizace těchto knihoven jsme navrhli algoritmus CoLiDe pro optimizaci aminokyselinových poměrů v rozsáhlých knihovnách náhodných proteinů pomocí kombinace degenerovaných kodonů. S použitím CoLiDe jsme připravili obě knihovny a otestovali jejích vlastnosti *in vitro* pomocí selektivní proteolýzy a vyhodnocení interakcí s chaperony. Předběžné výsledky naznačují vyšší přítomnost struktury v knihovně proteinu s prebiotickým aminokyselinovým složením a vysokou neuspořádanost knihovny složené ze všech 20 proteinogenních aminokyselin. V poslední studii této práce jsme vyhodnotili vliv substituce všech aromatických aminokyselin v sekvenci defosfo koenzym A kinázy jejími prebiotickými protějšky. Pomocí této modifikace jsme ukázali, že protein je schopen funkce při absenci aromatických aminokyselin i přes značnou destabilizaci terciární struktury. Pozoruhodným výsledkem byla výrazna změna struktury proteinu bez aromatických aminokyselin při interakci s ligandy jenž naznačuje klíčovou roli kofaktorů při stabilizaci raných proteinových struktur.

Táto práce je vhledem do evolučně nevyvinutého sekvenčního prostoru proteinů s důrazem na charakterizací rané proteinové abecedy. Výsledky disertace naznačují, že proteiny složené z raných aminokyselin disponují strukturními a funkčními vlastnostmi jenž mohly hrát důležitou roli v časech prvotního vývoje biosféry.

ABBREVIATIONS

10E	combinatorial protein library constructed with 10 amino acid alphabet
20F	combinatorial protein library constructed with 20 amino acid alphabet
AAA+	ATPases associated with a variety of cellular activities
ANS	8-anilinonaphthalene-1-sulfonic acid
CoA	coenzyme A
CoLiDe	combinatorial library design
dCoA	dephospho coenzyme A
DLS	dynamic light scattering
DPCK	dephospho coenzyme A kinase
dsDNA	double stranded DNA
ECD	electronic circular dichroism
EDTA	ethylendiaminetetraacetic acid
GOE	great oxidation event
Gy	billion years
HPLC-MS	high performance liquid chromatography coupled with mass spectrometric detection
HSP	heat shock protein
HTS	high throughput sequencing
IDP	intrinsically disordered protein
IDT	Integrated DNA Technologies
LUCA	last universal common ancestor
MALDI-TOF	matrix assisted laser desorption ionisation with time of flight detection
NBD	nucleotide binding domain
NBP	never born protein
NEB	New England Biolabs
NMR	nuclear magnetic resonance
PCR	polymerase chain reaction
PDB	protein data bank
PVDF	polyvinylidene fluoride
RBS	ribosome binding site
SBD	substrate binding domain
SCOP	structural classification of proteins
SDS-PAGE	polyacrylamide gel electrophoresis in presence of sodium dodecyl sulphate
ssDNA	single stranded DNA

TABLE OF CONTENTS

INTRODUCTION

1. Proteins

Proteins are versatile functional biopolymers utilised by all contemporary biological systems. Protein functions include chemical catalysis, interactions with ligands, processing of cellular genetic information, and mediation of intra and extracellular signalisation. Structurally, proteins are linear polyamide condensates of various lengths distinguished by their primary sequence, i.e., a specific succession of monomeric building blocks – amino acids. However, the most striking feature of proteins is their ability to fold into the compact three-dimensional structure. The ability of proteins to fold can be defined by local and global structural transitions of a polypeptide chain. On a local level, C=O carbonyl and N-H secondary amine groups of protein backbone interact to create secondary structures. Secondary structures manifest in helical backbone organisation if interactions are short-ranged or in pleated backbone stretches in the case of long-range interactions between distant residues of the protein chain. Further local arrangements of a single backbone combined with a global transition of protein chain form a unique spatial organisation

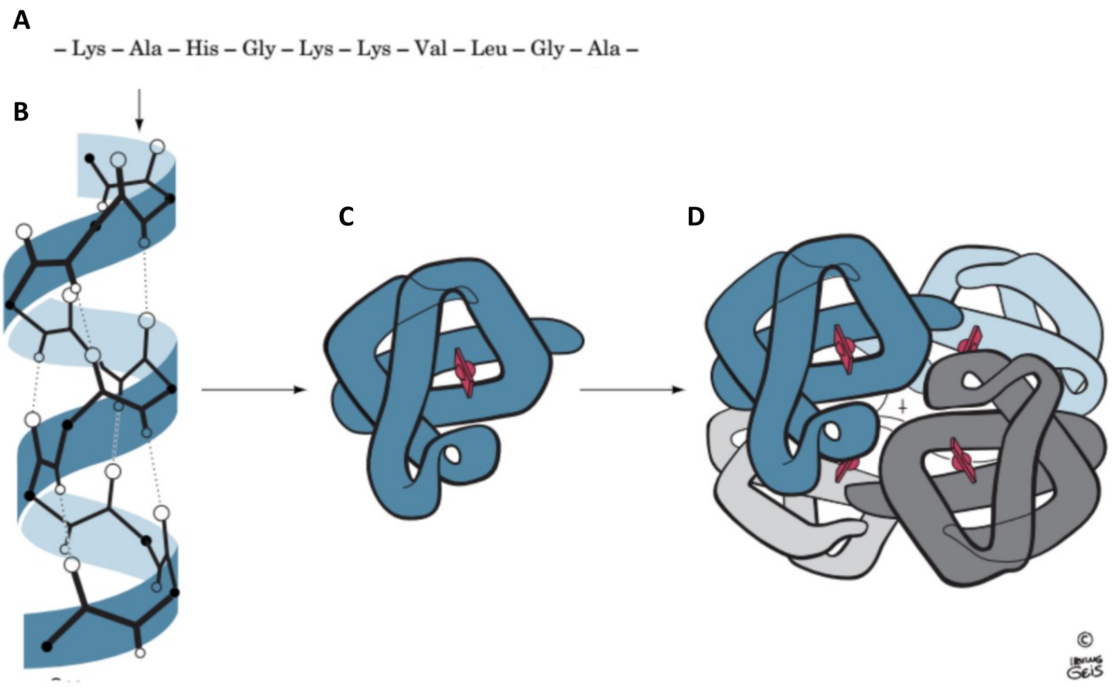

A
– Lys – Ala – His – Gly – Lys – Lys – Val – Leu – Gly – Ala –
B
C
D
©

Figure 1. Overview of protein structures. (A) Primary structure represented as amino acid sequence of protein. (B) Secondary structure is characterized by local arrangement of protein backbone. (C) Tertiary structure is a global conformation of protein chain and (D) quarternary structure is a composition of multiple protein chains. Figure adapted from [1]

known as tertiary structure. In some proteins, the final structure is composed of several tertiary units. This composed arrangement of several protein chains represents the quaternary protein structure (**Fig. 1**) [1].

Mechanistically, protein folding can take place independently, via interaction with ligands (folding upon binding) or with the help of specialised assisting proteins – chaperones. Independent protein folding relies on a compact core formation with hydrophobic amino acids situated in the protein interior and polar amino acids on the protein surface. The resulting collapsed structure is protected from spontaneous unfolding by hydrophobic interactions in the core as well as by polar interactions with the solvent on the exterior. Furthermore, this stabilisation can be established by the interactions with different ligands – proteins, nucleic acids, or small molecules. In fact, a particular class of proteins – intrinsically disordered proteins (IDPs) – undergo the folding process primarily upon interaction with a ligand. These conditional folders attracted significant attention from the scientific community in the last two decades because of their unconventional structural, functional, and material properties (reviewed in [2]). They contrast folded proteins in amino acid composition, absence of a stable hydrophobic core, and low sequence conservation levels [3]. Functionally, such proteins are often promiscuous (i.e., interact with several binding partners) and serve as conditional binders in transcriptional and signalling pathways [4]. The last and equally important track to protein structure is assisted folding by chaperones and chaperonins. These molecules evolved specifically to facilitate the complex folding pathways and played a significant role in the expansion of the protein structural universe.

2. Evolution of proteins

The evolution of proteins and their structure presents a thought-provoking conundrum. It is estimated that 10^7–10^8 existing species with 10^3–10^5 proteins per proteome cover approximately 10^{10}–10^{13} different protein variants on Earth [5,6]. These numbers represent a minuscule fraction of the 10^{321}–10^{469} possible amino acid chain permutations given by estimates of average protein lengths in contemporary genomes [7,8]. How did Nature discover that specific sequence sub-space? Considering 10^{30} prokaryotic cells on Earth with turnover rates of $\sim 8\times10^{29}$ cells per year, and mutation rates of 4×10^{-7} per cell per generation, we can estimate at most $\sim 2\times10^{32}$ total amino acid mutations in microbial proteins over 4 Gy history of life, which is still a negligible number in comparison to the vast protein sequence space [6,9,10].

The first conserved functional molecules were likely related to simple biopolymers that were available when life originated. In the case of proteins, experimental and theoretical considerations suggest that early sequences consisted of abiotically available amino acids [11,12]. The prebiotic condensation of peptide bonds could be achieved simply by the elimination of water. Synthesis could occur by means of ATP/Mg^{2+} dependent chemical catalysis as was suggested by Martinez-Bachs *et al.* or by wet-dry cycling of the amino acid mixture on the mineral surfaces [13,14]. Interestingly, short prebiotically-plausible dipeptides were shown to exhibit catalytic activities such as asymmetric aldol condensation and even peptide bond synthesis [15,16]. However, the identity of the first prebiotic peptide sequences remains obscured.

In order to identify the common precursors of contemporary proteins, Alva *et al.* exploited a combination of both sequence and structure similarity comparisons and derived 40 ancient sequences found among otherwise non-homologous protein families [17]. These peptides are hypothesised to be vestiges of early life, and, interestingly, 14 of them establish folds by repetition. A different approach to the identification of protein evolution relics was taken by Caetano-Annoles and coworkers. The authors of that study built on an observation that protein structure distribution in nature is driven by a power-law – few folds are used in most proteins, while most of the folds are utilised a few times. Based on phylogenetic analyses of the most widely occurrent folds, the authors discovered that most ancient protein structures included P-loop NTPases (SCOP fold c.37), TIM beta/alpha barrel (c.1), NAD(P)-binding Rossman fold domains (c.2), DNA/RNA binding 3-helical bundle (a.4), and oligonucleotide/oligosaccharide binding fold (b.40) (**Fig. 2**) [18–21].

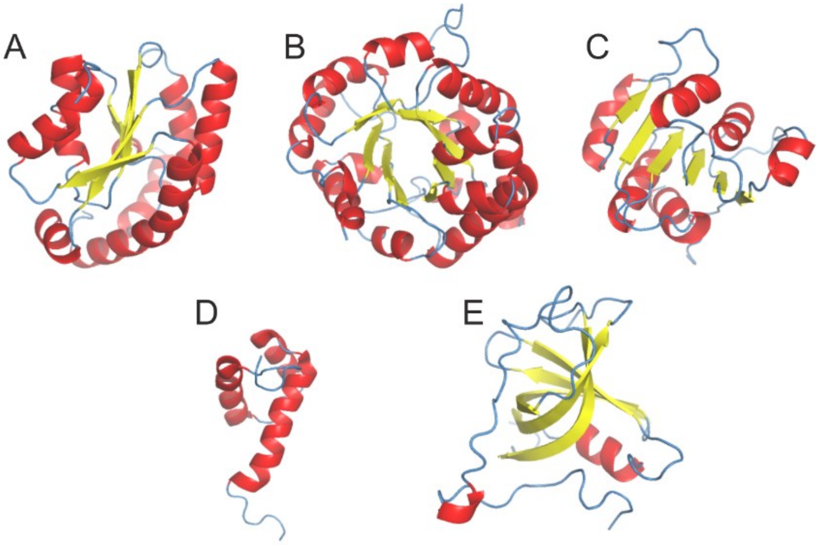


Figure 2. Most ancient protein folds by Caetano-Annoles and coworkers [18-21]. (A) P-loop NTPase, (B) TIM β/α barrel, (C) NAD(P)-binding Rossman fold, (D) DNA/RNA binding 3-helical bundle, (E) ololigonucleotide/oligosaccharide binding fold

The functions of these folds are notably connected with nucleotide and carbohydrate metabolism. Interestingly, some of the peptides identified by Alva *et al.* are prevalent in these ancient folds [17]. Goldman *et al.* investigated translation-related proteins as translation is considered an exemplary early protein function [22]. The study reported that 9 of the 10 most ancient folds identified by Caetano-Annoles *et al.* are widespread in these proteins; their functions involve RNA modification, binding, and phosphoryl transfer. Indeed, recent studies indicate that RNA-binding by flexible random coils was the ancestral function of proteins and that RNA facilitated the formation of first folded proteins [23,24]. Furthermore, protein structural flexibility is connected to higher sequence evolvability (i.e., the ability of a sequence to adopt new functions or structures rapidly); it has been shown that evolutionary younger protein families feature higher levels of structural flexibility [25–28].

Considering the sequences of early peptides, the first proteins were likely characterised by global physicochemical properties rather than specific sequence motifs. This condition could be described by Eigen's quasispecies model, which postulates that in the absence of precise repair mechanisms, the evolution is not directed towards the fittest sequence but towards the "cloud" of many sequences [29]. These earliest precursors of modern proteins were probably sufficient to provide the simplest functional and structural benefit to already existing catalytic structures of RNA. Subsequently, when the first protein-coding was established, these sequences reached opportunities governed by classic evolution mechanisms – mutate, recombine, and be selected up to the following generations.

3. Evolution of protein alphabet

The evolution of proteins is inevitably intertwined with the evolution of its coding DNA and, from the global perspective, with the evolution of genetic code itself. The amino acid repertoire of contemporary proteins is the result of constant organismal adaptation to the environment. Since biosynthetic possibilities of early life must have been limited, first proteins could emerge mainly from prebiotically available amino acids. These could be provided either by endogenous (formed on Earth) or exogenous (brought by extraterrestrial material, e.g., meteorites) sources. Despite the lack of direct evidence for the exact form of the early amino acid alphabet, the research community inclines to the scenario of alphabet divergence from simple prebiotic to more complex biosynthesised amino acids [11,12,17,24,30,31].

Miller and Urey's seminal experiment tested the abiogenesis of complex organic compounds from simpler prebiotic precursors [32]. The late analyses of Miller and Urey experiments showed that more than 22 different amino acids could be created by spark discharge in the mixture of water, methane, ammonia, and hydrogen gas. Among these 22, seven constitute part of the modern amino acid alphabet (V, G, A, D, S, E, and F). However, none of the contemporary cationic (H, K, R) nor the aromatic (W, Y, F) amino acids were detected. By contrast, many other non-proteinogenic amino acids were identified [32]. More elaborate syntheses such as the Strecker reaction and cyano sulphide base reaction demonstrated the abiotic formation of other amino acids not found in the spark discharge experiment. However, the connection of these syntheses to actual prebiotic conditions is yet to be established [33,34]. Similarly, a simulated chemical environment of oceanic hydrothermal vents was shown to be permissive for early amino acid formation under high pressure and temperature conditions [35,36]. In addition to endogenous sources, meteorite analyses consistently show a variety of non-biogenic and biogenic amino acids (G, S, A, T, D, V, E, L, I, F, and Y) [37,38]. These findings led to the two independent meta-analyses summarising numerous reports on prebiotically plausible amino acid sets from endogenous and exogenous sources. These reports agree on an early amino acid alphabet consisting of approximately 10 members – G, A, D, E, V, I, L, P, T, and S [39,40]. Aromatic amino acids were among the latest arrivals into the protein amino acid alphabet (**Table 1**) [39,40]. Interestingly, a combination of quantum chemical computation and biochemical experiments allowed Granold *et al.* to hypothesise that these late arrivals were introduced consequently with early oxygenation of the atmosphere to protect cells against destruction by oxygen free radicals [41].

Analyses of the contemporary protein world also provide some clues on the early alphabet evolution. Gulik *et al.* analysed PDB data and showed enrichment of prebiotic amino acids A, G, D and V in the most conserved parts of the enzymes – active centres [42]. Sobolevsky and Trifonov performed *in silico* translation of prokaryotic genomes and analysed most conserved short octapeptide sequences that allowed to derive a temporal order of amino acids, with the most conserved ones being the earliest [43].

Table 1. Temporal order of amino acid introduction according to Trifonov and Higgs&Pudritz [39,40]. Both studies derive a relative ages of amino acid introduction based on numerous criteria

amino acid	Trifonov relative age	rank	Higgs & Pudritz relative age	rank
G	3.5	1	1.1	1
A	4	2	2.8	2
D	6	3	4.3	3
V	6.3	4	8.5	5
P	7.3	5	10	9
S	7.6	6	8.6	6
E	8.1	7	6.8	4
T	9.4	8	11.7	10
L	9.9	9	9.4	8
R	11	10	13.3	13
N	11.3	11	14.2	14
I	11.4	12	9.1	7
Q	11.4	13	14.2	14
H	13	14	13.3	13
K	13.3	15	12.6	11
C	13.8	16	14.2	14
F	14.2	17	13.2	12
Y	15.2	18	14.2	14
M	15.4	19	14.2	14
W	16.5	20	14.2	14

From the opposite side of the evolutionary timeline, the appearance of late coming amino acids tryptophan and cysteine were approached in two studies [44,45]. In the former study, Fournier *et al.* have shown that tryptophan was introduced into the modern amino acid alphabet only after tyrosine, following the tyrosyl-tRNA synthetase divergence. This evolutionary fork allowed tryptophan to be incorporated into proteins through newly developed biosynthetic pathways [44]. In the latter study, Fujishima *et al.* demonstrated that the amino acid repertoire could be extended through biosynthetic pathways using enzymes composed of an alphabet lacking the newly synthesised amino acid. The cysteine metabolic pathway was successfully engineered to absent both sulphur-containing amino acids – cysteine and methionine [45].

The question if the early amino acid alphabet exhibits sufficient structure and function forming potential was investigated both computationally and experimentally. Virtual mutagenesis of modern protein sequences toward the early amino acid alphabet showed a surprising degree of structural information conservation [46]. Furthermore, the Akanuma group's experimental studies suggested that stable protein folds can be constructed from the prebiotically plausible 10 amino acid alphabet [47]. However, the function of such proteins was still dependent on the presence of the late amino acids. A recent study by Longo *et al.* showed that an ancestrally reconstructed peptide sequence derived from P-loop NTPase and assembled from prebiotically available amino acids maintains both structure and function. In addition, the reconstructed peptide exhibited coacervation in the presence of RNA [30].

In summary, the development of the protein alphabet and the emergence of the first proteins is still veiled by uncertainties brought by the constant works of evolution. Although novel proteins do emerge consistently from the genome's non-coding parts, their amino acid composition is given by modern genetic code architecture [48]. Contemporary experimental methods for "reverse evolution" of proteins towards the early alphabet are limited by already optimised scaffolds which do not necessarily represent the earliest structure precursors. While computational approaches such as ancestral sequence reconstruction could shed light on the evolutionary paths, the sequences beyond the last universal common ancestor (LUCA) are unreachable in this way [49]. In addition to these top-down methods, the bottom-up approach relying on creating new proteins within the defined amino acid repertoire, lacking any similarity to conserved natural sequences, could bring a desired insight into the early protein development. Here, as in many other fields, Richard Feynman's famous quote "What I cannot create, I do not understand." remains relevant and inspiring.

4. Cofactors and protein evolution

While a considerable amount of protein functions rely solely on amino acid chemistry, at least 30 % of all proteins in living cells operate through organic or metal-based cofactors. Four decades ago, it was proposed that some nucleotide-based cofactors (e.g., ATP, NAD(P)H, and coenzyme A) may represent remnants of ancient ribozymes [50]. This hypothesis is supported by the observation that some of the most ancient protein folds are related to the nucleotide-binding cofactors, i.e., P-loop NTPases bind ATP and GTP, ABC transporters bind ATP and Rossman fold proteins are able to bind NAD(P) and FAD [51–53]. Goldman *et al.* analysed previously reported 286 enzyme families dated back to LUCA and derived 10 ancient enzyme groups based on previously derived omnipresent sequence and structural motifs in all kingdoms of life. All of these enzyme families represented cofactor binding (specifically metal binding) proteins [43,54].

In addition to their role in catalysis, cofactors were shown to facilitate protein folding, illustrated in an example of ATP and NAD binding to glyceraldehyde-3-phosphate dehydrogenase [55,56]. From a thermodynamic point of view, free energy released by cofactor/protein binding is comparable to the free energies of protein folding (~10-15 kcal/mol *vs* 10-20 kcal/mol, respectively) [57]. Thus, it was hypothesised that protein function and its conformation could have been selected by ligand binding from the early pool of disordered proteins [26]. Furthermore, it was recently demonstrated that ATP promotes the peptide/DNA complex coacervation (or liquid-liquid phase separation), foreshadowing one of the potential mechanisms of the early compartments formation and global biological system organisation [58].

Besides the organic cofactors, transition and alkaline earth metals probably played a vital role in essential prebiotic processes. The most common metal cofactors of contemporary proteins are ions of zinc (Zn^{2+}), iron ($Fe^{3+/2+}$), manganese (Mn^{2+}), copper (Cu^{2+}), magnesium (Mg^{2+}), and calcium (Ca^{2+}). According to geological records, $Fe^{2+/3+}$ alongside Mn^{2+}, Co^{2+}, and Ni^{2+} was abundant in soluble forms in an anoxic archean ocean and might have been available to the early proteins [59,60]. The anoxic-to-oxic ocean transition triggered by the emergence of photosynthesis (Great Oxidation Event – GOE) decreased the bioavailability of soluble Mn^{2+}, $Fe^{2+/3+}$ and Co^{2+}, and introduced soluble forms of Cu^{2+} and Zn^{2+} into the environment (**Table 2**) [61,62]. These geochemical transitions are traceable in modern proteins. Dupont *et al.* analysed the metal cofactor usage (Mn-, Fe-, Zn-, and Co-binding metallomes) in all three kingdoms of life and suggested that prokarya and eukarya evolved in anoxic and oxic environments, respectively. Indeed, the earliest Mn, Fe, Zn, and Cu binding proteins appear to be the 14th, 74th, 105th and 164th in protein structural chronology established by Caetano-Annoles and coworkers [18,63,64].

The central role of Fe^{2+} in early protein evolution is supported by the recent study suggesting ferredoxin FeS-binding fold be the most ancient protein structure along with the Rossman fold [65]. Interestingly, it was suggested that Mg^{2+} in an oxic environment could play a similar role to the Fe^{2+} before the GOE [66]. The authors tested this hypothesis by replacing Mg^{2+} with Fe^{2+} in an anoxic environment and demonstrated that three enzymes under investigation not only retained their activity but that Fe^{2+} could be an even more effective cofactor than Mg^{2+} [66]. Similarly, a cofactor replacement experiment was performed by Bray *et al.*, investigating the function of the ribosome. Authors demonstrated that under anoxic conditions, Mg^{2+} could be replaced by Fe^{2+} and Mn^{2+} without impairment of ribosomal synthetic activity [67]. Apart from participation in protein functions, Mg^{2+}, Ca^{2+}, and polyamines are efficient chemical chaperones and promote folding of a 60-amino-acid-long ancestral protein [68].

All these findings suggest that cofactors have played a central role in protein function and structure evolution. Metal cofactors, being the only part of proteins that can not be biosynthesised, provide an important retrospective look at the ancient link between proteins and their environment.

Table 2. Molar concentrations of metals (M) in anoxic and oxic oceans [62]

	Anoxic (M)	**Oxic (M)**
Mg^{2+}	10^{-2}	10^{-2}
Fe	10^{-7} Fe(II)	10^{-19} Fe(III)
Mn^{2+}	10^{-6}	10^{-8}
Co^{2+}	10^{-9}	10^{-9}
Ni^{2+}	10^{-9}	10^{-9}
Cu	10^{-20} Cu(I)	10^{-10} Cu(II)
Zn^{2+}	10^{-12}	10^{-8}

5. Characteristics of unevolved protein space

The first protein sequences must have been simple, random polymers with defined amino acid composition, as discussed in chapter 3. Later, these proto sequences diverged into the modern variety of primary and tertiary protein structures. Current natural protein sequences are estimated to fold into ~2000 different topologies. This estimate raises several questions regarding protein evolution, engineering, and design. Are there other protein folds beyond the natural structure space? To what extent is protein structure space explored and how optimised are contemporary proteins? An attempt to reconstruct all possible protein folds computationally showed that all the natural topologies could be modelled by *ab initio* simulation of homopolymeric sequences. Vice versa, all modelled structures matched their natural counterparts suggesting the completeness of our knowledge of naturally occurring protein folds [69]. Nevertheless, whether other protein folds in the unnatural sequence space are possible, and the global trends of natural optimisation remain enigmatic [70].

Several computational and experimental studies approached the characterisation of unnatural sequence space, focusing either on structural and/or functional potential of random or never born proteins (NBP's). Minervini *et al.* and Prymula *et al.* analysed random libraries of canonical amino acid proteins using Rosetta tertiary structure predictor [71,72]. Their results suggested that structured molecules frequently occur in random sequence space. Moreover, the secondary and tertiary structure content was reported to be similar to the natural distribution.

From an experimental point of view, Chiarabelli *et al.* used selective proteolysis to investigate the folding propensity of 50-residue long NBP's with natural amino acid alphabet displayed on the phage library [73]. The authors demonstrated that 20 % of proteins in the library were protease-resistant, suggesting the occurrence of stable tertiary structure. In the following study, LaBean *et al.* showed that proteins with native conformation and cooperative denaturation profile could be isolated from a library of random 71-amino acid long proteins [74]. In the protein evolution-motivated inquiry of Tanaka *et al.*, the authors characterised random proteins built either of all 20 amino acids as well as prebiotically plausible amino acids [75]. Interestingly, proteins with a simplified amino acid alphabet showed higher solubility than their canonical counterparts. The study of Newton *et al.* agree with these observations and highlight that random proteins of various evolutionary relevant alphabets tend to be disordered [76]. Davidson and Sauer proceeded further in alphabet reduction and created random protein libraries with a 3-amino acid alphabet consisting of leucine, glutamine and arginine [77]. Analysis of these proteins confirmed the tertiary structure in 5 % of sequences. Other works suggest that random proteins are able to form hydrophobic cores and perform rudimentary enzymatic activities [78–80]. Activity centred investigation of Keefe and Szostak used mRNA-display to select ATP-binding proteins from a large library of random sequences [81]. The ATP binding was detected in 4 proteins in a 6×10^{12} random sequence pool. Interestingly, the structure of one ATP-binder was solved by another group and revealed a completely unknown flexible, metal-cofactor dependent protein fold (**Fig. 3**) [82].

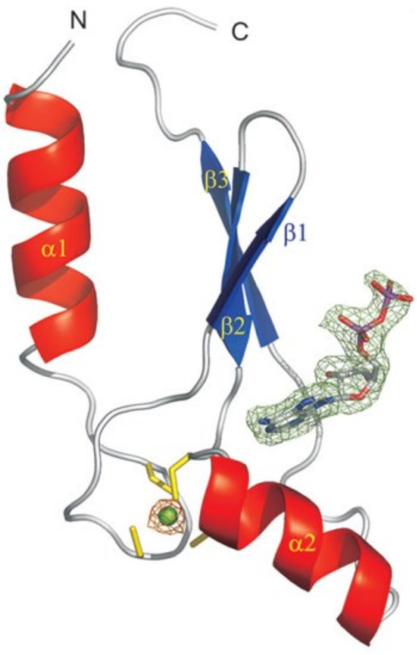


Figure 3. Structure of the novel flexible zinc-dependent ATP binding α/β protein selected by Keefe and Szostak and solved by Surdo *et. al.* [81,82]

Spontaneous appearance of function in random proteins is an attractive concept for *de novo* gene emergence – transcription and translation of a protein from a previously non-coding part of the genome. Such proteins are not homologous to any known conserved sequence; thus, the term "orphan gene" was coined for their DNA template sequences. In that context, random or pseudo-random sequences could provide an organism with an evolutionary "wild card" – Neme *et al.* showed that expression of random proteins in *E. coli* cell culture could positively and negatively affect organismal fitness [83].

Although studies of random sequences are scarce, they provide a rich source of thought-provoking topics and gedankenexperiments. Furthermore, random protein’s biological and biophysical characteristics are being revised in light of a rapidly growing field of *de novo* protein emergence.

6. Protein libraries for sequence space exploration

Traditionally, protein sequence space exploration is accomplished by experimental library construction and screening for desired functional protein or peptide variants. The most straightforward approach is an oligopeptide synthesis via solid-phase chemistry, allowing for fast, controlled and pure peptide product preparation [84]. Unfortunately, synthesis costs are high, proteins need to be refolded, and the length of chemically synthesised products is currently limited to ~120 residues [85]. Nevertheless, solid-phase synthesis is the method of choice for short peptide and peptide library preparation, especially when unnatural amino acid incorporation is required. Alternatively, a biochemical method for protein library preparation is through DNA template modification by error-prone PCR [86]. The principle of the method relies on the amplification of protein-coding DNA via error-prone DNA polymerase. The DNA polymerase introduces 1-5×10^3 mutations/base in each amplification cycle resulting in a DNA template library with randomly mutated positions. The method is cost-effective and robust. However, the degree of sequence space exploration is low, with the possibility of introducing a specific kind of mutation (PCR-bias) [87]. Efficient sequence space exploration can be achieved with a direct DNA template modification. By reason, the simplest way to design a protein library on a DNA level is to synthesise combinatorial templates chemically from the amino acid coding triplets – codons. Triplet synthesis is the golden standard for a combinatorial library preparation – it brings complete control over the mutagenised positions and frequencies of amino acids on a given position [88]. However, the method is rarely utilised because of its immense costs and reduced stability of protected trinucleotide precursors. An alternative approach utilises combinatorial nucleotide synthesis. An equimolar nucleotide mixture is added during the coupling step of synthesis, which results in a degenerate position on the synthesised DNA strand (**Table 3.**). This synthetic degeneracy allows designing a specific pattern of three degenerated positions – degenerate codon.

Table 3. IUPAC nomenclature for degenerate nucleotides

symbol	bases	symbol	bases
A	A	**K**	G/T
C	C	**S**	C/G
G	G	**W**	A/T
T	T	**H**	A/C/T
I	I	**B**	C/G/T
R	A/G	**V**	A/C/G
Y	C/T	**D**	A/G/T
M	A/C	**N**	A/C/G/T

Considering 4-letter DNA code and equimolar ratios of synthesised nucleotides, the total number of all possible degenerate codons is $(2^4 – 1)^3 = 3375$ (where 2^4 represent all A, C, T, G combinations and -1 stands for the empty combination set). A typical example: degenerate codon NNN where all three positions are randomised with all four nucleotide types allows to encode all 64 codons. However, if such codon is included in a degenerate position of DNA strand, frequencies of amino acid occurrence will be driven by the architecture of genetic code and include mostly unwanted STOP codons. Alternatively, codon NNK (G or T in a third position) can encode all 20 amino acids using only 32 different codon combinations with fewer STOP codon occurrences. Nevertheless, the frequency of amino acid occurrence in NNK based mutagenesis is still uneven, with a higher incidence of amino acids coded by multiple codons (e.g. leucine, serine, and arginine). Degenerate codons can be combined to embed a specific amino acid distribution in a mutated position. In that scheme, several templates with degenerate codons are synthesised and mixed in appropriate ratios in one tube. An example of such a method is the "22c trick", which combines three or four degenerate codons and achieves uniform coding of all 20 amino acids using 22 or 20 codons, respectively [89]. If even more precise control over the amino acid distribution is required, one can utilise spiked codons – triplets consisting of degenerate nucleotides with non-equimolar nucleotide ratios [90]. This modification allows for fine-tuning the amino acid frequencies coded by traditional degenerate codons and potentially cover a sufficient protein space at lower costs (fewer templates to combine in one tube). Unfortunately, the rational design of spiked codons is non-trivial, and design software is not publicly available.

Overall, degenerate codon synthesis offers a cost-efficient and versatile solution for site-specific library construction. The selection of 3375 different degenerate codons, their spiked variants and their combinations provide sufficient space for amino acid distribution approximation. The method is widely used in protein engineering studies. Although it offers satisfactory results for small targeted libraries, it does not provide a solution for a combinatorial design of diverse random protein libraries with precise amino acid distributions – an essential instrument of random protein space exploration.

7. Algorithms for degenerate libraries design

Traditionally, algorithms for degenerate codon library design aim to construct a compact DNA template with the lowest codon diversity. This allows for efficient protein library screening as the molecular diversity for *in vitro* screening is limited by experimental constraints (in most cases, reaching 10^{12}–10^{14} different species). These constraints dictate the choice of degenerate codons with the highest degeneracy (more amino acids per degenerate codon) and the lowest amino acid redundancy (fewer codons to code each amino acid). Pines *et al.* designed a codon compression algorithm that provides an optimal codon combination according to a user-defined amino acid alphabet [91]. The algorithm is implemented in two versions – the first one is suitable for codon compression with the lowest amino acid redundancy, and the second mode is optimised for maximum codon degeneracy. These approaches have their distinct advantages in optimising screening efficiency or cost-effectiveness, respectively, and both are implemented as web applications [92]. Parker *et al.* built an alignment-based codon optimiser using a linear programming algorithm to automatise the library design process [93]. The algorithm analyses all positions from the input multiple sequence alignment and outputs the most suitable degenerate codon for each variable position respecting their mutual co-variation. Similarly, web implemented algorithm SwiftLib uses dynamic programming to identify optimal codon choice for a given amino acid alphabet [94]. SwiftLib can fit amino acid distributions via multiple degenerate codon combinations; however, it does not employ mutual residue information from multiple sequence alignments.

A further step in degenerate codon library design is the inclusion of codons with variable nucleotide ratios (spiked codons). Spiked codons allow for fine-tuning intrinsic amino acid distribution of a canonical degenerate codon. Wolf *et al.* and Craig *et al.* designed spiked codon optimisation schemes based on numerical optimisation and genetic algorithms [95,96]. Both algorithms are non-deterministic and provide only one of the several possible solutions for the amino acid distribution. Unfortunately, neither of these algorithms is implemented in a publicly available form.

The task of degenerate codon optimisation was shown to be mathematically straightforward. Finding the most fitting codon for a given distribution can be solved precisely by optimal methods (e.g., dynamic programming or linear integer programming). However, the task becomes increasingly complex when combinations of all 3375 degenerate codons come into consideration. In this work, we have designed a heuristic algorithm for the diverse library construction composed of several types of degenerate codons that provide protein templates with a specific amino acid composition (**Fig. 4**).

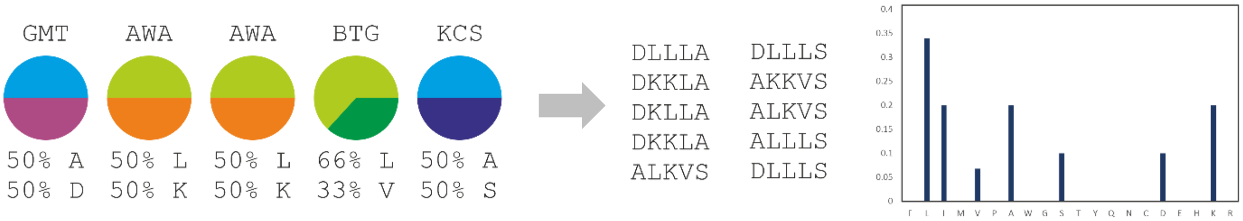


Figure 4. (left) degenerate codon string encodes various peptide sequences (right), the amino acid distribution of the whole peptide set is given by identity of degenerate codons. In a case of a short sequences, degenerate codons describe the distribution of the whole protein library rather than amino acid distribution of a single peptide. With extension of coding template, amino acid distribution of each protein approaches the global amino acid distribution of the whole protein library

8. Chaperones from an evolutionary perspective

Protein evolution is the process of sequence adjustment which under selection pressure grants the selected individual the highest reproduction or survival advantage in a given environment. These sequence/structure adjustments can perturb the complex pathway of protein folding and lead to spontaneous aggregation and organismal dysfunction before reaching the optimal phenotype [97]. Thus, from the evolutionary perspective, aggregation suppression is imperative. Indeed, numerous studies have shown that minimisation of aggregation propensity is one of the specific evolutionary optimisation outcomes [98–103]. It was pointed out that most essential cellular proteins have the lowest aggregation propensities [101]. Foy *et al.* analysed protein families differing in their degrees of evolutionary optimisation (e.g., evolutionary age). They demonstrated that although both old and young protein families possess a hydrophobic core, younger proteins tend to avoid aggregation by increased dispersion of hydrophobic residues throughout their sequences [103]. In contrast, evolutionary older proteins show frequent hydrophobic amino acid clusters, suggesting carefully optimised sequence parameters to achieve independent and stable folding [103]. In fact, the tendency of proteins to aggregate strongly correlates with global biological properties such as cellular protein abundance and subcellular localisation [102].

One of Nature's evolved mechanisms of aggregation prevention is protein-dependent folding assistance. The proteins which facilitate the folding process are known as chaperones, and they can be divided into three functional classes – chaperonins, assisting enzymes and stabilisers [104].

(i) Chaperonins are large protein complexes that recognise unfolded or misfolded proteins via solvent-exposed hydrophobic residues. They act on already translated polypeptide chains in the cytoplasm and utilise ATP hydrolysis to either unfold and release the protein or encapsulate it into the molecular cage and promote the refolding through conformational changes of the polar environment of the cage.

(ii) Assisting enzymes include proteins with very specific functions such as protein disulfide isomerase or peptidyl prolyl *cis-trans* isomerase, which promote the formation and reorganisation of disulfide bridges and catalyse *cis-trans* shift of proline peptide bond in the protein.

(iii) Stabilizers bind to the nascent protein chains via their exposed hydrophobic residues, prevent translated proteins from aggregation and serve as a hub to the downstream folding machinery. Examples of such proteins are heat shock proteins (HSPs) which can function either in ATP independent or dependent ways.

In this work, we have exploited the stabilising activity of ATP-dependent Hsp70 protein (known as DnaK in *E. coli*) which cooperates with its co-chaperone DnaJ and nucleotide-exchange factor GrpE (**Fig. 5**) [105,106]. Although DnaK and its co-chaperones are not essential for *E. coli,* the DnaK lacking mutants can not withstand elevated temperatures due to overwhelming protein misfolding [107]. DnaK protein is composed of nucleotide-binding (NBD) and substrate-binding (SBD) domains with a total molecular weight of 60 kDa. NBD domain contains a MgATP/MgADP binding site and exhibits ATPase activity [108]. The SBD domain of DnaK displays an unspecific binding preference towards mis/unfolded and partially folded proteins with exposed hydrophobic residues [109]. The substrate recognition begins with SBD binding to a stretch of ~5–7 hydrophobic amino acids, which subsequently induces a conformational change in SBD. The affinity to the substrate is regulated by ATP – when bound to NBD, the chaperone persists in a low-affinity conformation known as an open state. When ATP is hydrolysed to ADP, the affinity of SBD is increased and leads to the closed conformation [110]. This shuffling between the conformations is arranged by co-chaperones DnaJ and GrpE. While DnaJ acts as a substrate recognition protein and closed conformation inducer, GrpE releases NBD-bound ADP and recycles the chaperone. The mechanism of refolding consists in kinetic restriction of the substrate folding, allowing the molecule to fold its distinct parts independently [111].

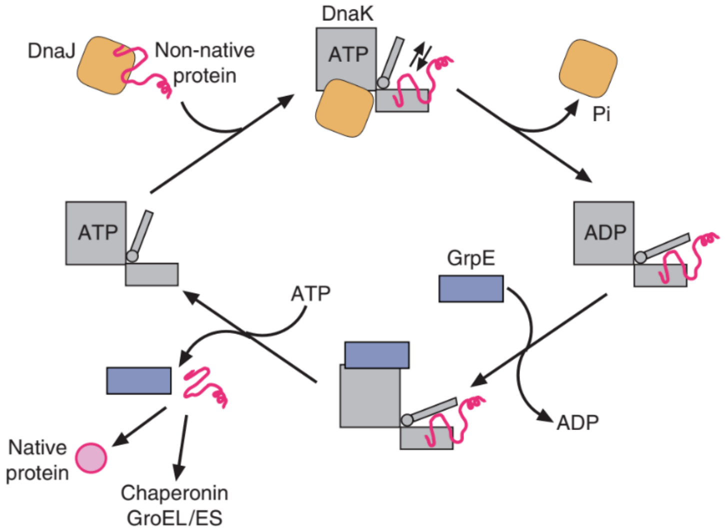


Figure 5. DnaK/DnaJ/GrpE chaperone cycle. DnaJ-bound unfolded protein interacts with SBD of DnaK which shifts to the closed state via ATP hydrolysis. Subsequently GrpE releases bound ADP and is exchanged by an ATP which causes the release of protein in native conformation or towards downstream processing [111]

The indispensability of chaperones for contemporary organisms is marked by their abundance amounting to ~0.3 % of all genes in the genomes [112]. Besides the evident function of chaperones in housekeeping and stress-induced folding assistance, they can serve as evolutionary capacitors facilitating otherwise impossible evolutionary trajectories. It was demonstrated that overexpression of either DnaK or GroEL/GroES chaperone system in *E. coli* leads to survival and adaptation of the cells with impaired essential functions [113,114]. These results indicate that protein solubility/stability is a major evolutionary constraint that can be buffered by a chaperone system. Furthermore, Aguilar-Rodriguez *et al.* studied evolutionary rates of DnaK chaperone clients and revealed a direct correlation between interaction levels with chaperones and the rate of protein evolution [114]. The study of Kadibalban *et al.* confirmed this correlation and particularised that the most potent chaperon binders evolve 4.3× faster than the least potent chaperon-binding proteins [115]. In a more recent study, Alverez-Ponce *et al.* analysed the effect of chaperone dependence on evolutionary divergence of proteins and confirmed that the genes encoding chaperone clients had diverged faster than genes encoding non-client proteins [116].

The connection between chaperones, natural protein world and alphabet evolution was pointed out by Houben *et al.* They studied the effects of basic and acidic amino acids on protein structures and showed that although positively charged residues are more compatible with folded conformation, they also tend to drive protein into the aggregated state [100]. The authors hypothesised that chaperones co-evolved with the introduction of basic amino acids into the genetic code and allowed the expansion of the structural variety of the protein universe. This conclusion agrees with the evolutionary analysis of proteome expansion across the tree of life performed by Rebeaud *et al.* [112]. The authors showed that from the simplest archaea to eukaryotes, the total number of proteins per proteome expanded 200×, proteins became larger, and aggregation-prone proteins became 6× more frequent. They proposed that the proteome expansion network was supported by higher abundances of intracellular chaperones and their interconnection into the misfolding prevention network.

In summary, evolution carefully selected proteins able to perform specific functions and reach compact conformations. The role of chaperones in protein folding is similar to that of DNA-repair enzymes in replication – both maintain the functionality of a precisely evolved state to guarantee the smooth transition of biological information to the next generations. Moreover, chaperones play an indispensable role in allowing evolution to experiment with potentially beneficial protein phenotypes, which would not be achievable without their assistance.

9. Dephospho Coenzyme A Kinase

As was described in chapter 3., the amino acid alphabet appears to evolve concurrently with the development of proteins. While the latest additions to the protein amino acid alphabet are considered essential for the fold stabilisation, the earliest, prebiotically available amino acids are among the strongest promoters of protein disorder [117]. Here, on the example of enzyme dephospho coenzyme A kinase (DPCK), we investigate the effect of complete aromatic amino acid replacement with their prebiotically plausible counterparts. Dephospho coenzyme A kinase (ATP:dephospho-CoA 3'-phosphotransferase) catalyses the final step in coenzyme A biosynthesis – γ-phosphate group transfer from ATP to 3-hydroxyl ribose moiety of dephospho CoA (**Fig. 6**) [118].

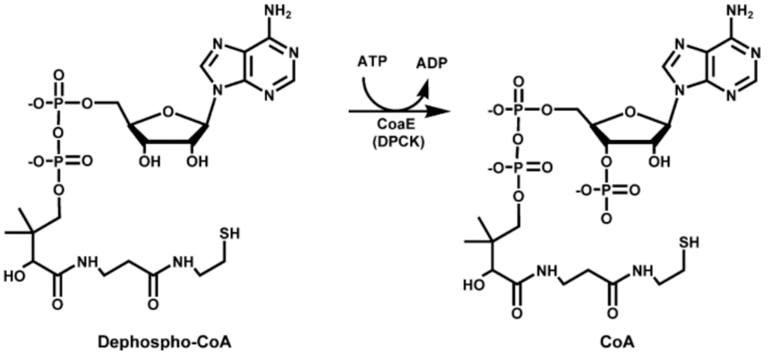


Figure 6. Reaction catalyzed by DPCK [118]

DPCK is a well folded and conserved protein with high α-helical content; it is divided into three functional parts – CoA binding domain, ATP binding domain and lid domain (**Fig. 7**)

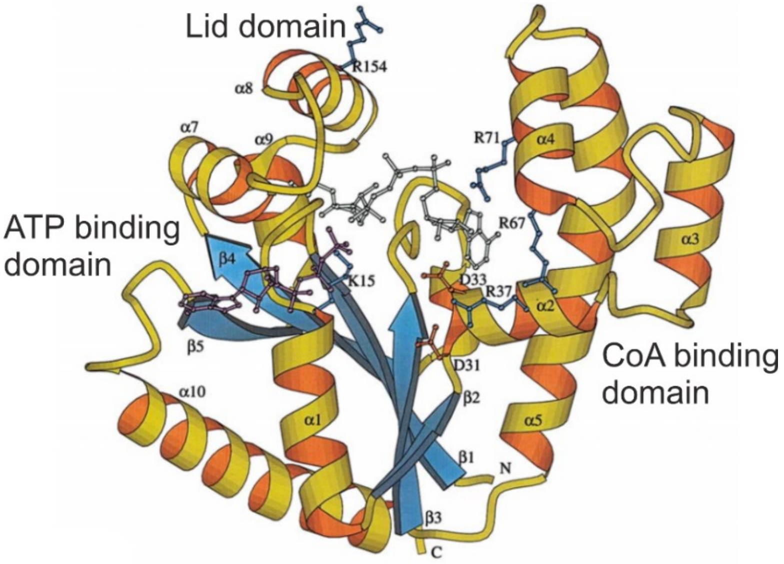


Figure 7. Structure of DPCK with highlighted lid, ATP, and CoA domains and both bound substrates. Illustration adapted from [119]

While ATP and CoA domains bind and activate their substrates by hydrophobic interaction and an extensive network of hydrogen bonds, the lid domain protects the active site from water molecules, which would hydrolyse the phosphoanhydride bond of the activated ATP. The γ-phosphate transfer from ATP to dephospho CoA requires a significant conformational change of the mobile CoA and lid domains; hence the enzyme exists in an open and closed conformation in substrate unbound and bound states, respectively [119].

Both substrates of the enzyme – ATP and CoA are among the essential nucleotide-derived cofactors of life (as was summarised chapter 4). Coenzyme A and its derivatives are indispensable for ~4 % of all cellular enzymes and engage in a diverse repertoire of biochemical pathways. On the protein side – the ATP binding domain is an example of an ancient P-loop NTPase fold which comprises ~10-18 % of the predicted gene products in both prokaryotic and eukaryotic genomes [120]. The P-loop NTPases appear to have emerged in a very early stage of protein evolution and were already present in the last universal common ancestor [121,122]. Moreover, recent studies suggest that the P-loop NTPase and Rossman fold might be the most ancient phosphor-binding proteins and carriers of the earliest protein function [30].

Here we study the impact of evolutionary youngest amino acids removal from the contemporary protein structure. We investigate the structural properties of the aromatics-less DPCK mutant via enzyme kinetics, circular dichroism spectroscopy, nuclear magnetic resonance, dynamic light scattering, and limited proteolysis.

10. Selective proteolysis as a tool for protein conformation screening

Selective proteolysis is a simple and low-resolution technique for protein conformation assessment. The principle of the method relies on a differing sensitivity of compact and folded parts of a protein in comparison to its unfolded chains. While unfolded proteins are degraded rapidly, proteolysis of folded molecules shows slower kinetics. Nevertheless, even folded proteins can be hydrolysed by proteases by initial nicking of locally unfolded part and subsequent proteolysis of disordered fragments (**Fig. 8**) (reviewed in [123]).

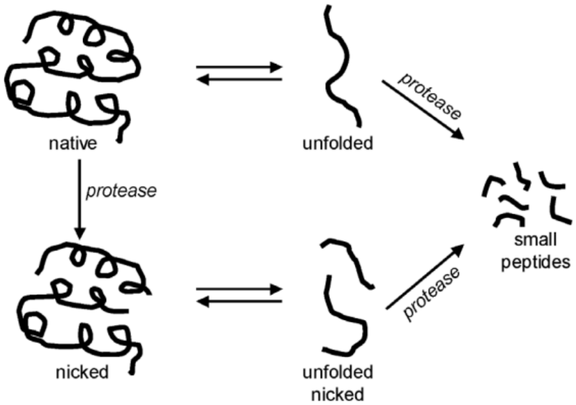


Figure 8. Proteolysis of a folded protein can be followed by global protein unfolding or by initial nicking of flexible protein parts and subsequent degradation of resulting disordered fragments [123]

Limited proteolysis can be carried out by proteases with broad specificity as well as specific enzymes. The former allows analysing the global conformational properties of the studied protein substrate. If the amino acid sequence of the substrate protein is known, broad specificity proteolysis followed by the mass spectrometric detection of proteolytic fragments can provide insight on the sites of increased protein flexibility [124]. By contrast, site-specific proteolysis offers insight into either naturally occurring or engineered regions of a protein substrate. Here we utilised both approaches to characterise global conformational states of random proteins as well as the local protease accessibility to engineered cleavage site in the centre of a random protein sequence. We exploited the broad specificity of *E. coli* Lon protease and site-specificity of bovine thrombin.

a. Thrombin

Thrombin is the key protease of the blood coagulation cascade in the vertebrates. This enzyme belongs to the serine protease family along with trypsin, with which it shares ~50 % sequence similarity. The mechanism of serine proteases depends on the catalytic triad of histidine, aspartate, and serine. Both thrombin and trypsin preferentially cleave peptide bonds following basic amino acid residue

Unlike trypsin, thrombin exerts significant specificity, as was demonstrated on variant oligopeptide substrates [125–129]. These investigations showed that thrombin exhibits a preference for an aliphatic residue at the P4 position, for Pro at P2, Arg at P1 and for a basic residue at P3’. Furthermore, acidic residues are unfavored at all positions from P3 to P3’ (**Fig. 9**).

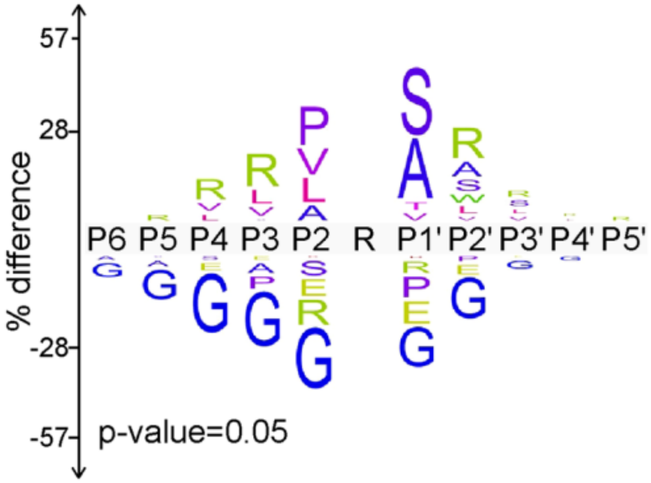


Figure 9. Substrate specificity of thrombin with preferred (positive % difference) and undesirable (negative % difference) residues on position P6 to P5‘ of substrate sequence [129]

Here, in order to investigate the collective folding/aggregation behaviour of random proteins, we engineered the DNA sequence of the thrombin-specific cleavage site into the central part of the library template. The rationale for the experiment was adopted from the study of Chiarabelli *et al.*, where a similar specific sequence was engineered into the phage-displayed random sequence library [73]. Authors of that study subjected phage displayed library to proteolysis, collected the resistant proteins via affinity chromatography and examined the genotypes of the resistant variants by high throughput sequencing of phage embedded DNA.

b. Lon protease

Lon protease is one of the central regulatory proteases in bacterial cells. In *E. coli*, it participates in processes of bacterial communication, biofilm formation and general stress response (heat, acidic, nutritional or pathogen-induced, reviewed in [130]). Since the protease is involved in ~50 % of total abnormal protein degradation in *E. coli*, it was proposed and experimentally demonstrated that Lon broadly recognises exposed patches of hydrophobic residues that are normally buried in properly folded proteins (reviewed in [131]). Interestingly, protein-bound ligands were shown to protect the polypeptides against Lon degradation.

Functionally, the enzyme belongs to the family of AAA+ proteins (ATPases Associated with a variety of cellular Activities). It is composed of ATP binding and proteolytic domains situated on one polypeptide chain. Additionally, as shown by limited proteolysis studies, bacterial Lon proteases possess a large N-terminal domain believed to be responsible for substrate recognition [132]. In *E. coli*, the active form of Lon protease is a homohexameric ring protein with a central cavity known as the proteolytic chamber. The hydrolysis is mediated through a unique Ser-Lys dyad, unrelated to the serine protease His-Asp-Ser catalytic triad. Since the specificity of Lon protease is broad and takes place in a confined proteolytic chamber, it is hypothesised that the substrate is degraded by rounds of substrate binding, cleavage, release and rebinding [133].

Here we exploited the broad specificity of the Lon protease against exposed hydrophobic residues to assess the folding state of random library proteins. The approach was adopted from a recently published methodology on cell-free based protein folding assay [134]. In comparison with the limited proteolysis of purified proteins, proteolysis by Lon allows to evaluate the protein folding propensity directly during the *in vitro* translation and to assess the immediate dynamics of protein folding.

AIMS OF THE THESIS

The overall aims of this work were to (i) investigate properties of random protein space and relation of random protein sequences to the natural proteins and (ii) study the effect of amino acid alphabet on protein structure and function.

The specific goals were:

- To analyze structure occurrence in random protein sequence pool and characterize selected random proteins *in vitro*.

- To build a computational tool for degenerate protein library construction capable to design a diverse library of random proteins with specified amino acid occurrencies.

- To experimentally characterize libraries of random proteins with different amino acid alphabets.

- To investigate the effect of latest amino acids on a selected protein structure and function.

METHODS

The research papers included in this Ph.D. thesis provide a detailed description of methods and experimental procedures used together with details necessary for the reproduction of the presented results. Therefore, this chapter only lists experimental techniques used throughout the thesis and selection of methods concerning the presented unpublished data.

List of used research methods:

- Recombinant expression in *E. coli*
- Protein purification via affinity, ion exchange and gel chromatography
- *In vitro* transcription
- Cell free protein synthesis
- Electronic circular dichroism
- Dynamic light scattering
- Nuclear magnetic resonance
- Amino acid analysis
- High throughput sequencing
- Bioinformatic analysis of protein secondary structure, solubility and aggregation

Selection of methods supporting the unpublished material

1. Preparation of template DNA library

Synthetic DNA oligonucleotides of 197 (Oligonucleotide 1) and 198 (Oligonucleotide 2) bases in length were ordered from Integrated DNA technologies (IDT). The oligonucleotides were dissolved in RNAse free water to a final concentration 100 µM. The annealing reaction was mixed as follows:

Component	**Volume (µl)**
NEB buffer 2 (New England Biolabs)	5
Oligonucleotide 1 (100 µM)	1
Oligonucleotide 2 (100 µM)	1
dNTPs (10 mM)	2
H_2O	39

The mixture was incubated at 95 °C for 5 minutes in a thermocycler and cooled down by turning the closed appliance off. After 30 minutes, 2 µl of Klenow polymerase (10 units, NEB) was added and the reaction was incubated for 1 hour at 25 °C. Subsequently, Klenow polymerase was inactivated by 5-minute incubation at 85 °C and the annealed DNA template was purified using the DNA Clean&Concentrator Kit-25 (Zymo Research) according to the manufacturer inctructions.

2. *In vitro* transcription of DNA library template and mRNA purification

The annealed and purified DNA template was transcribed into mRNA using the HiScribe™ T7 Quick High Yield RNA Synthesis Kit (NEB) according to the manufacturer's recommendations; 1 µg of the template DNA was added to one 20 µl reaction. Transcription underwent for 2 hours at 37 °C. Subsequently, one reaction volume of 5 M ammonium acetate (Sigma-Aldrich) was added into the mixture, the reaction was mixed and left on ice for 15 minutes. The formed RNA precipitate was collected by centrifugation at 21 000 ×g at 4 °C for 30 minutes, pellet was washed twice with 70 % ethanol and let dry on air. Dried RNA precipitate was dissolved in RNAse free water to the final concentration of 3 µg/ul.

3. Agarose gel electrophoresis of nucleic acids

Synthesized ssDNA, annealed dsDNA and *in vitro* transcribed RNA oligonucleotides were analyzed for its molecular size and homogeneity by electrophoresis in 1 % (w/w) agarose. The agarose gel was prepared as follows: 0.2 g of powdered agarose (Sigma-Aldrich) was dissolved in 20 ml of TBE buffer by boiling in microwave oven. Solution was cooled down to 50 °C and 2 µl of 10000× GelRed® Nucleic Acid Gel Stain (Biotium) was added. The liquid was poured into the gel holder, provided with a plastic comb, and let cool until solid. DNA samples were resuspended with ¹/₅ volume of 6×DNA Sample buffer, RNA samples were resuspended with ½ volume of 2× RNA Loading Dye (NEB) and boiled for 5 minutes at 95 °C. Samples were put into the wells and electrophoresed for 40 minutes under constant voltage of 100 V in TBE buffer filled electrophoresis tank. Separated products were visualized under UV illumination at 340 nm.

4. Cell free protein expression

Protein libraries were expressed using PUREfrex 2.0 cell free expression kit (GeneFrontier). The mixtures including variable additives were prepared as follows:

Component	**Volume (µl)**
Solution I (buffer, NTPs and amino acids)	10
Solution II (proteins and tRNAs)	1
Solution III (ribosomes)	2
Triton X-100 (5 % in dH$_2$O)	0.2
Library mRNA (3 µg/µl)	1
Lon protease (2 µM hexamer)	1
DnaK chaperone mix	1

Reactions were incubated at 30 °C for 2 hours (unless stated otherwise) and quenched by addition of 4 µg of RNAse A (NEB) per reaction.

5. Solubility assay of the protein libraries in presence of additives

Protein libraries were expressed according to the protocol described above. In order to analyze quantity of total protein product, 1 µl of reaction mixture was diluted with 14 µl of water and denatured with 5 µl of 6× SDS-PAGE sample buffer. The rest of the reaction mixture was centrifuged at 21 000 ×g for 30 minutes at 21 °C and 1 µl of soluble content in supernatant was prepared for SDS-PAGE analysis simalarly to the total reaction sample. Both fractions were analyzed by SDS-PAGE and Western blotting.

6. Purification of the protein libraries

Protein libraries were expressed as described in paragraph 4, diluted 10× by buffer A and incubated for 12 hours with 2.5 µl of TALON affinity purification resin (Clontech) pre-equlibrated with buffer A. Subsequently, resin with the immobilized library was washed 3× by 1 ml of buffer A and library was eluted to buffer A supplemented with 10 mM EDTA equlibrated to pH 8.5.

7. Folding assay of the purified library by thrombin digestion

A library sample purified by protocol from paragraph 6 was split into 10 µl aliquots representing digested and non-digested sample. Digestion was initiated by addition of 0.2 µl (0.2 U/ml) thrombin protease (Sigma-Aldrich) and proceeded for 4 hours at 37 °C. Reaction was quenched by addition of 3 µl of 6× SDS-PAGE sample buffer and products were analyzed by SDS-PAGE and Western blotting.

8. Acrylamide gel electrophoresis of proteins in presence of sodium dodecyl sulphate

Proteins were prepared for acrylamide gel electrophoresis by boiling the sample in the presence of $^1/_5$ sample volume of 6× Sample Buffer at 95 °C for 5 minutes. In this study we utilized the commercial Novex Tris-Tricine gels (Thermo Fisher) with gradiet acrylamide concentration ranging from 10 to 20 %. Gels were unpacked and inserted into the electrophoresis apparatus according to the manufacturer instructions. Electrophoresis tank was filled with 10× diluted Tricine running buffer and electrophoresis was proceeded for 1.5 hours in constant voltage of 100 V.

9. Protein western blotting with chemiluminiscent detection

Acrylamide gels with separated protein samples were equilibrated in 50 ml of Western Blot transfer buffer for 10 minutes. PVDF membrane (0.22 µm pore size) was activated by incubation in methanol for 30 seconds and equilibrated in 50 ml of Western Blot transfer buffer for 5 minutes. The western blot apparatus was assembled according to the manufacturer recommendation, the transfer cell was filled with 1 L of Western Blot transfer buffer and proteins were electroblotted onto PVDF membrane for 1 hour under the constant voltage of 100 V. Subsequently membranes were transferred into 5 % BSA (Sigma-Aldrich) solution in PBST buffer and incubated for 1 hour at 21 °C. Membranes were briefly washed with PBST buffer and incubated for 1 hour in 3 ml of PBST buffer supplemented with Monoclonal ANTI-FLAG M2-Peroxidase (HRP) antibody (Sigma-Aldrich) diluted 10 000×. Next, membranes were washed 3× for 5 minutes with PBST, overlayed with 1 ml of Immobilon Forte Western HRP substrate (Merck-Millipore) and visualized with Amersham™ ImageQuant™ 800 biomolecular imager.

10. Expression and purification of the recombinant Lon protease

Plasmid with coding sequence of recombinant Lon protease with hexahistidine affinity tag was provided by prof. Hideki Taguchi from Tokyo University of Technology, protein was prepared following the published instructions [134]. In summary, *E. coli* BL21-(DE3) chemically competent cells were transformed with 10 ng of plasmid. Starter culture for expression was prepared by picking a colony from the plate and by incubation in 5 ml of ampicillin supplemented LB media (Sigma-Aldrich) overnight. Starter culture was subsequently added to 1 L of ampicillin supplemented 2×YT media (Sigma-Aldrich) and cells were grown at 37 °C until OD_{600} = 1 was reached. Culture was equilibrated to 25 °C and induced by addition of IPTG (Sigma-Aldrich) to the final concentration of 1 mM. Expression proceeded for 12 hours at 25 °C, cells were harvested by centrifugation at 5000 ×g for 30 minutes and resuspended in 20 ml of Lysis buffer. Cells were disrupted by sonication with alternating 5/10 seconds pulse on/off cycles at 15 W power. Lysate was centrifuged at 4 °C at 30 000 ×g for 30 minutes and supernatant was loaded directly to 2 ml of QIAgen NiNTA Superflow resing equilibrated with Lysis buffer. Lysate was incubated with resin for 1.5 hours at 4 °C and washed 3× with 25 ml of 100 mM imidazole supplemented Lysis buffer. Lon protease was eluted by 20 ml of Lysis buffer supplemented with 300 mM imidazole which was subsequently removed by 3× concentration/dilution cycles with Lysis buffer on Amicon Ultra-15 Centrifugal Unit (Merck-Millipore) with 50 kDa cut-off. Protein concentration was assessed by Bradford assay and adjusted to hexamer concentration of 2 µM. Protein was aliquoted and frozen at -80 °C.

11. High throughput sequencing data processing

Paired end Illumina reads were joined by fastq-join utility at Galaxy web server, DNA sequences were translated and statistically analyzed by a set of MatLab scripts available from Heinis lab [135,136].

List of buffers

TBE buffer	90 mM Tris base, 90 mM boric acid, 2 mM EDTA
DNA sample buffer	0.25 % Bromophenol Blue, 50 % (v/v) glycerol
SDS-PAGE sample buffer	0.375M Tris pH 6.8, 12% SDS, 60% glycerol, 0.6M DTT, 0.06% bromophenol blue
Buffer A	50 mM Tris, 100 mM NaCl, 100 mM KCl, 0.05 % (v/v) Triton X-100, pH 7.5
Western Blot transfer buffer	25 mM Tris, 192 mM glycine, 20 % methanol, pH 8.5
Tricine running buffer (10X)	1M Tris base, 1M Tricine, 1 % (w/v) SDS
Lysis buffer	20 mM HEPES, 400 mM NaCl, 20 % (v/v) glycerol, pH 7.5
PBST buffer	137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 0.1 % (v/v) Tween-20

RESULTS AND DISCUSSION

1. Scarce sampling of random sequence space

Results in this section were included in the attached ***paper I*** *- Tretyachenko V. et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Scientific reports. 2017 Nov 13;7(1):1-9.*

Contemporary proteins are the result of 4 Gy of evolutionary optimization. Our knowledge of protein structure, function and evolution heavily relies on theoretical and experimental analyses of natural proteins. However, the behaviour of proteins lacking evolutionary background and optimization remains largely unexplored. Therefore, we performed a systematic computational and experimental investigation of random proteins (never born proteins, NBP's) with canonical amino acid alphabet and studied their relevance to naturally evolved sequences.

In silico random protein library construction and analysis

We generated 4 datasets with 10 000 protein sequences of 100 amino acids in length each to investigate properties of random proteins and compare them to their natural counterparts. We used 5 secondary structure predictors, 3 protein disorder predictors and protein aggregation predictor to compare libraries of (A) random sequences with natural-like amino acid occurrences (Random), (B) fragments of natural proteins from the TOP8000 database of non-redundant structurally characterized proteins from PDB database (PDB), (C) natural protein fragments from the UniProt database (Uni) and (D) fragments of natural intrinsically disordered proteins from the disprot database (Dis) [137–139]. Predictions for the random sequences (A) were also performed with the libraries extended by a 9 amino acid tag that was later used during recombinant expression of selected random proteins to verify that the tag addition did not alter the structural properties. Random protein library was searched against known natural sequences using blastp method and only low-significant matches were found.

Prediction of secondary structure and disorder

Statistical analysis of consensual secondary structure and disorder predictions showed that the overall occurrence of secondary structure and motif distribution were comparable for the random and PDB/Uni proteins (**Fig. 10**). The overall occurrence of secondary structure was approximately 5 % lower for the random sequence library than for the PDB/Uni dataset, thus we did not identify any profound differences between them. This finding is in contrast with previous reports which revealed statistically significant differences in structure content between random and natural sequences. However, these conclusions were based on reports from a single secondary or tertiary structure predictor [72].

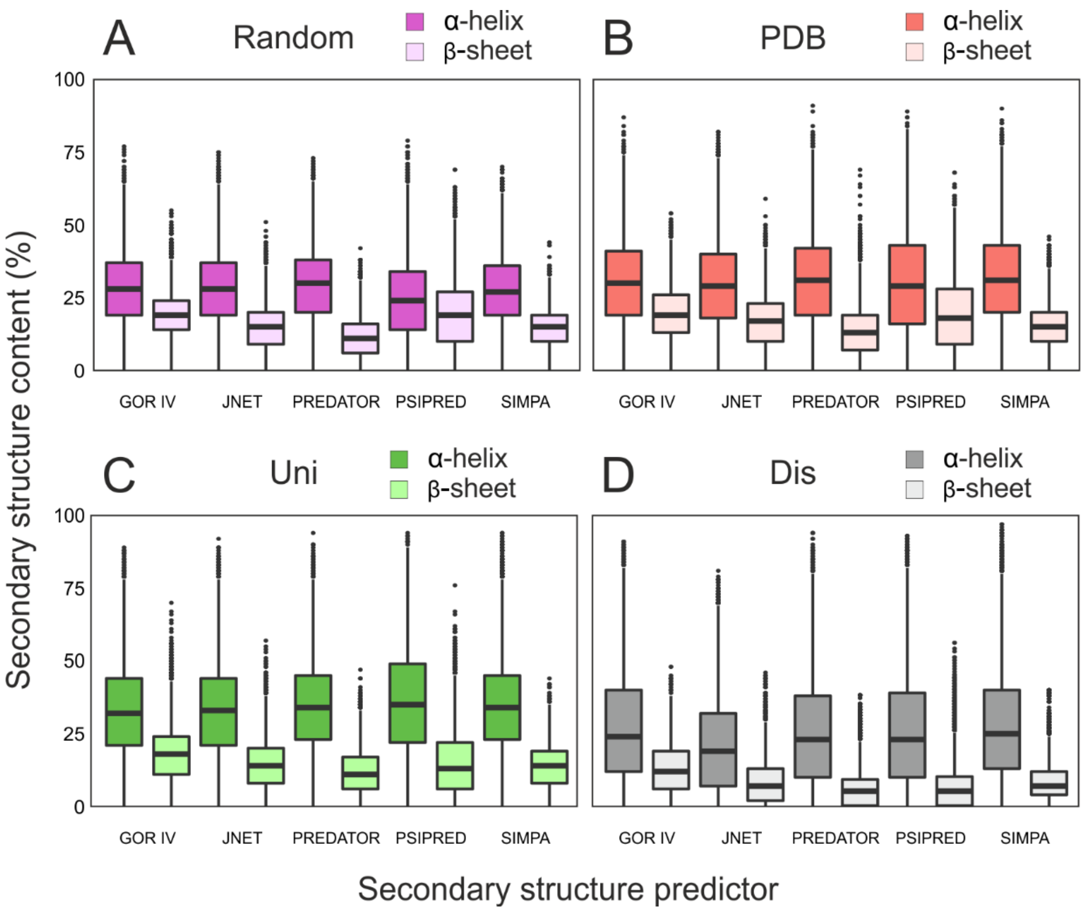


Figure 10. Predictions of secondary structure occurrence in the (A) Random, (B) PDB, (C) Uni, and (D) Dis datasets. α-helical and β-sheet content was determined by five different predictors. The center of the box represents the median; upper, and lower borders are 3rd and 1st quartile respectively. The solid lines illustrate maximum and minimum contents, which are shown as dots. The Dis dataset is included as a negative reference

Aggregation prediction analysis

Prediction of the aggregation propensity indicated that with exception of the Dis dataset, overall trends are similar in random and natural datasets. However, differences appear when aggregation predictions are correlated with the predictions of secondary structure. Each dataset was divided into three sequence subsets with (i) high structure, low disorder; (ii) average structure, average disorder, (iii) low structure, high disorder predicted contents. First, aggregation propensity in the random dataset shows positive correlation between structure content and aggregation tendency (**Fig. 11A**). This correlation is maintained in the Uni dataset, however sequences with average aggregation propensity appear to be enriched. This enrichment can be explained by the natural tendency of proteins to suppress aggregation while maintaining the structural content (**Fig. 11C**). Interestingly, protein fragments in the Uni dataset exhibit higher content of aggregation prone sequences than random dataset, the identity of these were examined by ontology analysis which showed that they belong to the membrane proteins. Similar analysis of the Dis dataset showed that disordered sequences tend to be least aggregating which can be explained by their unique amino acid composition enriched by polar and charged residues (**Fig. 11D**). The PDB dataset shows an expectable clustering of all sequences in average aggregation zone - experimentally characterized proteins tend to behave reasonably for the purposes of *in vitro* analyses suggesting that conclusions on protein sequence space derived solely from experimentally determined structures tend to be biased by the proteins itself (**Fig. 11B**).

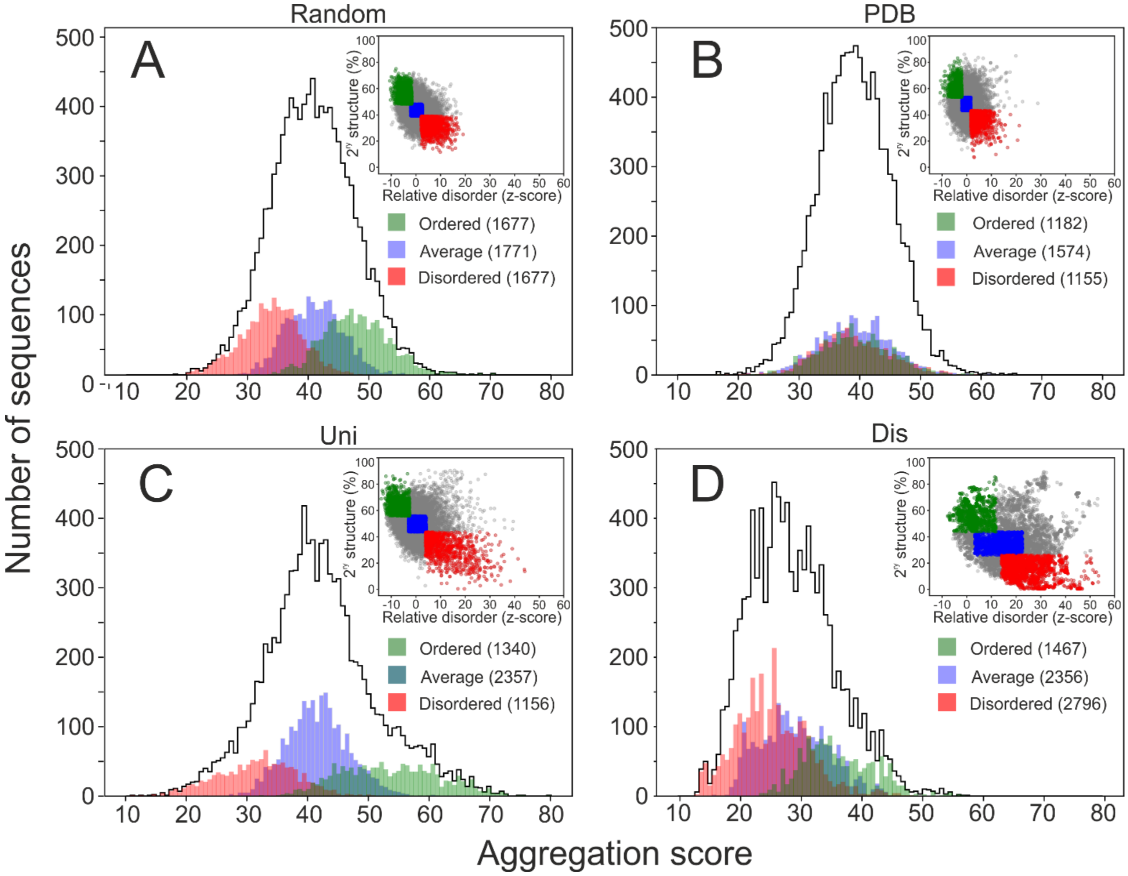


Figure 11. Aggregation propensity of the datasets depending on secondary structure analysis. Aggregation analysis was performed on the whole Random (A), PDB (B), Uni (C) and Dis (D) datasets as well as on the subsets defined by various predicted structure content. The subsets are indicated in green (high structure), blue (average structure content) and red (unstructured). Scatterplots in the top right corner of the distribution illustrate the total sequence pool for each dataset (grey dots) with the structural subsets highlighted. The values in brackets indicate total number of sequences in each subset

Biophysical characterization of random proteins

To perform detailed characterization of random proteins, we selected 45 sequences from the random dataset for subsequent expression, purification, and biophysical analysis. Selection of experimental candidates was based on secondary structure, disorder, and solubility predictions. Three groups of 15 proteins were selected: GROUP 1 - sequences with **high** predicted structure (either α-helical and β-sheet content) and solubility, GROUP 2 - sequences selected **randomly** from the whole dataset and GROUP 3 - soluble sequences with **low** occurrence of secondary structure and high disorder.

DNA sequences encoding all selected random proteins were synthesized to include N-terminal methionine and C-terminal hexa-histidine tag. Two more amino acids (Leu-Glu) were introduced into C-terminus of the resulting proteins because of the restriction cloning pipeline. Synthesized genes were expressed in *E. coli* BL21 (DE3) cells and proteins were analyzed for the expression levels and solubility. Out of 15 proteins in each group the following expressed/soluble ratios were observed: GROUP 1 - 13/4, GROUP 2 - 8/6 and GROUP 3 - 14/14. Notably, protein solubility decreased with the increasing predicted secondary structure content of the proteins (**Fig. 12, left**). In total 22 recombinant proteins were successfully produced and purified for subsequent biophysical characterization. Group 1 proteins showed pronounced ellipticity and minima between 205 and 220 nm in their electronic circular dichroism (ECD) spectra typical of proteins with high secondary structure content while Group 3 ECD spectra indicated low structural content (**Fig. 12, right**). Unlike for Group 3 proteins, presence of denaturing agent moderately decreased ellipticity of Group 1 proteins. Furthermore, addition of trifluoroethanol, a secondary structure inducer, caused induction of structure in Group 3 proteins. 1D^{1}H-NMR of Group 1 proteins further suggested the presence of a hydrophobic core and a certain degree of aggregation. The narrow and less dispersed signals of Group 3 and 2 proteins spectra are typical for IDPs.

Proteins were also subjected to dynamic light scattering (DLS) to reveal the aggregation tendencies of random proteins in different groups. Determined hydrodynamic radii correlated with the aggregation propensity predictions supporting our *in silico* observation that random proteins with high structural content tend to aggregate more than their unstructured counterparts (**Fig. 13**).

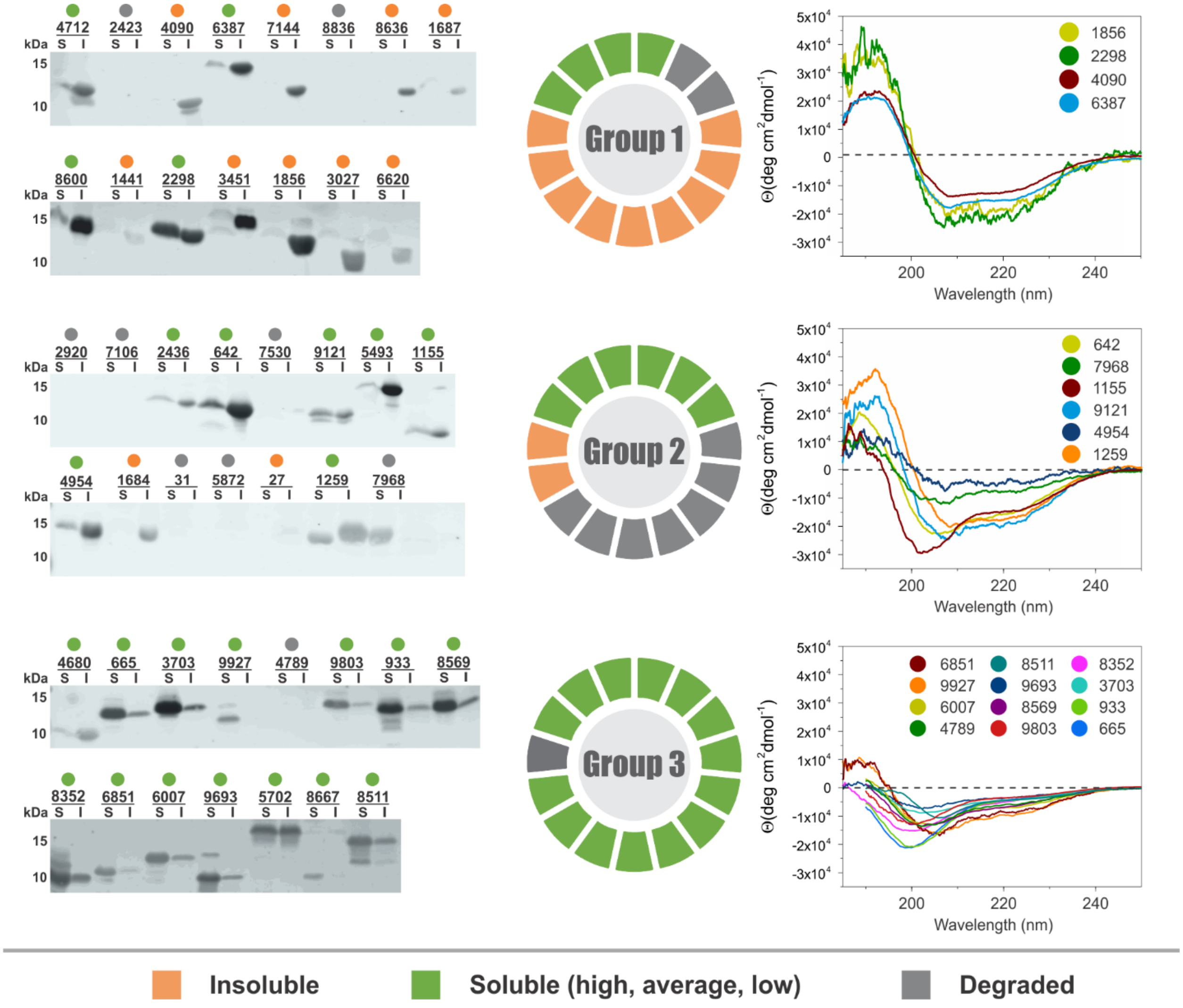


Figure 12. Summary of expression/solubility analyses and circular dichroism spectra of random proteins from Group 1, 2 and 3. **(Left)** western blot expression analysis of NBP's in *E. coli*. S – soluble fraction of the lysate, I - insoluble fraction; **(Middle)** a pie graph summarizing the solubility of NBP's based on western blot profiles; **(Right)** electronic circular dichroism spectra of successfully overexpressed and purified proteins from groups 1-

In summary, we concluded that in terms of secondary structure occurrence and overall aggregation propensity random proteins do not differ significantly from natural proteins. However, natural proteins do exhibit significantly higher levels of optimization towards aggregation suppression while maintaining comparable structure content. This optimization is based on a protein sequence modification rather than on amino acid composition tuning. While in young proteins the clusters are short and dispersed, in more evolved sequences, they tend to be longer and concentrated in a few specific positions which seed the formation of a hydrophobic core [103]. As surprising came the discovery of random disordered proteins to be tolerated *in vivo.* It has long been suspected that unevolved sequences might be toxic or degraded by cells as their disordered nature can make them an excellent substrate for intracellular proteases or lead to a nonspecific aggregation via exposed hydrophobic residues. Here we showed that random proteins have essential properties (low aggregation, high solubility, intracellular tolerance) to be suitable precursors for the following evolutionary optimization as novel proteins. Indeed, several *de novo* proteins have recently been confirmed to continuously arise from the non-coding genome regions (reviewed in [140]).

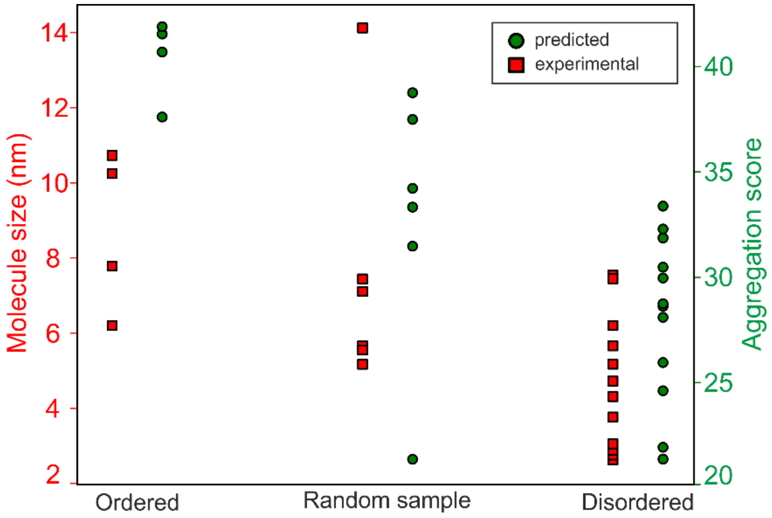


Figure 13. Scatterplot summarizing the experimentally determined molecular sizes of purified NBPs **(red)** and predicted aggregation scores for the corresponding sequences **(green)**

Author's contribution: I designed the protein expression and purification protocols, performed all biochemical experiments and analyzed the bioinformatic prediction data.

2. Development of combinatorial library design tool (CoLiDe)

Results in this section were included in the attached ***paper II*** *- Tretyachenko V. et al. CoLiDe: Combinatorial Library Design tool for probing protein sequence space. Bioinformatics. 2020 Sep 21.*

Following the scarce characterization of random protein space, we decided to undertake an investigation of collective protein structural features via library approach. Unfortunately, the existing and available design tools are not optimal for the stated tasks. While current algorithms allow for an efficient design of small targeted libraries for protein engineering, our objective was to construct a diverse protein library with each protein sequence constrained only by its amino acid composition rather than sequence. For that reason, we implemented the combinatorial library design (CoLiDe) tool which is optimized for a computationally efficient and accessible diverse library construction.

CoLiDe design principle and implementation

The purpose of CoLiDe is to compute such a combination of degenerate codons which, when combined into one DNA template, will produce a protein-coding library with user defined amino acid ratios. Inputs to the algorithm are the length of the target library, its amino acid composition, the degeneration level (maximum number of amino acids per codon) and the expressing organism codon preferences. Moreover, the algorithm allows to remove specified non-degenerate codons from inclusion into the library or to reassign certain codons to the user defined amino acids and include them into the target distribution. The primary output of the CoLiDe is a degenerate codon string which encodes the target protein library with defined amino acid distribution. The program graphically illustrates library degenerate codons with color coded amino acid coding content as well as provides statistics on mean GC content of the DNA template, average molecular weight of the protein library and differences between target and calculated distribution. CoLiDe was designed to operate either in degenerate or spiked codon optimization mode, the inputs/outputs for both are identical.

The principle of CoLiDe consists in a simplified evolutionary algorithm - specimens are optimized via random mutations, however since only one template is optimized during the calculation there is no population to select from. After initialization of the input parameters, the pool of total 3375 possible degenerate codons is filtered to contain only those which code for amino acids from the input distribution. Subsequently the program generates an initial random set of filtered degenerate codons in the size of the library's protein amino acid length. The deviation of the amino acid distribution given by this initial codon set from the target distribution is calculated as the sum of squared errors for each amino acid. Next, one degenerate codon in the initial set is exchanged for a randomly picked codon from the filtered set and the error is recalculated. If the error decreases, the exchanged codon is kept, otherwise the exchange is rejected. This cycle is repeated until 1000 × *l* subsequent rejections (where *l* stands for a protein library length in amino acids) are reached. This set, where no other exchanges provide a decrease in a sum of squared errors is returned as a solution. The computational pipeline is depicted in **Fig. 14**.

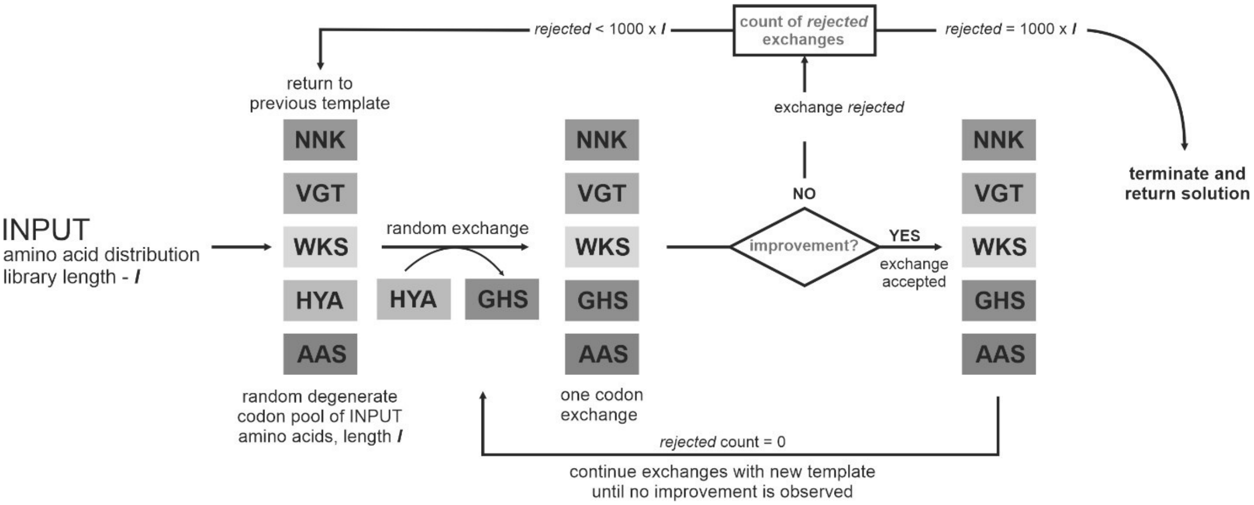


Figure 14. Schematic representation of CoLiDe computational pipeline. Input is a user defined library length and amino acid distribution of the protein library. Program filters degenerate codon pool and excludes all codons which code non defined amino acids. Subsequently algorithm randomly generates a string of degenerate codons of the library length and introduces codon exchanges until the input amino acid distribution is approximated

We tested the algorithm's performance on four library designs with differing amino acid alphabets and library lengths of 5, 10, 15, 20, 40, 60, 80 and 100 amino acids. Mean squared errors were highest on the short templates ranging from 0.11 to 0.17 and decreased with the growing template length reaching 0.005. Variance in precision showed a similar trend - coefficient of variation in short libraries ranged between 10^{-2} and 10^{-3} and decreased to 10^{-5} in longer templates. Solutions using spiked codons showed better precision with a similar variance within each group. CoLiDe runtimes were tested on all four alphabets with template sizes ranging from 5 to 400 degenerate codons. Reported runtimes of algorithm ranged between 3 to 600 s on an Intel i5-8250U equipped laptop.

Experimental validation of the protein libraries

To validate the computational methodology, we expressed and purified 45 amino acid long protein library with 33 variable amino acid positions. The rest of the construct consisted of affinity purification tag and unstructured linking sequence. The DNA template construct as well as the reverse-transcribed mRNA were characterized by high-throughput sequencing (HTS), protein library was expressed in a cell free system and characterized by MALDI-TOF mass spectrometry and by amino acid analysis of the purified product.

Both DNA and mRNA showed good agreement with the designed template. However, we did observe a bias in purine/pyrimidine base content introduced through oligonucleotide synthesis. Upon *in silico* translation, this bias resulted in enrichment of valine, leucine, and isoleucine (2.9, 2.2 and 1.6 % respectively) and depletion of proline, threonine, and alanine (3, 2.2 and 2.4 %) (**Fig. 15**.). Overall mean squared error of amino acid composition remained around 0.02. Statistical analysis of the sequencing data showed that most of all sequences (99.9 %) are unique. These results indicate that while CoLiDe can provide a precise amino acid distribution for a designed library, one should be aware of the nucleotide bias that might be introduced during oligonucleotide synthesis of highly degenerate DNA templates. Such nucleotide composition bias of DNA libraries depends on each synthesis provider (unpublished observation).

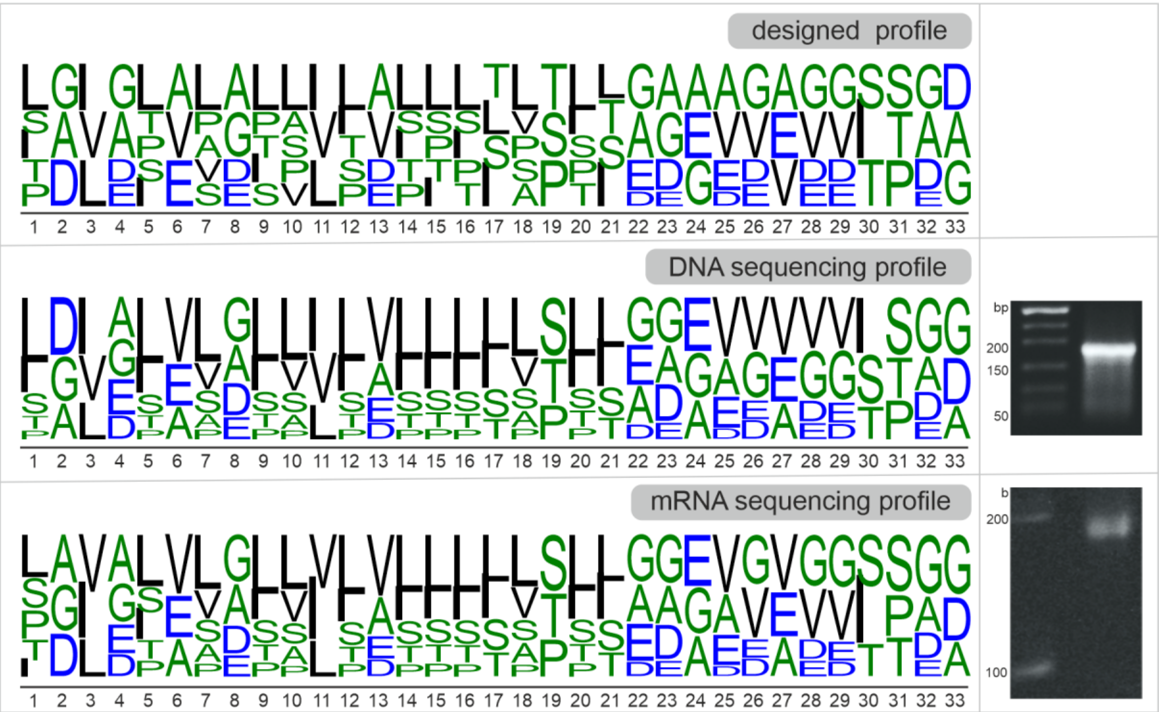


Figure 15. Agarose gels of DNA (middle right) and mRNA (bottom right) libraries and sequence logos representing the amino acid occurrence on all variable positions of designed library (top left), experimental DNA library (middle left) and experimental mRNA library (bottom left)

Following the initial sequence variability, the degenerate DNA template was transcribed, and library was expressed via cell free system, purified by His-tag affinity chromatography under the denaturing conditions and its molecular weight distribution was analysed by MALDI-TOF mass spectrometry. Mass spectrometric evaluation of molecular weight distribution is in a good agreement with the expected values obtained by *in silico* translation of 600 000 randomly generated proteins and 600 000 sequences from sequenced DNA and mRNA templates. The experimental spectrum is represented by normal weight distribution with a mean value 5 029 Da and a standard deviation of 120.6. Expected mean value from the template design is 4 902 Da and an expected mean from the HTS is 4 957 Da (**Fig. 16B**). Thus, while part of the shift can be explained by nucleotide bias of the synthesized DNA template, the rest of it lies on the additional compositional bias introduced by translation and purification. Amino acid analysis of the purified library showed under-representation in alanine, aspartic acid and threonine (by 2-4 % of the target amount) and enrichment in glutamic acid and glycine (by ~5 % from the input), likely due to their impact on protein solubility and possibly also as a result of contamination by carry-over proteins from the cell free expression system (**Fig. 16C**).

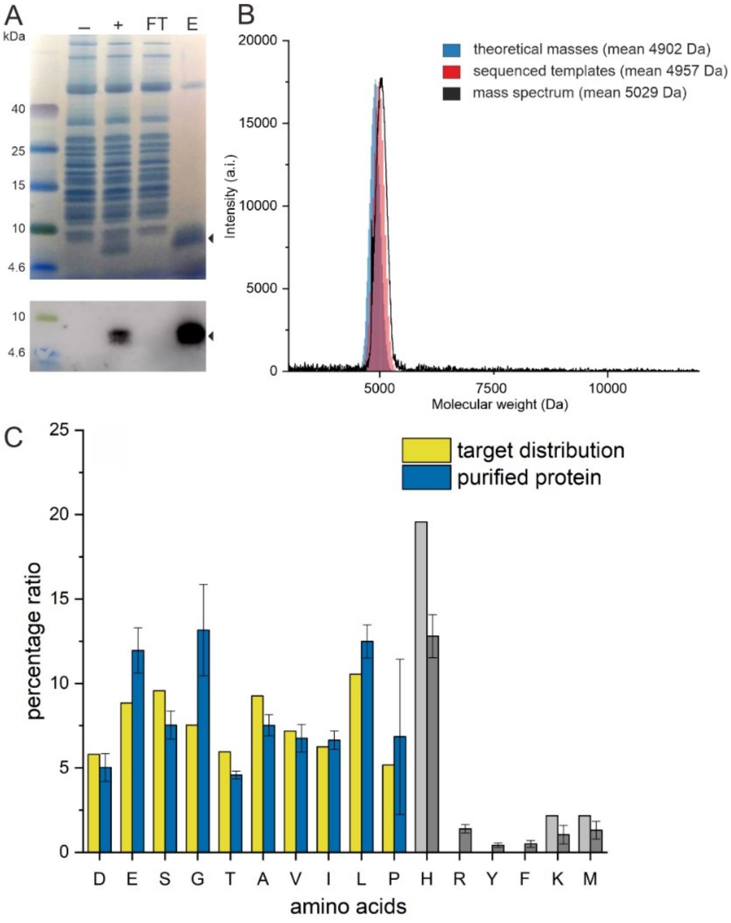


Figure 16. Experimental analysis of the cell free expressed protein library by SDS-PAGE gel/western blot (A), MALDI-TOF mass spectrometry (B) and amino acid analysis (C)

Author's contribution: I wrote an intial code, performed all biochemical experiments, and analyzed the HTS data.

3. Characterization of combinatorial protein libraries with distinct amino acid alphabets

Results in this section represent preliminary progress on combinatorial protein library characterization. This study builds on the results from the previous two sections.

The previously described sampling of 45 random proteins allowed us to focus on individual NBP's in detail, however, to infer general characteristics of random protein space as well as to deduce the impact of amino acid alphabet on protein structure, high-throughput approaches are necessary. Here we utilized CoLiDe to design two libraries with 20 (canonical set) and 10 (early set) amino acid alphabets and studied their behaviour in the presence of contemporary protein folding enhancers (DnaK, DnaJ and GrpE apparatus). We assessed libraries solubility, aggregation propensity and sensitivity against two different proteases.

Library design, preparation and validation

Combinatorial libraries of 100 amino acids in length and two different alphabets in composition were designed using the CoLiDe algorithm. The first library is constructed with a full 20 amino acid alphabet (20F), while the second one consists of a prebiotically plausible subset of this alphabet (10E, i.e. A, S, D, G, L, V, E, T, I and P). The amino acid ratios of both libraries reflected the natural amino acid distribution adopted from the UniProt statistics. Additionally, sequences for affinity purification were inserted on N'- (FLAG tag) and C'- (6xHis tag) termini and coding sequence of the thrombin cleavage site (F+ and E+, sequence ALVPRGS) and corresponding negative control non-cleavable site (F- and E-, sequence ALVGSGS) in the middle of the construct (**Fig. 17.**).

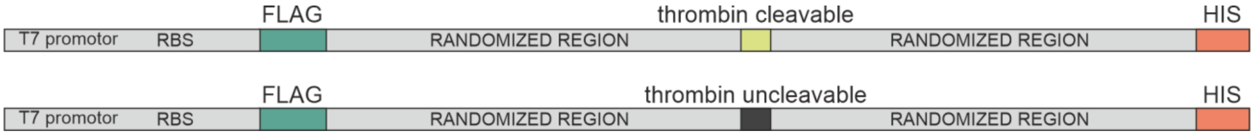


Figure 17. Schematic illustration of random protein library coding DNA template. Template contains T7 promoter sequence for *in vitro* transcription, ribosome binding site (RBS) for translation, FLAG-tag and His-tag encoding sequences on N'- and C'- termini of proteins, thrombin cleavage site (ALVPRGS) encoding sequence in the center of the template and randomized part encoding 85 amino acids

The libraries were synthesized in the form of two oligonucleotides, annealed on the cleavage-site coding sequences and filled in with the Klenow polymerase fragment (**Methods, 1; Fig. 18**). The linear double stranded construct contained all functional sequences required for *in vitro* transcription and translation. The identity of constructs was validated by HTS (**Fig. 19**). Assembled templates were *in vitro* transcribed and translated using a cell free protein synthesis system (**Methods, 2 and 4**), purified via His-tag affinity chromatography (**Methods, 6**) and analysed by MALDI-TOF mass spectrometry (**Fig. 20**).

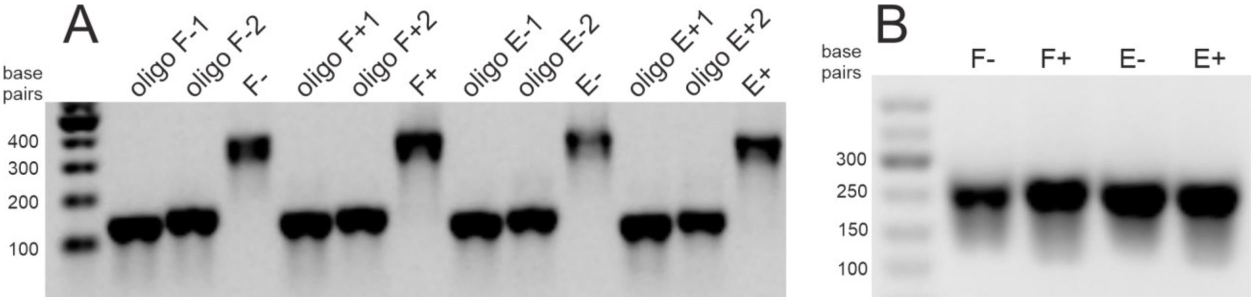


Figure 18. Agarose gels illustrating DNA library assembly (A) and RNA product of *in vitro* transcription (B). Library was assembled of 2 degenerate oligonucleotides within overlapping thrombin cleavage site encoding region

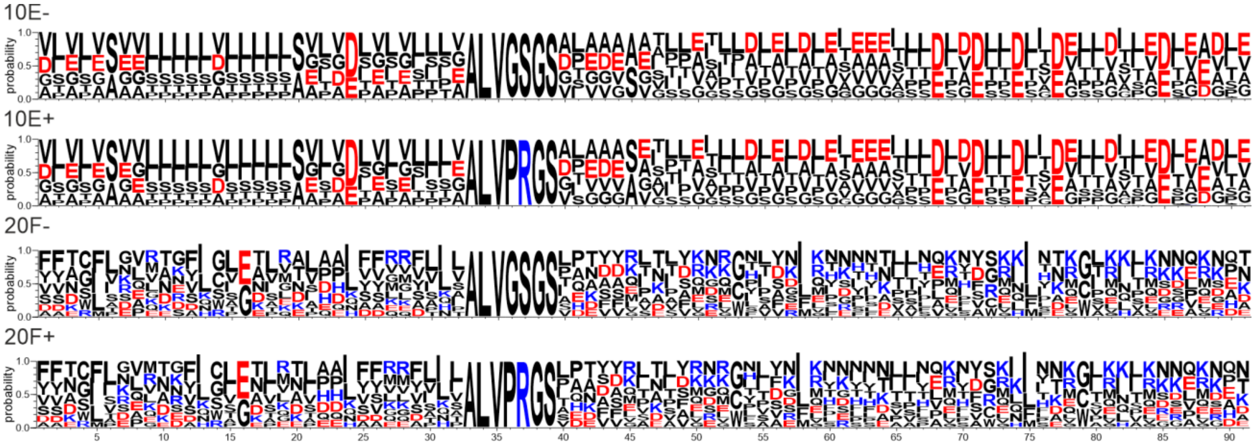


Figure 19. Sequence logos based on a high throughput sequencing data analysis illustrating the frequencies of amino acid occurrencies on all positions of the library

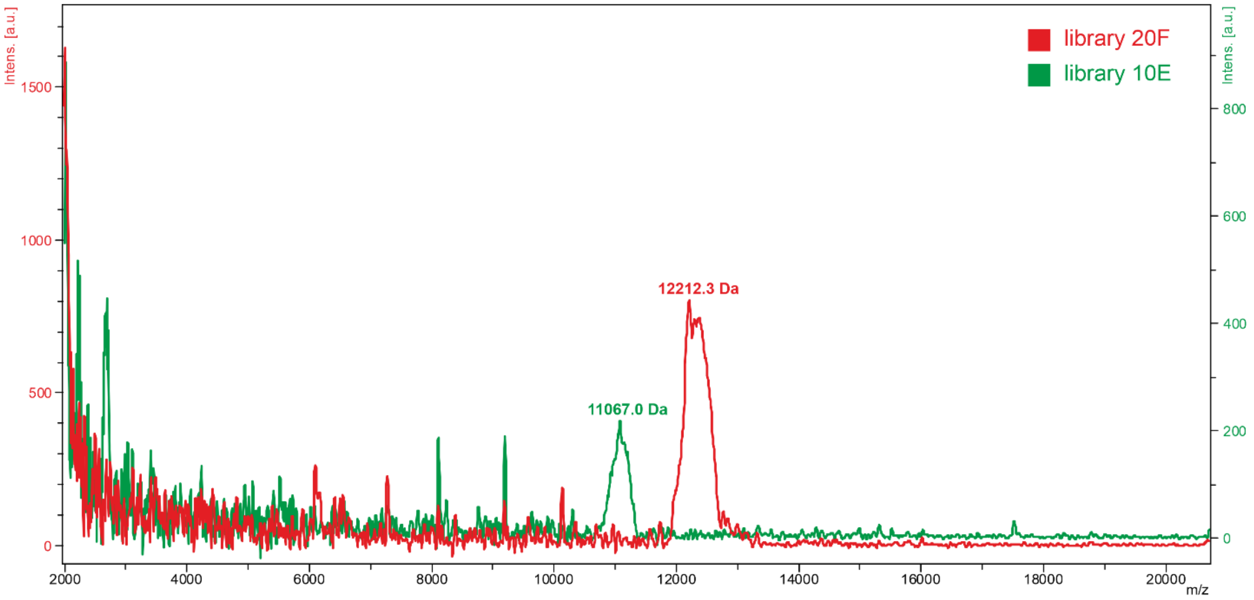


Figure 20. MALDI-TOF analysis of one-step purified libraries 20F (red) and 10E (green). Wide mass distribution signifies the broad variability of the produced library

Effect of DnaK chaperone on library solubility

Chaperones, as was pointed out in Chapter 8, play a major role in the contemporary protein world helping proteins to reach native conformations and to allow evolution to test potentially beneficial mutations without taking aggregation hazards. Here, we tested whether random, unevolved sequences would interact with chaperones similarly as their natural counterparts and we studied the effects of this interaction on protein folding and solubility. First, we screened DnaK chaperone effect on expression levels and solubility of library 20F and 10E in 25, 30 and 37 °C (**Methods, 5; Fig. 21**). Western blot analysis showed a pronounced effect of chaperones on library 20F solubilization in all tested temperatures (**Fig. 21B**). Marked solubilization effect of chaperones on the library can be of significance for *de novo* gene formation, protein products of which can be aggregation-prone provided their unevolved nature. In contrast, solubility of library 10E was not affected by the presence of chaperones and moreover, library expression was suppressed (**Fig. 21A**). This effect can be partially explained by the absence of positively charged residues in 10E library which in natural proteins mark the aggregation-prone sequence stretches in order to navigate the chaperone binding [100]. However, basic residues are not required for successful interaction with chaperones which associate unspecifically with hydrophobic amino acid patches in a protein sequence. Moreover, amino acid distributions of 10E produces proteins with increased overall hydrophobicity when compared to 20F proteins, hence with increased number of potential chaperone binding sites. An alternative explanation for the observed effect is that highly soluble 10E proteins are already structured and do not require chaperone-assisted solubilization and folding. However, this hypothesis needs to be elaborated by subsequent experiments.

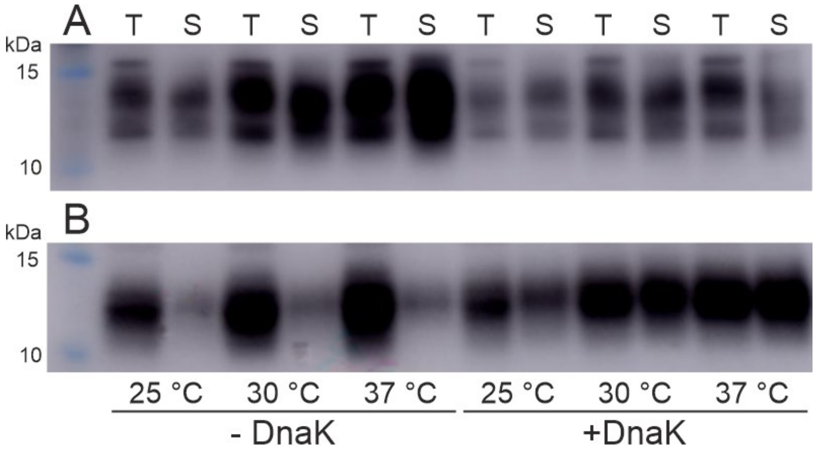


Figure 21. Western blot analysis of 10E (A) and 20F (B) solubilities in 25, 30 and 37 °C. Reactions were performed in absence (-DnaK) and presence (+DnaK) chaperones. Equal volumes of total (T) and soluble (S) reaction products were analysed

Analysis of antagonistic effects of DnaK and Lon protease on library proteins

To investigate the dynamic nature of random protein folding, libraries were expressed in the presence of either DnaK chaperone, Lon protease or both. DnaK and Lon participate in natural protein misfolding response. While DnaK stabilizes and kinetically restricts protein, helping it to fold into the native conformation, Lon protease degrades unfolded proteins reducing the intracellular stress levels. Hence, both proteins act antagonistically and inclusion of Lon protease into the cell free expression system can help to determine whether chaperone-assisted solubilization leads to soluble unfolded proteins or rather to compacted structures protected from the digestion by Lon. All four possible inclusions of chaperone and protease were tested and assessed by western blot solubility assay.

The solubility analysis confirmed the previously described observations (pronounced chaperone effect on 20F library, inhibition of expression of 10E library, **Fig. 21**) and in addition provided an interesting insight into the library intrinsic behaviour (**Fig. 22**). Proteins of the 20F library are degraded by Lon protease in both *DnaK+/Lon+* and *DnaK-/Lon+* samples suggesting an increased levels of protein disorder within the library (**Fig 22B**). A marked decrease in the soluble fraction of the 20F proteins in *DnaK-/Lon+* indicates formation of insoluble aggregates with a minute fraction of folded proteins and/or soluble aggregates (**Fig 22B**). On the other hand, soluble fraction of the 20F proteins in *DnaK+/Lon+* is equal to the total fraction of expressed proteins revealing the dominant role of chaperones on protection against Lon protease digestion (**Fig 22B**). In contrast, similar analysis of 10E proteins indicates comparable levels of

Lon degradation in both *DnaK+/Lon+* and *DnaK-/Lon+* supporting previously stated hypothesis of the independence of 10E proteins on chaperone assisted folding (**Fig 22A**).

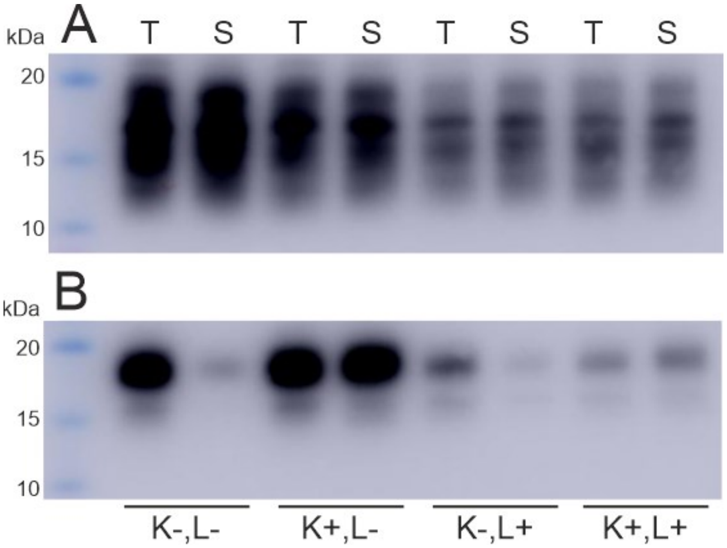


Figure 22. Western blot analysis of 10E (A) and 20F (B) solubilities in co-translational presence/absence of chaperones (K+/K-) and Lon protease (L+/L-). Same volumes of total (T) and soluble (S) reaction products were analysed

Analysis of the purified protein libraries by specific thrombin digestion

To assess posttranslational folding state of protein libraries, proteins were purified by affinity chromatography and subjected to selective proteolysis by thrombin. Purification step allows to control the proteolytic reaction conditions as well as to enrich monomeric and low-oligomeric fraction of the expressed proteins due to potentially increased accessibility of affinity tag in non-aggregated forms of proteins.

Analysis of the purified library shows high and moderate thrombin sensitivity of the 20F proteins in *DnaK-/Lon-* and *DnaK+/Lon-,* respectively (**Fig. 23B**). This observation agrees with the previous Lon digestion analysis which showed that both *DnaK+/Lon+* and *DnaK-/Lon+* samples of library 20F are enriched in unfolded proteins (**Fig. 22B**) – both thrombin and Lon degraded unfolded fraction of the library. However, while Lon protease digests unfolded proteins co-translationally, thrombin acts on a fully translated protein. This nuance can explain higher proportion of the uncleaved product in *DnaK+/Lon-* in comparison to *DnaK-/Lon-* where proteins had an opportunity to interact with the chaperones post-translationally (**Fig. 23B**). In addition, both *DnaK+/Lon+* and *DnaK-/Lon+* library samples show moderate level of degradation suggesting the flexible nature of some of the 20F proteins central part (**Fig. 23B**). A possible interpretation of the additional proteolysis by thrombin suggests partial folding of library proteins and inaccessibility of hydrophobic core residues. While Lon might not recognize such substrates, thrombin will cleave on unstructured and accessible specific site.

Library 10E exhibits weak sensitivity to proteolysis in all tested samples except *DnaK-/Lon+* (**Fig. 23A**). The proteolysis of 10E proteins in *DnaK-/Lon-* can be explained by simple digestion of library's unfolded fraction and absence of it in *DnaK-/Lon+* by co-translational library filtering for the unfolded proteins. However, the interpretation of the *DnaK+/Lon+* sample cleavage is non-trivial. A possible explanation could be in unproductive interaction (stabilization of soluble aggregates and unfolded proteins) of chaperones with unfolded and aggregation-prone 10E proteins. This interaction can also explain higher purification efficiency of chaperone-treated 10E library which previously exhibited lower expression levels (**Fig. 21A, Fig. 22A**). Proteins from the 10E library could be expressed in the form of soluble aggregates which are not purified efficiently and thus the yields of *DnaK-/Lon-* library is lower in comparison to *DnaK+/Lon-* sample. However, this anti-aggregation activity does not lead to the folded proteins and so library is subjected to thrombin digestion. In summary, this proposed scenario leads to the conclusion that (i) Lon co-translational and thrombin post translational proteolyses lead to the isolation of the folded fraction of the library and (ii) *DnaK+/Lon-* and *DnaK-/Lon-* samples cleaved by thrombin represent the combined folded and aggregated fractions with the enrichment in soluble aggregates in *DnaK+/Lon-* library.

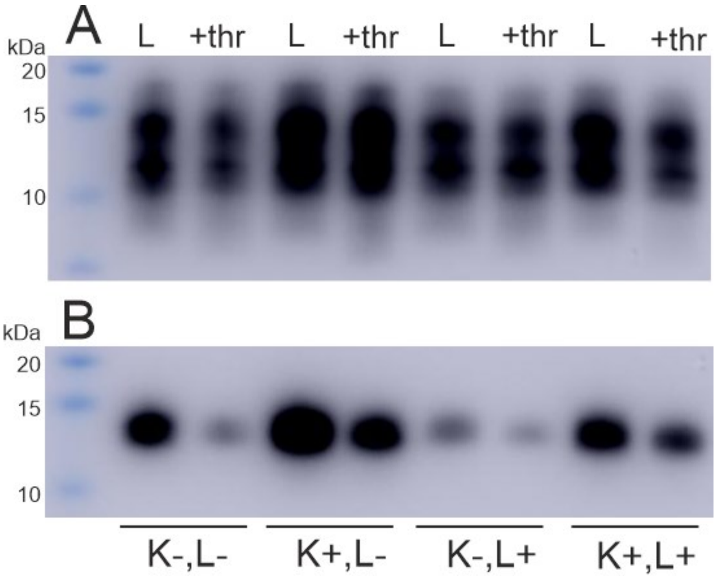


Figure 23. Western blot analysis of 10E (A) and 20F (B) thrombin sensitivity assay on affinity purified libraries in co-translational presence/absence of chaperones (K+/K-) and Lon protease (L+/L-). Same volumes of total (T) and soluble (S) reaction products were analysed

Data presented in this chapter represent a qualitative characterization of protein libraries and served to form an initial hypothesis on library collective structural characteristics. The next stage of the characterisation will gather quantitative input from Western Blot assays – library chemiluminescent signal will be calibrated to ensure linear response to concentration changes and signals will be averaged from multiplicate measurements. Moreover, levels of protein compactness in the combinatorial libraries will be assessed by analytical size exclusion chromatography.

Author's contribution: I designed and performed all the libraries biochemical experiments and analyzed the HTS data

4. Characterization of aromatic-less variant of dephospho coenzyme-A kinase (DPCK)

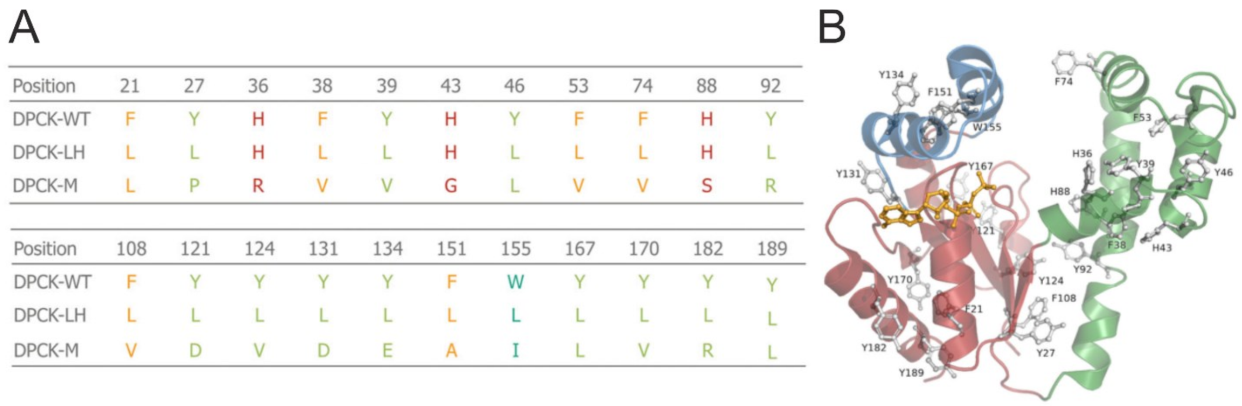
Results in this section were included in the attached ***paper III*** *– Makarov M. et al. Enzyme catalysis prior to aromatic residues: Reverse engineering of a dephospho-CoA kinase. Protein Science. 2021; 1– 13.*

Aromatic amino acids are hypothesized to be among the latest arrivals into the amino acid alphabet and at the same time are the strongest structure promoters of contemporary proteins [3]. However, early proteins probably served their function in their absence. To test the hypothesis that aromatic amino acids might be dispensable for the basal protein function we designed an aromatics-less version of a contemporary enzyme dephospho coenzyme A kinase (DPCK). Since none of the aromatic amino acids in the protein are known to be essential for enzymatic function, DPCK represents an ideal candidate for alphabet/structure relationship investigation.

Mutant generation, expression and purification

We searched the PDB database for solved structures of confirmed and putative DPCKs from different thermophilic bacterial species and selected DPCK from *Aquifex aeolicus* based on an initial expression/solubility screening. Mutant variants were designed as follows. All aromatic amino acids were substituted by (i) Leu residues (DPCK-LH) and (ii) non-aromatic amino acids (DPCK-MH) based on the best preservation of thermodynamic stability using the Hot Spot Wizard server [141]. Additionally, all His residues (besides the aromatics Tyr, Phe, Trp) were substituted by Leu or other non-aromatic residues using the same logic (DPCK-L and DPCK-M). In comparison with the wild type DPCK, 10 % (DPCK-LH/MH) and 11 % (DPCK-L/M) of aromatic residues were substituted (**Fig. 24A**). Upon preliminary expression, purification in *E. coli* and activity assessment, only DPCK-LH and DPCK-M variants were selected since DPCK-L and DPCK-M did not have any measurable phosphotransferase or ATPase activities.

DPCK WT, -LH and -M variants were purified to homogeneity using a three-step purification protocol and their identity was confirmed by mass spectrometry. Additionally, analytical size-exclusion chromatography showed that while DPCK-WT and LH eluted as monomers corresponding to their molecular weights, DPCK-M variant resembled either dimeric or disordered monomeric form in the elution profile.



Position	21	27	36	38	39	43	46	53	74	88	92
DPCK-WT	F	Y	H	F	Y	H	Y	F	F	H	Y
DPCK-LH	L	L	H	L	L	H	L	L	L	H	L
DPCK-M	L	P	R	V	V	G	L	V	V	S	R

Position	108	121	124	131	134	151	155	167	170	182	189
DPCK-WT	F	Y	Y	Y	Y	F	W	Y	Y	Y	Y
DPCK-LH	L	L	L	L	L	L	L	L	L	L	L
DPCK-M	V	D	V	D	E	A	I	L	V	R	L

Figure 24. Summary of mutated aromatic residues and their positions in comparison to the DPCK-WT sequence (A) and (B) structure of the DPCK enzyme with highlighted aromatic amino acids (grey) and bound ATP substrate (yellow)

Characterization of aromatic-less mutants and comparison with the wild type enzyme

We characterized the enzymatic activity and specificity of both DPCK mutants and compared them to the wild-type enzyme. The measured catalytic efficiency of DPCK-WT for ATP and dCoA (3.4×10^4 and 5.7×10^4 $M^{-1} \cdot s^{-1}$ for dCoA and ATP, respectively) were similar to the previously reported values of DPCK from *E. histolytica*. Unlike DPCK-WT, both mutants showed the ability to hydrolyze ATP in absence of dCoA with 1000× lower catalytic efficiencies (355 and 118 $M^{-1} \cdot s^{-1}$ for ATP, respectively). Comparing both mutants, DPCK-M has lost the phosphotransferase activity, e.g. only ATPase activity was observed. On the other hand, DPCK-LH maintained the dCoA-dependent phosphotransferase activity above 80 mM dCoA with K_M greater than 200 mM. We performed HPLC-MS analysis to confirm the identity of reaction products catalyzed by both DPCK-WT and DPCK-LH. Analysis proved CoA formation in both, however we detected 100× less CoA in the reactions catalyzed by DPCK-LH (summarized in **Table 4**).

Table 4. Summary of measured kinetic properties of wild type and aromatic-less mutants of DPCK

Enzyme	Substrate	K_M (μM)	V_{max} (μM · min^{-1})	k_{cat} (s^{-1})	k_{cat}/K_M (M^{-1} · s^{-1})
DPCK-WT	dCoA (with 200 μM ATP)	24.3 ± 1.7	1.57 ± 0.17	0.817 ± 0.088	33621.4
	ATP (with 200 μM dCoA)	12.7 ± 0.3	1.41 ± 0.05	0.730 ± 0.026	57480.3
	ATP (without dCoA)	n.d.	n.d.	n.d.	n.d.
DPCK-LH	dCoA (with 200 μM ATP)	>200	n.d.	n.d.	n.d.
	ATP (with 200 μM dCoA)	65.9 ± 4.5	0.30 ± 0.010	0.0234 ± 0.0008	355.1
	ATP (without dCoA)	40.2 ± 3.8	0.11 ± 0.013	0.0086 ± 0.0010	213.9
DPCK-M	dCoA (with 200 μM ATP)	n.d.	n.d.	n.d.	n.d.
	ATP (with 200 μM dCoA)	32.3 ± 3.4	0.088 ± 0.003	0.0038 ± 0.0001	117.6
	ATP (without dCoA)	29.2 ± 2.5	0.085 ± 0.002	0.0037 ± 0.0001	126.7

The structure content of all three studied proteins was assessed by electronic circular dichroism (ECD) measurement followed by numerical analysis of the spectra (**Fig. 25A**). DPCK-WT showed high α-helical content in correspondence with available PDB structure. Variants DPCK-LH and -M demonstrated higher β-sheet content and in case of DPCK-M enrichment in protein disorder. The stability of mutants was investigated by urea-titrations and ECD spectroscopy (**Fig. 25B**). While DPCK-WT and -LH spectra remained relatively constant upon urea titration up to 2 M concentration, DPCK-M variant started to lose its structural integrity already in very low urea concentrations (> 100 mM). Structural similarity of DPCK-LH and -WT was further confirmed by 1D and 2D HN NMR spectroscopy (**Fig. 25C**). Yet, while DPCK-WT showed clear signals near 1 ppm indicative of methyl groups in the hydrophobic core, these were absent in DPCK-LH spectra. However, the signal dispersion in the -NH- region implies that the -LH variant is at least partially folded in contrast to DPCK-M which showed lack of a tertiary structure formation.

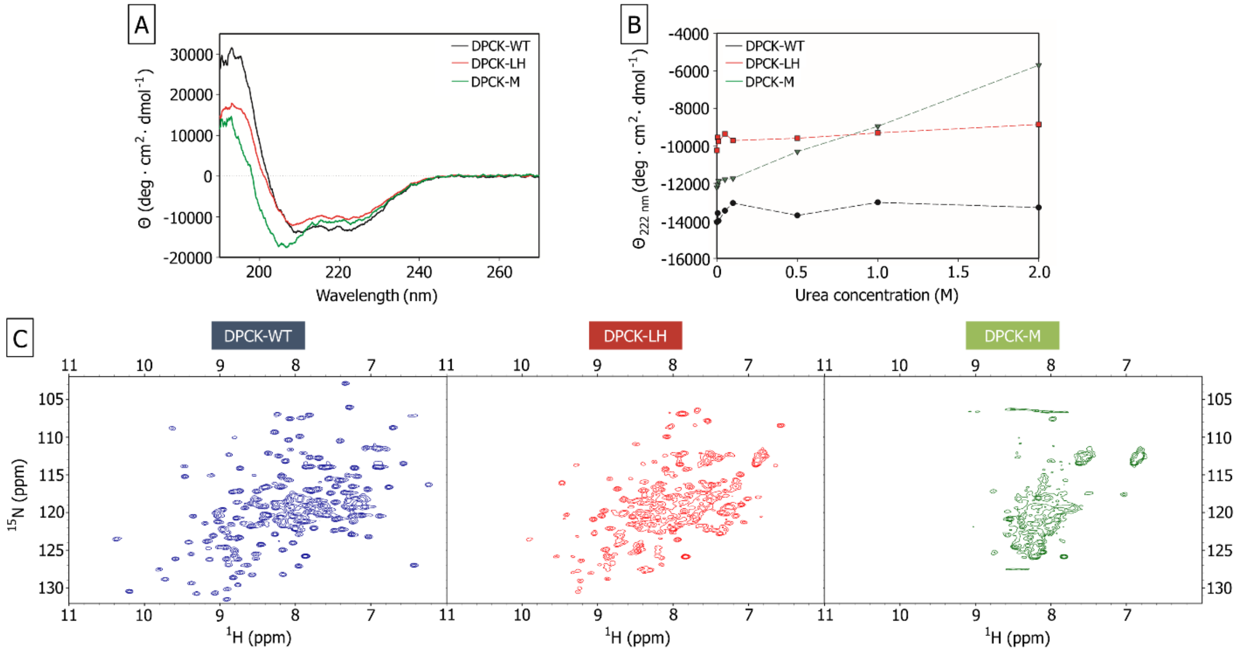


Figure 25. Secondary and tertiary structure characterization of dephospho CoA kinase (DPCK) variants. (A) Far-UV CD spectra of DPCK proteins. (B) Change in ellipticity at 222 nm upon 0–2 M urea titration of DPCK proteins. (C) 2D NMR of DPCK proteins

Furthermore, structural integrity of all three proteins was assessed by a limited proteolysis by LysC proteinase (**Fig. 26**). While DPCK-WT was protease resistant, both mutant variants were gradually digested by time of the experiment. Mutants showed different proteolysis dynamics - while DPCK-LH was hydrolysed to yield large fragments with the approximate sizes of 15 kDa, DPCK-M variant was fully digested indicating lack of the intradomain folding patterns and overall tertiary structure.

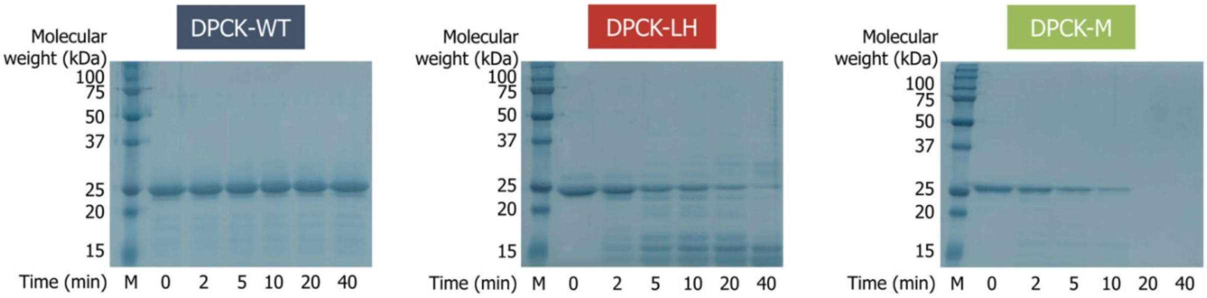
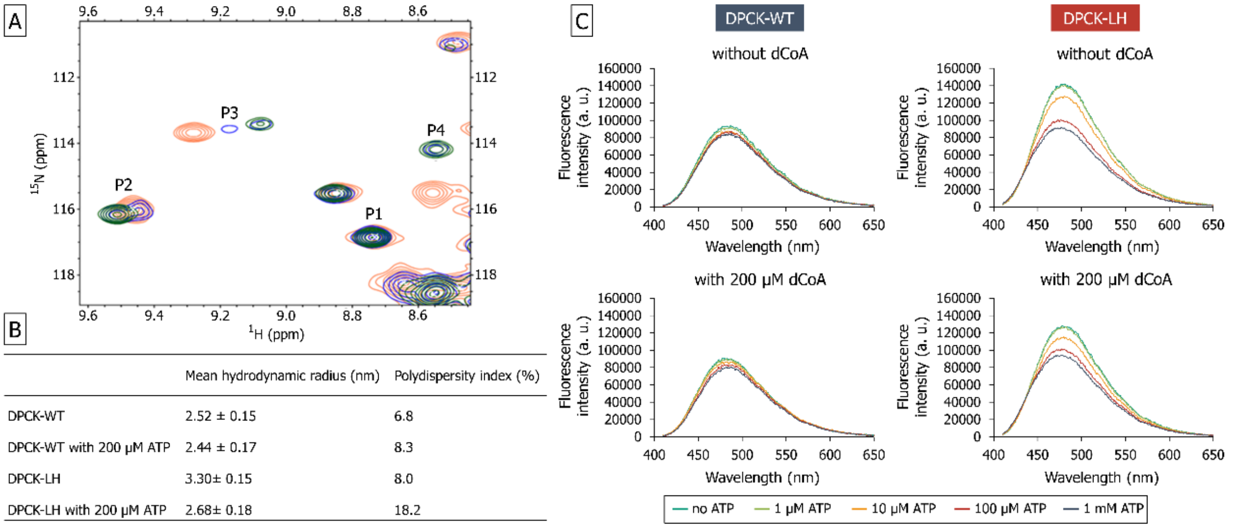


Fig 26. 14% SDS-polyacrylamide gels of limited proteolysis of dephospho CoA kinase (DPCK) proteins visualized by imidazole-zinc staining after SDS-PAGE with the protein samples exposed to LysC endoproteinase for different times

In addition, we investigated the dynamics of the proteins upon ATP binding via NMR, dynamic light scattering and titration by 8-anilinonaphthalene-1-sulfonic acid (ANS). These approaches confirmed the molten globular nature of the ATP-unbound form of DPCK-LH and interestingly, indicated protein compaction upon the ligand binding. According to the DLS measurements, hydrodynamic radius of DPCK-LH is reduced by ~20 % and reaches that of value DPCK-WT value upon ATP addition (**Fig. 27B,C**). This observation supports the previously stated hypotheses that cofactors might play a crucial role in early protein structure stabilization.



	Mean hydrodynamic radius (nm)	Polydispersity index (%)
DPCK-WT	2.52 ± 0.15	6.8
DPCK-WT with 200 μM ATP	2.44 ± 0.17	8.3
DPCK-LH	3.30± 0.15	8.0
DPCK-LH with 200 μM ATP	2.68± 0.18	18.2

Figure 27. Structural characterization of dephospho-CoA kinase (DPCK)-WT and -LH upon substrate binding. (a) An exemplary closeup of DPCK-LH 2D NH NMR spectra induced by ATP binding (red—free protein [100 µM], blue—300 µM ATP, green—1000 µM ATP); labeled peaks: (P1) N-H signal not influenced by protein-ATP interaction. (P2, P3) N-H signal undergoing medium-slow to slow exchange on NMR chemical shift time scale (µs-ms). (P4) N-H signal documenting a slow exchange process. (b) Mean hydrodynamic radius of DPCK-WT and -LH variants with and without 200 µM ATP measured by dynamic light scattering. (c) The steady-state fluorescence spectra of ANS binding at excitation wavelength 380 nm. The spectra were measured at different concentrations of ATP (with and without 200 µM dCoA), and each spectrum is the average of three individual scans. The fluorescence was recorded between 410 and 650 nm after exciting the protein solution at 380 nm

Author's contribution: participated on initial project outline, protein expression/purification and structural characterization assays design. Moreover, I supervised the first author of the study.

SUMMARY

The overall aims of this work were to (i) investigate properties of random protein space and their relationship to natural proteins and (ii) study the effect of amino acid alphabet on protein structure and function.

The following results were obtained and included in the three attached scientific publications and preliminary data which will lead to a subsequent publication.

- Computational analysis of random protein library showed similar secondary structure content but different aggregation tendencies in comparison to natural proteins.

- Experimental characterization of 45 random proteins showed agreement with computational analysis in the secondary structure content and aggregation propensity and revealed that disordered random sequences are better tolerated in intracellular mileu than their structure-rich counterparts.

- Combinatorial library design tool (CoLiDe) was implemented and made available to the broad scientific community.

- The CoLiDe algorithm was validated experimentally. The validation demonstrated the biases in library preparations for further experiments.

- Combinatorial protein libraries with different amino acid compositions were prepared and purified *in vitro* and their biochemical characterization suggest different structural tendencies within the random sequence space.

- Characterizations of aromatic-less variants of dephospho coenzyme A kinase supported the role of aromatic amino acids in achieving the structural stability of contemporary proteins but demonstrated that enzyme activity can still be gained even in their absence.

- Enhanced compaction upon the interaction of aromatic-less mutant of dephospho coenzyme A kinase with its ligands indicated the plausible importance of cofactor on early protein structure stabilization.

LIST OF PUBLICATIONS

Publications directly supporting this doctoral thesis:

1) **Tretyachenko V**, Vymětal J, Bednárová L, Kopecký V, Hofbauerová K, Jindrová H, Hubálek M, Souček R, Konvalinka J, Vondrášek J, Hlouchová K. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Scientific reports. 2017 Nov 13;7(1):1-9. (IF 3.998)
2) **Tretyachenko V**, Voráček V, Souček R, Fujishima K, Hlouchová K. CoLiDe: Combinatorial Library Design tool for probing protein sequence space. Bioinformatics. 2020 Sep 21. (IF 5.610)
3) Makarov, M, Meng, J, **Tretyachenko, V**, et al. Enzyme catalysis prior to aromatic residues: Reverse engineering of a dephospho-CoA kinase. Protein Science. 2021; 1– 13. (IF 3.876)

Other publications by the author:

1) Makukhin N, **Tretyachenko V**, Moskovitz J, Míšek J. A ratiometric fluorescent probe for imaging of the activity of methionine sulfoxide reductase A in cells. Angewandte Chemie International Edition. 2016 Oct 4;55(41):12727-30. (IF 12.959)
2) Kadek A, **Tretyachenko V**, Mrazek H, Ivanova L, Halada P, Rey M, Schriemer DC, Man P. Expression and characterization of plant aspartic protease nepenthesin-1 from Nepenthes gracilis. Protein expression and purification. 2014 Mar 1;95:121-8. (IF 1.695)
3) Fejfarová K, Kádek A, Mrázek H, Hausner J, **Tretyachenko V**, Koval T, Man P, Hašek J, Dohnálek J. Crystallization of nepenthesin I using a low-pH crystallization screen. Acta Crystallographica Section F: Structural Biology Communications. 2016 Jan 1;72(1):24-8. (IF 0.968)

References

1. Voet, D., Voet, J. & Pratt, C. *Fundamentals of Biochemistry life at the molecular level*. *Annals of Internal Medicine* vol. 14 (2011).

2. Wright, P. E. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2014).

3. Uversky, V. N. Intrinsically disordered proteins and their 'Mysterious' (meta)physics. *Front. Phys.* **7**, (2019).

4. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).

5. Bull, A. T., Goodfellow, M. & Slater, J. H. Biodiversity as a Source of Innovation in Biotechnology. *Annu. Rev. Microbiol.* **46**, 219–246 (1992).

6. William, B. W., David, C. C. & William, J. W. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6578–83 (1998).

7. Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **33**, 3390–3400 (2005).

8. Kurland, C. G., Canbäck, B. & Berg, O. G. The origins of modern proteomes. *Biochimie* **89**, 1454–1463 (2007).

9. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).

10. Wang, M., Caetano-Anolles, D. & Caetano-Anolles, G. The origin , evolution and structure of the protein world. **637**, 621–637 (2009).

11. Frenkel-Pinter, M., Frenkel-Pinter, M., Samanta, M., Ashkenasy, G., Leman, L. J. & Leman, L. J. Prebiotic Peptides: Molecular Hubs in the Origin of Life. *Chem. Rev.* **120**, 4707–4765 (2020).

12. Runnels, C. M., Lanier, K. A., Williams, J. K., Bowman, J. C., Petrov, A. S., Hud, N. V. & Williams, L. D. Folding, Assembly, and Persistence: The Essential Nature and Origins of Biopolymers. *J. Mol. Evol.* **86**, 598–610 (2018).

13. Martínez-Bachs, B. & Rimola, A. Prebiotic Peptide Bond Formation Through Amino Acid Phosphorylation. Insights from Quantum Chemical Simulations. *Life* **9**, 75 (2019).

14. Ross, D. S. & Deamer, D. Dry/wet cycling and the thermodynamics and kinetics of prebiotic polymer synthesis. *Life* **6**, (2016).

15. Weber, A. L. & Pizzarello, S. The peptide-catalyzed stereospecific synthesis of tetroses: A possible model for prebiotic molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12713–12717 (2006).

16. Wieczorek, R., Adamala, K., Gasperi, T., Polticelli, F. & Stano, P. Small and random peptides: An unexplored reservoir of potentially functional primitive organocatalysts. the case of seryl-histidine. *Life* **7**, (2017).

17. Alva, V., Söding, J. & Lupas, A. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**, (2015).

18. Caetano-Anollés, G. & Caetano-Anollés, D. An evolutionarily structured universe of protein

architecture. *Genome Res.* **13**, 1563–1571 (2003).

19. Caetano-Anollés, G. & Caetano-Anollés, D. Universal sharing patterns in proteomes and evolution of protein fold architecture and life. *J. Mol. Evol.* **60**, 484–498 (2005).

20. Kim, H. S., Mittenthal, J. E. & Caetano-Anollés, G. MANET: Tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics* **7**, (2006).

21. Wang, M., Yafremava, L. S., Caetano-Anollés, D., Mittenthal, J. E. & Caetano-Anollés, G. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17**, 1572–1585 (2007).

22. Goldman, A. D., Samudrala, R. & Baross, J. A. The evolution and functional repertoire of translation proteins following the origin of life. *Biol. Direct* **5**, (2010).

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

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

25. Pohorille, A., Wilson, M. A. & Shannon, G. Flexible proteins at the origin of life. *Life* **7**, 1–20 (2017).

26. Tokuriki, N. & Tawfik, D. S. Protein dynamism and evolvability. *Science (80-.).* **324**, 203–207 (2009).

27. Edwards, H., Abeln, S. & Deane, C. M. Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies. **9**, (2013).

28. Banerjee, S. & Chakraborty, S. Molecular BioSystems with gene age in different eukaryotic lineages †. 2044–2055 (2017) doi:10.1039/c7mb00230k.

29. Eigen, M. & Schuster, P. *The Hypercycle. A Principle of Natural Self-Organization*. (Springer, 1979).

30. Longo, L. M., Despotović, D., Weil-Ktorza, O., Walker, M. J., Jabłońska, J., Fridmann-Sirkis, Y., Varani, G., Metanis, N. & Tawfik, D. S. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15731–15739 (2020).

31. Eck, R. V. & Dayhoff, M. O. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science (80-.).* **152**, 363–366 (1966).

32. Miller, S. L. A production of amino acids under possible primitive earth conditions. *Science (80-.).* **117**, 528–529 (1953).

33. Cleaves, H. J., Chalmers, J. H., Lazcano, A., Miller, S. L. & Bada, J. L. A reassessment of prebiotic organic synthesis in neutral planetary atmospheres. *Orig. Life Evol. Biosph.* **38**, 105–115 (2008).

34. Bada, J. L. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chem. Soc. Rev.* **42**, 2186–2196 (2013).

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

36. Aubrey, A. D., Cleaves, H. J. & Bada, J. L. The role of submarine hydrothermal systems in the synthesis of amino acids. *Orig. Life Evol. Biosph.* **39**, 91–108 (2009).

37. Pizzarello, S. & Shock, E. The organic composition of carbonaceous meteorites: the evolutionary story ahead of biochemistry. *Cold Spring Harb. Perspect. Biol.* **2**, (2010).

38. Sephton, M. A. Organic compounds in carbonaceous meteorites. *Nat. Prod. Rep.* **19**, 292–311 (2002).

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

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

41. Granold, M., Hajieva, P., Toşa, M. I., Irimie, F. D. & Moosmann, B. Modern diversification of the amino acid repertoire driven by oxygen. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 41–46 (2018).

42. van der Gulik, P., Massar, S., Gilis, D., Buhrman, H. & Rooman, M. The first peptides: The evolutionary transition between prebiotic amino acids and early proteins. *J. Theor. Biol.* **261**, 531–539 (2009).

43. Sobolevsky, Y. & Trifonov, E. N. Conserved sequences of prokaryotic proteomes and their compositional age. *J. Mol. Evol.* **61**, 591–596 (2005).

44. Fournier, G. P. & Alm, E. J. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. *J. Mol. Evol.* **80**, 171–185 (2015).

45. Fujishima, K., Wang, K. M., Palmer, J. A., Abe, N., Nakahigashi, K., Endy, D. & Rothschild, L. J. Reconstruction of cysteine biosynthesis using engineered cysteine-free enzymes. *Sci. Rep.* **8**, (2018).

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

47. Shibue, R., Sasamoto, T., Shimada, M., Zhang, B., Yamagishi, A. & Akanuma, S. Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* **8**, (2018).

48. Schmitz, J. & Bornberg-Bauer, E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Research* **6**, (2017).

49. Gumulya, Y. & Gillam, E. M. J. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: The 'retro' approach to protein engineering. *Biochem. J.* **474**, 1–19 (2017).

50. White, H. B. Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7**, 101–104 (1976).

51. Hanukoglu, I. Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. *Biochem. Mol. Biol. Educ.* **43**, 206–209 (2015).

52. Jones, P. M. & George, A. M. The ABC transporter structure and mechanism: Perspectives on recent research. *Cell. Mol. Life Sci.* **61**, 682–699 (2004).

53. Saraste Matti., Sibbald Peter R. & Wittinghofer Alfred. The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**, 430–434 (1990).

54. Goldman, A. D., Baross, J. A. & Samudrala, R. The enzymatic and metabolic capabilities of early life. *PLoS One* **7**, (2012).

55. Liu, P. F. & Park, C. Selective stabilization of a partially unfolded protein by a metabolite. *J. Mol. Biol.* **422**, 403–413 (2012).

56. Chen, C. & Park, C. Chaperone action of a cofactor in protein folding. *Protein Sci.* **29**, 1667–1678 (2020).

57. Ji, H. F., Kong, D. X., Shen, L., Chen, L. L., Ma, B. G. & Zhang, H. Y. Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.* **8**, (2007).

58. Shakya, A. & King, J. T. DNA Local-Flexibility-Dependent Assembly of Phase-Separated Liquid Droplets. *Biophys. J.* **115**, 1840–1847 (2018).

59. Holland, H. D. The oxygenation of the atmosphere and oceans. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 903–915 (2006).

60. Reinhard, C. T., Planavsky, N. J., Gill, B. C., Ozaki, K., Robbins, L. J., Lyons, T. W., Fischer, W. W., Wang, C., Cole, D. B. & Konhauser, K. O. Evolution of the global phosphorus cycle. *Nature* **541**, 386–389 (2017).

61. Saito, M. A., Sigman, D. M. & Morel, M. M. The bioinorganic chemistry of the ancient ocean : the co-e v olution of cyanobacterial metal requirements and biogeochemical cycles at the Archean Á Proterozoic boundary ? *Inorganica Chim. Acta* **356**, 308–318 (2003).

62. Williams, R. J. P. & da Silva, J. J. R. F. The Chemistry of Evolution. *Chem. Evol.* (2006) doi:10.1016/B978-0-444-52115-6.X5042-8.

63. Ji, H. F. & Zhang, H. Y. Bioinformatic identification of the most ancient copper protein architecture. *J. Biomol. Struct. Dyn.* **26**, 197–201 (2008).

64. Ji, H. F., Chen, L., Jiang, Y. Y. & Zhang, H. Y. Evolutionary formation of new protein folds is linked to metallic cofactor recruitment. *BioEssays* **31**, 975–980 (2009).

65. Raanan, H., Poudel, S., Pike, D. H., Nanda, V. & Falkowski, P. G. Small protein folds at the root of an ancient metabolic network. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 7193–7199 (2020).

66. Okafor, C. D., Lanier, K. A., Petrov, A. S., Athavale, S. S., Bowman, J. C., Hud, N. V. & Williams, L. D. Iron mediates catalysis of nucleic acid processing enzymes: Support for Fe(II) as a cofactor before the great oxidation event. *Nucleic Acids Res.* **45**, 3634–3642 (2017).

67. Bray, M. S., Lenz, T. K., Haynes, J. W., Bowman, J. C., Petrov, A. S., Reddi, A. R., Hud, N. V., Williams, L. D. & Glass, J. B. Multiple prebiotic metals mediate translation. *Proc. Natl. Acad. Sci.* **115**, 12164–12169 (2018).

68. Despotović, D., Longo, L. M., Aharon, E., Kahana, A., Scherf, T., Gruic-Sovulj, I. & Tawfik, D. S. Polyamines mediate folding of primordial hyperacidic helical proteins. *Biochemistry* **59**, 4456–4462 (2020).

69. Zhang, Y., Hubner, I. a, Arakaki, A. K., Shakhnovich, E. & Skolnick, J. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2605–2610 (2006).

70. Pan, X., Thompson, M. C., Zhang, Y., Liu, L., Fraser, J. S., Kelly, M. J. S. & Kortemme, T. Expanding the space of protein geometries by computational design of de novo fold families.

Science (80-.). **369**, 1132–1136 (2020).

71. Minervini, G., Evangelista, G., Villanova, L., Slanzi, D., De Lucrezia, D., Poli, I., Luisi, P. L. & Polticelli, F. Massive non-natural proteins structure prediction using grid technologies. *BMC Bioinformatics* **10**, (2009).

72. Prymula, K., Piwowar, M., Kochanczyk, M., Flis, L., Malawski, M., Szepieniec, T., Evangelista, G., Minervini, G., Polticelli, F., Wiśniowski, Z., Sałapa, K., Matczyńska, E. & Roterman, I. In silico structural study of random amino acid sequence proteins not present in nature. *Chem. Biodivers.* **6**, 2311–2336 (2009).

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

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

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

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

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

78. Yomo, T., Prijambada, I. D., Yamamoto, K., Shima, Y., Negoro, S. & Urabe, I. Properties of artificial proteins with random sequences. *Ann. N. Y. Acad. Sci.* **864**, 131–135 (1998).

79. Hayashi, Y., Sakata, H., Makino, Y., Urabe, I. & Yomo, T. Can an arbitrary sequence evolve towards acquiring a biological function? *J. Mol. Evol.* **56**, 162–168 (2003).

80. Yamauchi, A., Nakashima, T., Tokuriki, N., Hosokawa, M., Nogami, H., Arioka, S., Urabe, I. & Yomo, T. Evolvability of random polypeptides through functional selection within a small library. *Protein Eng. Des. Sel.* **15**, 619–626 (2002).

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

82. Lo Surdo, P., Walsh, M. A. & Sollazzo, M. A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nat. Struct. Mol. Biol.* **11**, 382–383 (2004).

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

84. Merrifield, B. Solid phase synthesis. *Science (80-.).* **232**, 341–347 (1986).

85. Loibl, S. F., Harpaz, Z., Zitterbart, R. & Seitz, O. Total chemical synthesis of proteins without HPLC purification. **7**, 6753–6759 (2016).

86. McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random mutagenesis by error-prone PCR. *Methods Mol. Biol.* **634**, 103–109 (2010).

87. Arnold, F. H., Georgiou, G., Cirino, P. C., Mayer, K. M. & Umeno, D. Generating Mutant Libraries Using Error-Prone PCR. *Dir. Evol. Libr. Creat.* 3–10 (2003) doi:10.1385/1-59259-395-

x:3.

88. Yagodkin, A., Azhayev, A., Roivainen, J., Antopolsky, M., Kayushin, A., Korosteleva, M., Miroshnikov, A., Randolph, J. & Mackie, H. Improved synthesis of trinucleotide phosphoramidites and generation of randomized oligonucleotide libraries. *Nucleosides, Nucleotides and Nucleic Acids* **26**, 473–497 (2007).

89. Kille, S., Acevedo-Rocha, C. G., Parra, L. P., Zhang, Z. G., Opperman, D. J., Reetz, M. T. & Acevedo, J. P. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol.* **2**, 83–92 (2013).

90. Arkin, A. P. & Youvan, D. C. Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis. *Bio/Technology* **10**, 297–300 (1992).

91. Pines, G., Pines, A., Garst, A. D., Zeitoun, R. I., Lynch, S. A. & Gill, R. T. Codon compression algorithms for saturation mutagenesis. *ACS Synth. Biol.* **4**, 604–614 (2015).

92. Halweg-Edwards, A. L., Pines, G., Winkler, J. D., Pines, A. & Gill, R. T. A web interface for codon compression. *ACS Synth. Biol.* **5**, 1021–1023 (2016).

93. Parker, A. S., Griswold, K. E. & Bailey-Kellogg, C. Optimization of combinatorial mutagenesis. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **6577 LNBI**, 321–335 (2011).

94. Jacobs, T. M., Yumerefendi, H., Kuhlman, B. & Leaver-Fay, A. SwiftLib: Rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res.* **43**, 1–10 (2015).

95. Wolf, E. & Kim, P. S. Combinatorial codons: A computer program to approximate amino acid probabilities with biased nucleotide usage. *Protein Sci.* **8**, 680–688 (2008).

96. Craig, R. A., Lu, J., Luo, J., Shi, L. & Liao, L. Optimizing nucleotide sequence ensembles for combinatorial protein libraries using a genetic algorithm. *Nucleic Acids Res.* **38**, (2009).

97. Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 1–4 (2003).

98. Rousseau, F., Serrano, L. & Schymkowitz, J. W. H. How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.* **355**, 1037–1047 (2006).

99. Niwa, T., Kanamori, T., Ueda, T. & Taguchi, H. Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 8937–8942 (2012).

100. Houben, B., Michiels, E., Ramakers, M., Konstantoulea, K., Louros, N., Verniers, J., der Kant, R., De Vleeschouwer, M., Chicória, N., Vanpoucke, T., Gallardo, R., Schymkowitz, J. & Rousseau, F. Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues. *EMBO J.* **39**, 1–22 (2020).

101. Chen, Y., Ding, F., Nie, H., Serohijos, A. W., Sharma, S., Wilcox, K. C., Yin, S. & Dokholyan, N. V. Protein folding: Then and now. *Arch. Biochem. Biophys.* **469**, 4–19 (2008).

102. Carija, A., Pinheiro, F., Iglesias, V. & Ventura, S. Computational Assessment of Bacterial Protein Structures Indicates a Selection Against Aggregation. *Cells* **8**, (2019).

103. Foy, S. G., Wilson, B. A., Bertram, J., Cordes, M. H. J. & Masel, J. A shift in aggregation avoidance strategy marks a long-term direction to protein evolution. *Genetics* **211**, 1345–1355 (2019).

104. Kessel, A. & Ben-Tal, N. *Introduction to proteins*. (2011).

105. Calloni, G., Chen, T., Schermann, S. M., Chang, H. C., Genevaux, P., Agostini, F., Tartaglia, G. G., Hayer-Hartl, M. & Hartl, F. U. DnaK Functions as a Central Hub in the E. coli Chaperone Network. *Cell Rep.* **1**, 251–264 (2012).

106. Liberek, K., Marszalek, J., Ang, D., Georgopoulos, C. & Zylicz, M. Escherichia coli DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 2874–2878 (1991).

107. Bukau, B. & Walker, G. C. Cellular defects caused by deletion of the Escherichia coli dnaK gene indicate roles for heat shock protein in normal metabolism. *J. Bacteriol.* **171**, 2337–2346 (1989).

108. Bauer, D., Bauer, D., Bauer, D., Bauer, D., Merz, D. R., Merz, D. R., Merz, D. R., Merz, D. R., Pelz, B., Pelz, B., Pelz, B., Pelz, B., Theisen, K. E., Theisen, K. E., Theisen, K. E., Theisen, K. E., Yacyshyn, G., Yacyshyn, G., Yacyshyn, G., *et al.* Nucleotides regulate the mechanical hierarchy between subdomains of the nucleotide binding domain of the Hsp70 chaperone DnaK. *Proc. Natl. Acad. Sci. U. S. A.* (2015).

109. Schlecht, R., Erbse, A. H., Bukau, B. & Mayer, M. P. Mechanics of Hsp70 chaperones enables differential interaction with client proteins. *Nat. Struct. Mol. Biol.* **18**, 345–351 (2011).

110. Mayer, M. P. Intra-molecular pathways of allosteric control in Hsp70s. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, (2018).

111. Imamoglu, R., Balchin, D., Hayer-Hartl, M. & Hartl, F. U. Bacterial Hsp70 resolves misfolded states and accelerates productive folding of a multi-domain protein. *Nat. Commun.* **11**, (2020).

112. Rebeaud, M. E., Mallik, S., Goloubinoff, P. & Tawfik, D. S. On the evolution of chaperones and co-chaperones and the exponential expansion of proteome complexity. *bioRxiv* (2020) doi:10.1101/2020.06.08.140319.

113. Tokuriki, N. & Tawfik, D. S. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* **459**, (2009).

114. Aguilar-Rodríguez, J., Sabater-Muñoz, B., Montagud-Martínez, R., Berlanga, V., Alvarez-Ponce, D., Wagner, A. & Fares, M. A. The molecular chaperone DnaK is a source of mutational robustness. *Genome Biol. Evol.* **8**, 2979–2991 (2016).

115. Kadibalban, A. S., Bogumil, D., Landan, G. & Dagan, T. DnaK-dependent accelerated evolutionary rate in prokaryotes. *Genome Biol. Evol.* **8**, 1590–1599 (2016).

116. Alvarez-Ponce, D., Aguilar-Rodríguez, J., Fares, M. A. & Papp, B. Molecular Chaperones Accelerate the Evolution of Their Protein Clients in Yeast. *Genome Biol. Evol.* **11**, 2360–2375 (2019).

117. DeForte, S. & Uversky, V. N. Order, disorder, and everything in between. *Molecules* **21**, (2016).

118. Leonardi, R., Zhang, Y. M., Rock, C. O. & Jackowski, S. Coenzyme A: Back in action. *Prog. Lipid Res.* **44**, 125–153 (2005).

119. Obmolova, G., Teplyakov, A., Bonander, N., Eisenstein, E., Howard, A. J. & Gilliland, G. L. Crystal structure of dephospho-coenzyme A kinase from Haemophilus influenzae. *J. Struct. Biol.* **136**, 119–125 (2001).

120. Koonin, E. V., Wolf, Y. I. & Aravind, L. Protein fold recognition using sequence profiles and its application in structural genomics. *Adv. Protein Chem.* **54**, 245–275 (2000).

121. Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G. & Little, E. Determining divergence times of the

major kingdoms of living organisms with a protein clock. *Science (80-.).* **271**, 470–477 (1996).

122. Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.* **317**, 41–72 (2002).

123. Fontana, A., Laureto, P. P. De, Spolaore, B., Frare, E., Picotti, P. & Zambonin, M. Probing protein structure by limited proteolysis *. **51**, 299–321 (2004).

124. Schopper, S., Kahraman, A., Leuenberger, P., Feng, Y., Piazza, I., Müller, O., Boersema, P. J. & Picotti, P. Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nat. Protoc.* **12**, 2391–2410 (2017).

125. Lottenberg, R., Hall, J. A., Blinder, M., Binder, E. P. & Jackson, C. M. The action of thrombin on peptide p-Nitroanilide substrates. Substrate selectivity and examination of hydrolysis under different reaction condtions. *Biochim. Biophys. Acta (BBA)/Protein Struct. Mol.* **742**, 539–557 (1983).

126. Backes, B. J., Harris, J. L., Leonetti, F., Craik, C. S. & Ellman, J. A. Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. *Nat. Biotechnol.* **18**, 187–193 (2000).

127. Le Bonniec, B. F., Myles, T., Johnson, T., Knight, C. G., Tapparelli, C. & Stone, S. R. Characterization of the P2′ and P3′ specificities of thrombin using fluorescence-quenched substrates and mapping of the subsites by mutagenesis. *Biochemistry* **35**, 7114–7122 (1996).

128. Le Bonniec, B. F., MacGillivray, R. T. A. & Esmon, C. T. Thrombin Glu-39 restricts the P’3 specificity to nonacidic residues. *J. Biol. Chem.* **266**, 13796–13803 (1991).

129. Kretz, C. A., Tomberg, K., Van Esbroeck, A., Yee, A. & Ginsburg, D. High throughput protease profiling comprehensively defines active site specificity for thrombin and ADAMTS13. *Sci. Rep.* **8**, 1–14 (2018).

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

131. MR., M. Proteases and protein degradation in Escherichia coli. *Experientia.* **48(2):178-**, 181 (1992).

132. Melnikov, E. E., Andrianova, A. G., Morozkin, A. D., Stepnov, A. A., Makhovskaya, O. V., Botos, I., Gustchina, A., Wlodawer, A. & Rotanova, T. V. Limited proteolysis of E. coli ATP-dependent protease Lon - A unified view of the subunit architecture and characterization of isolated enzyme fragments. *Acta Biochim. Pol.* **55**, 281–296 (2008).

133. Sauer, R. T. & Baker, T. A. AAA+ Proteases: ATP-fueled machines of protein destruction. *Annu. Rev. Biochem.* **80**, 587–612 (2011).

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

135. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).

136. Rebollo, I. R., Sabisz, M., Baeriswyl, V. & Heinis, C. Identification of target-binding peptide

motifs by high-throughput sequencing of phage-selected peptides. *Nucleic Acids Res.* **42**, e169–e169 (2014).

137. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 899–907 (2002).

138. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L.-S. L. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115-9 (2004).

139. Piovesan, D., Tabaro, F., Micetic, I., Necci, M., Quaglia, F., Oldfield, C., Aspromonte, M., Davey, N., Davidovic, R., Dosztanyi, Z. & Elofsson, A. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2016).

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

141. Bendl, J., Stourac, J., Sebestova, E., Vavra, O., Musil, M., Brezovsky, J. & Damborsky, J. HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res.* **44**, W479–W487 (2016).

SCIENTIFIC REPORTS

OPEN

Random protein sequences can form defined secondary structures and are well-tolerated *in vivo*

Received: 7 August 2017
Accepted: 31 October 2017
Published online: 13 November 2017

Vyacheslav Tretyachenko[1,2], Jiří Vymětal[1,2], Lucie Bednárová[2], Vladimír Kopecký Jr.[3], Kateřina Hofbauerová[3], Helena Jindrová[1,2], Martin Hubálek[2], Radko Souček[2], Jan Konvalinka[1,2], Jiří Vondrášek[2] & Klára Hlouchová[1,2]

The protein sequences found in nature represent a tiny fraction of the potential sequences that could be constructed from the 20-amino-acid alphabet. To help define the properties that shaped proteins to stand out from the space of possible alternatives, we conducted a systematic computational and experimental exploration of random (unevolved) sequences in comparison with biological proteins. In our study, combinations of secondary structure, disorder, and aggregation predictions are accompanied by experimental characterization of selected proteins. We found that the overall secondary structure and physicochemical properties of random and biological sequences are very similar. Moreover, random sequences can be well-tolerated by living cells. Contrary to early hypotheses about the toxicity of random and disordered proteins, we found that random sequences with high disorder have low aggregation propensity (unlike random sequences with high structural content) and were particularly well-tolerated. This direct structure content/aggregation propensity dependence differentiates random and biological proteins. Our study indicates that while random sequences can be both structured and disordered, the properties of the latter make them better suited as progenitors (in both *in vivo* and *in vitro* settings) for further evolution of complex, soluble, three-dimensional scaffolds that can perform specific biochemical tasks.

The proteinogenic amino acid alphabet has remained largely unchanged during the past ~3 billion years of astonishing evolutionary diversification. The 20 amino acid building blocks could be combined to construct a plethora of polypeptides, yet only a fraction of potential sequences are found in life on Earth[1,2]. For example, 10^{130} possible sequences for a 100-residue polypeptide can be formed from the canonical alphabet, but the number of existing proteins is estimated to be at most 10^{15} [1].

It appears that a finite, relatively small library of protein domains has evolved[3–5]. Structural classification databases (such as SCOP and CATH) have amassed ~1,500 different domain families that account for more than 70% of genomic sequences[6–8]. The prevailing assumption is that once evolution arrived at a set of stable protein folds, evolutionary pressure was dominated by functional constraints[9]. In most cases, a protein's structure determines its functional properties. This raises the question of whether a defined secondary or tertiary structure is a unique property of the sequences found in nature, or whether random sequences also have the potential to form defined structures. Understanding how the structural potential of natural protein sequences differs from that of sequences not subjected to billions of years of evolutionary constraints could provide insights into evolutionary history.

Contrary to early assumptions, a few recent studies suggest that there are unknown functional folds outside the natural protein space, but estimates of their frequency differ[10–12]. Systematic characterization of the folding potential of random sequences has been attempted using tertiary structure prediction algorithms such as Rosetta, but parallel studies questioned the reliability of these algorithms for random sequences unrelated to those found in nature[13,14]. In their Rosetta *ab initio* study, Minervini *et al.* reported that random sequences (with equal relative content of each amino acid) have higher α-helical content (by nearly 10%) and lower β-sheet content than biological sequences. Using a single predictor of secondary structure occurrence, Yu *et al.* recently reported an opposite

[1]Department of Biochemistry, Faculty of Science, Charles University, Hlavova 2030, 128 00, Prague 2, Czech Republic. [2]Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Flemingovo náměstí 2, 166 10, Prague 6, Czech Republic. [3]Institute of Physics, Faculty of Mathematics and Physics, Charles University, Ke Karlovu 5, 121 16, Prague 2, Czech Republic. Correspondence and requests for materials should be addressed to K.H. (email: klara.hlouchova@natur.cuni.cz)

conclusion of the α-helical/β-sheet preference, although they used nearly the same input parameters as Minervini *et al.*[15]. Besides a random dataset with equal relative content of each amino acid, Yu *et al.* additionally included a random dataset with natural occurrence of amino acids, both reporting a comparable distribution of secondary structure content. A few experimental studies have used random 50- to 80-residue sequences to assess secondary structure outside the natural protein space, but these studies had to rely on relatively sparse sampling[2,16,17]. The researchers estimated that compact folding is a property for 5–20% of random sequences[16,17]. Taken together, all of these studies agree that formation of secondary and tertiary structures seem to be general features of polypeptide chains. However, there is a clear lack of correlation among the available bioinformatics/experimental studies, making it difficult to draw conclusions about protein structure evolution.

Here, we present a systematic computational and experimental exploration of the amino acid alphabet and the structural and biophysical consequences of random sequence formation. We generated an *in silico* library (10^4 sequences) of 100-amino-acid proteins and evaluated the occurrence of secondary structure by 5 different prediction algorithms, comparing the properties of random polypeptides with those of natural proteins. Next, we selected 3 × 15 candidate proteins from the library based on their predicted properties (high, low, or random secondary structure occurrence) and experimentally characterized the individual proteins. Because they stem from identical input parameters, the outcomes of these two approaches can be directly compared, allowing us to assess the prediction algorithm accuracy when applied to the unevolved sequence space.

Results and Discussion

Frequencies of secondary structure motifs are similar in random sequences and biological proteins. We used numerous bioinformatic predictors of secondary structure and protein disorder to compare four polypeptide libraries: (A) **random** sequences in which the ratios of individual amino acids reflect those found in natural proteins, (B) fragments of natural proteins from the TOP8000 database of non-redundant structurally characterized proteins extracted from the **PDB** database, (C) a selection of fragments of natural proteins from the **Uni**Prot database, and (D) fragments of natural intrinsically disordered proteins (IDPs) from the **Dis**Prot database[18–20]. The four libraries each comprise 10^4 100-residue sequences (the predictions were performed with the same outcome also with 109-residue sequences including additional residues that were added for the purpose of recombinant expression). Additionally, we investigated the similarity of the random and characterized protein sequences. Only low-significant matches were found by BLAST method for the whole set (Fig. S1) as well as for sequences chosen for experimental characterization (Table S1).

According to statistical analyses of the bioinformatic predictions, both the overall occurrence of secondary structure and the distribution of motifs were comparable for the random and Uni/PDB protein sequence space (Fig. 1 and Table S2). The total occurrence of secondary structure motifs was approximately 5% lower for the random sequence library than for the Uni/PDB natural protein datasets (Table S2). Therefore, our results did not identify any profound differences between random and biological sequences secondary structure formation and thus contrast with previous reports that were based on a single secondary or tertiary structure prediction and which reported statistically significant differences[13,15]. The overall α-helical and β-sheet content predicted by the different algorithms correlate well for all libraries in our study, with an average Pearson correlation coefficient of approximately 0.7 (Table S3).

Experimental sampling of random sequences confirms frequent occurrence of secondary structure and demonstrates tolerance *in vivo*. Based on the bioinformatic analyses, three groups of 15 proteins each were selected from the random sequence library based on the following criteria (Fig. 2):

GROUP 1: (i) High occurrence of predicted secondary structure (samples with both α-helices and β-sheets) and low disorder, (ii) high predicted solubility

GROUP 2: Random selection

GROUP 3: (i) Low occurrence of predicted secondary structure and high disorder, (ii) high predicted solubility

DNA sequences encoding the selected never-born proteins (NBPs) were synthesized so that each NBP has methionine as the N-terminal residue and a 6×His tag at the C-terminus. The bioinformatic predictions were repeated with sequences including the methionine and 6×His tag, to confirm that there were no variations between the predictions of the unmodified and modified sequences. The NBPs were recombinantly expressed in *E. coli* BL21(DE3), and the protein expression level and solubility were analyzed. Out of 15 proteins in each group, the following expressed/soluble ratios were observed: 13/4 in group 1, 8/6 in group 2, and 14/14 in group 3 (Fig. 3). Notably, protein overexpression and solubility in cells increased from group 1 (most structured) to group 3 (least structured). In total, 22 proteins were successfully overexpressed and purified for further characterization. While group 1 proteins have pronounced ellipticity and minima between 205–220 nm in their electronic circular dichroism (ECD) spectra (typical of proteins with high secondary structure content), group 3 ECD spectra indicate proteins with low structural content (Fig. 3). Low concentration of denaturing agent (0.4 M guanidinium hydrochloride) moderately decreases ellipticity for group 1 proteins unlike for group 3 proteins. Conversely, upon addition of a helical structure inducer (50% trifluoroethanol), structure is significantly induced in group 3 spectra, indicative of its original lack of structure unlike for group 1 proteins (Fig. S2). This observation is further supported by the 1D^{1}H NMR spectra (Fig. S3). The overall signal dispersion in the spectra obtained for group 1 proteins (panels A and B) suggests the presence of a hydrophobic core. In addition, the relatively broad signals are indicative of a certain degree of aggregation. The narrow and significantly less dispersed signals observed in the spectra of group 2 and 3 proteins are typical for IDPs.

To compare the performance of bioinformatic predictors and the experimental sampling, the ECD spectra were subjected to hierarchical cluster analysis. Two dominant clusters clearly distinguished groups 1 and 3, the groups that were differentiated based on bioinformatic predictions. The ECD spectra of group 2 (randomly selected from the bioinformatics dataset) were evenly distributed between these major clusters (Fig. 4). Despite

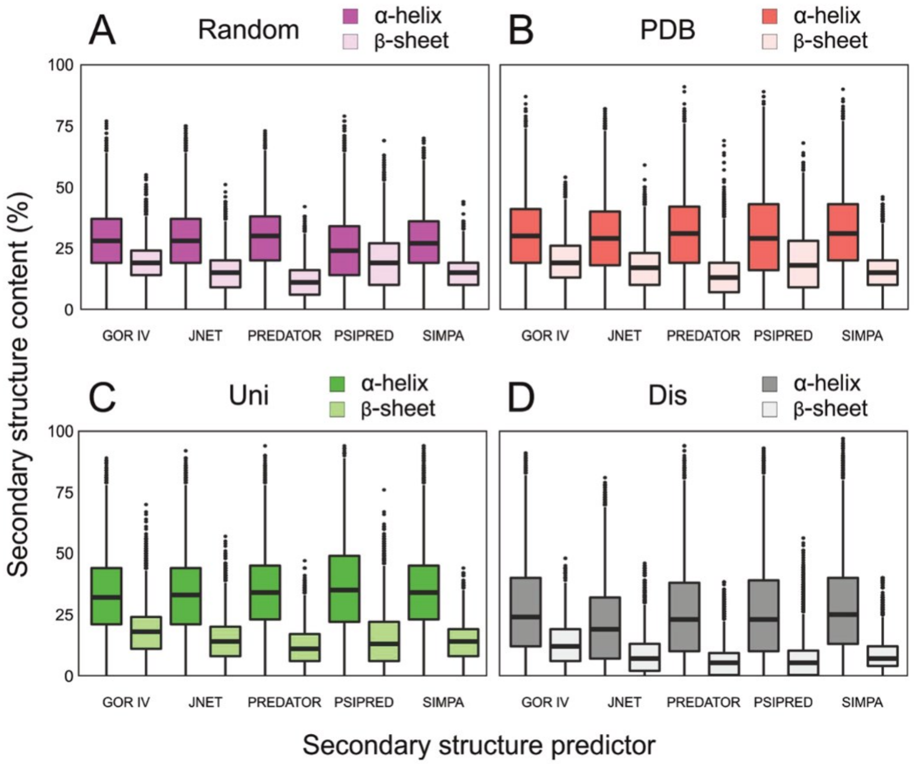


Figure 1. Predictions of secondary structure occurrence in the (**A**) random, (**B**) PDB, (**C**) Uni, and (**D**) Dis libraries. α-helical and β-sheet content determined by five different predictors are shown with statistical information. The center of the box represents the median, and the upper and lower borders represent the 3rd and 1st quartile, respectively. The solid lines illustrate the maximal value and minimal value, excluding outliers, which are shown as dots. The Dis dataset secondary structure prediction is included as a negative reference.

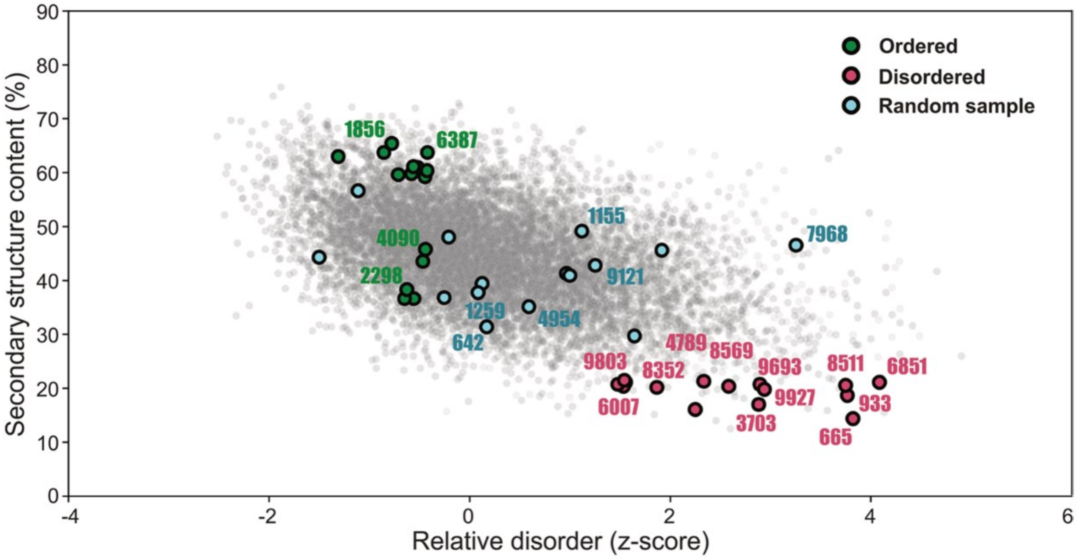


Figure 2. Selection of sequences from the random dataset for experimental characterization. Scatter plot of the secondary structure (y-axis) and disorder prediction (x-axis) in which the selection of group 1 (green), 2 (blue), and 3 (red) proteins for experimental sampling is highlighted as circles (circles with 4-digit codes assigned were successfully purified and characterized). The secondary structure prediction is based on the overall average structure content stemming from five different predictors. The final disorder score is based on the z-value of the average rank of individual sequences from four different disorder predictors.

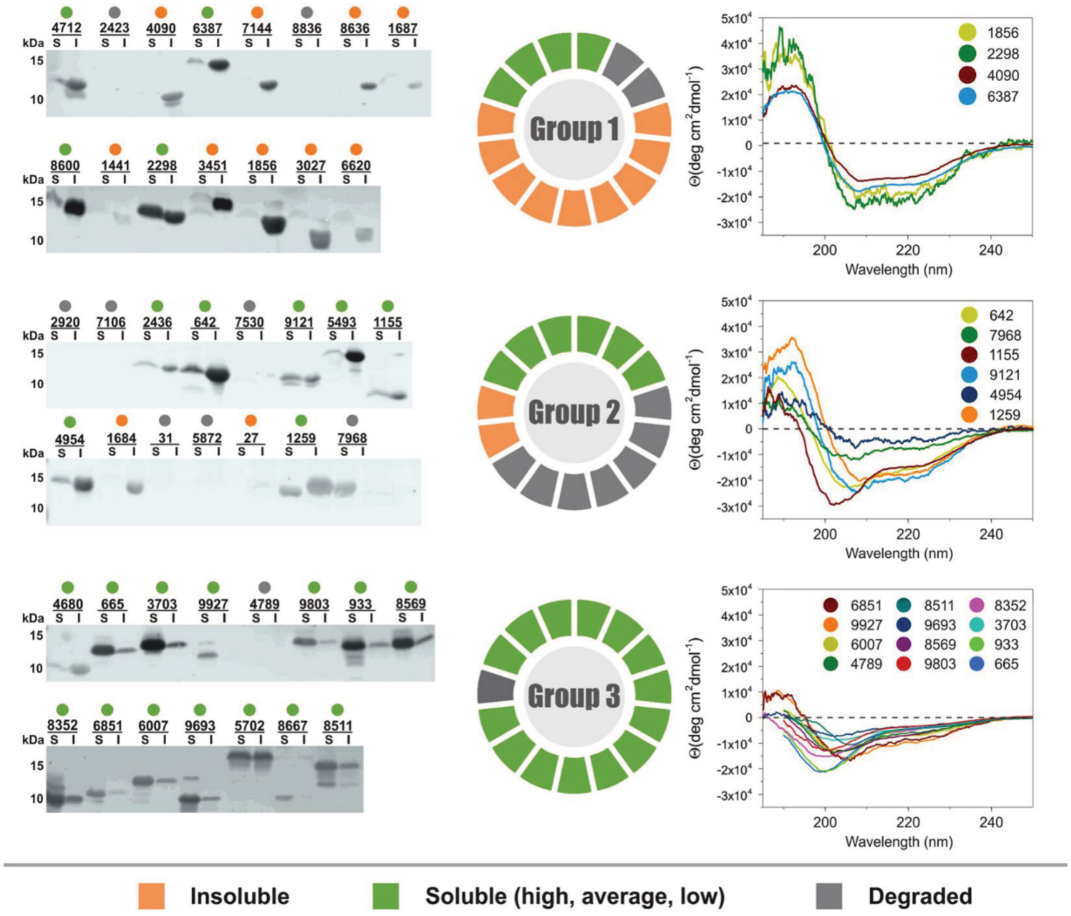


Figure 3. Results of experimental sampling. **Left** – Western blot solubility assay showing never-born proteins (NBPs) expressed in *E. coli* in the insoluble (I) and soluble (S) - including all high, average and low levels of solubility - fractions; **Middle** – a pie graph reporting the expression profiles for group 1–3 NBPs; **Right** – electronic circular dichroism spectra of group 1–3 NBPs that were successfully overexpressed and purified.

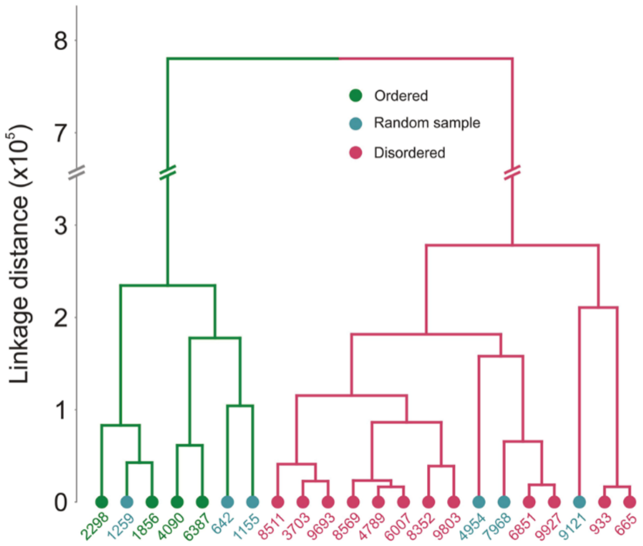


Figure 4. Hierarchical clustering of the electronic circular dichroism spectral data.

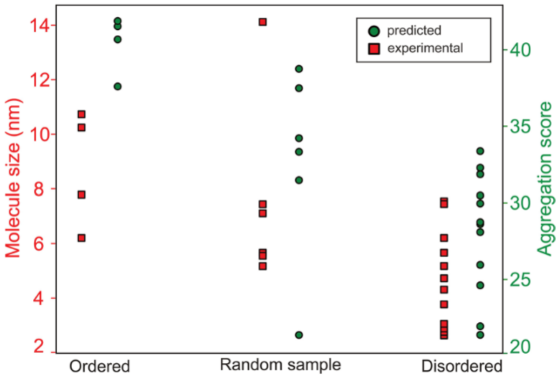


Figure 5. Aggregation propensity of selected never-born proteins. Results of dynamic light scattering analysis (left y-axis and red squares) and aggregation prediction using the ProA-RF algorithm (right y-axis and green circles) for the 22 experimentally sampled group 1–3 proteins.

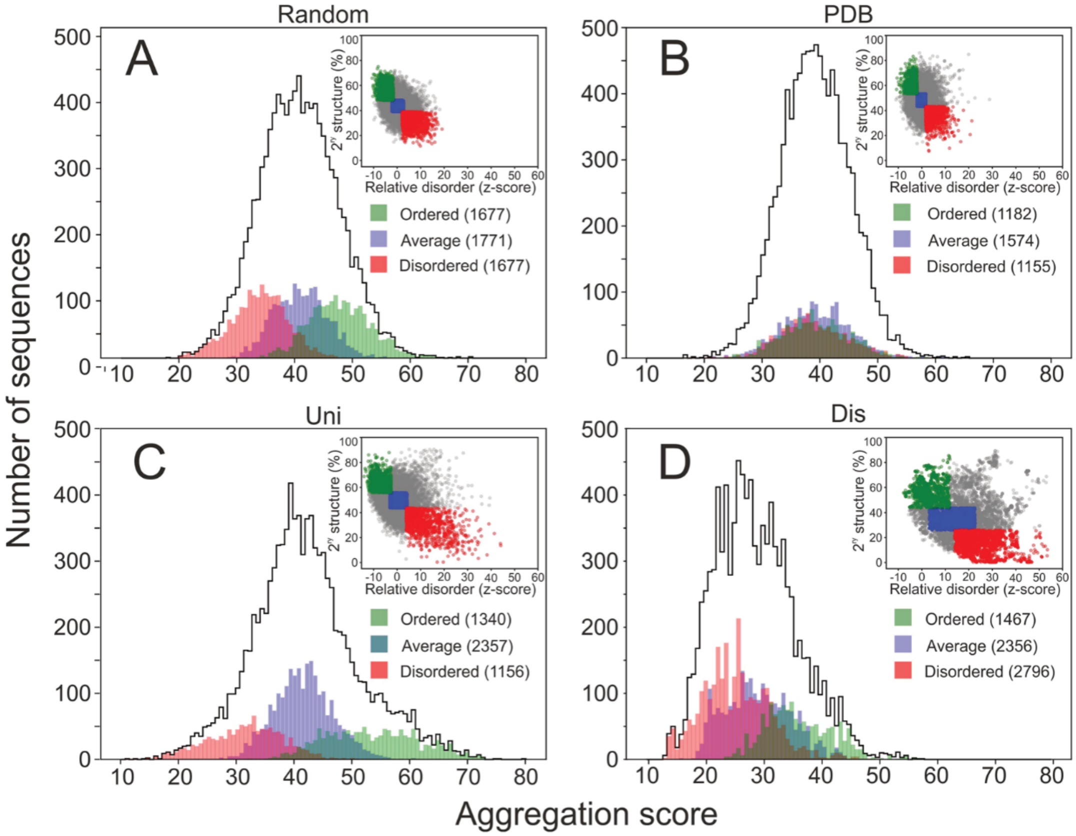


Figure 6. Aggregation propensity of the datasets depending on secondary structure analysis. Distributions of aggregation propensities for the entire random (**A**), PDB (**B**), Uni (**C**) and Dis (**D**) datasets showing the "ordered", "average" and "disordered" subsets (<0.5 SD value, SD ± 0.5 SD, and >0.5 SD values, respectively, from mean values of predicted disorder and secondary structure). The top right corners graphically demonstrate the "ordered", "average" and "disordered" selections from the secondary structure predictions (y-axis, total % of secondary structure content) and disorder (x-axis, z-score units) – equivalent to Fig. 2. Values in brackets for individual subsets in the legend indicate the population numbers.

the small number of proteins sampled experimentally, the experimental data provide reasonable support for the power of bioinformatic predictors when applied to random sequences.

Bioinformatic and experimental analyses highlight the evolutionary potential of random disordered sequences. Dynamic light scattering (DLS) experiments, expression/solubility profiles, and analysis of physicochemical properties relevant to aggregation using the ProA-RF algorithm all suggested that group 1 proteins (most structured) are prone to oligomerization and aggregation. In comparison, along with their better *E. coli* expression profile, group 3 proteins (most unstructured) form smaller particles in solution and do not tend to aggregate (Fig. 5)[21]. The same trend persisted for the entire random dataset when the ProA-RF algorithm was used to predict aggregation of the "ordered" and "disordered" segments of the dataset (Fig. 6A). Therefore, random sequences with higher structural content have a greater tendency to aggregate than those with less structural content.

It has long been suspected that random and IDP-like sequences would tend to aggregate and be toxic to cells. However, a recent computational study reported that random proteins do not have an increased aggregation propensity compared with existing proteins, which is in full accordance with our experimental and bioinformatic results[22]. Several studies have suggested that while natural IDP sequences were expected to be aggregation prone (because of the increased likelihood of exposing aggregation-prone residues to solvent at the absence of a hydrophobic core), it is not so probably as a result of strong anti-aggregation evolutionary pressure[23,24]. Our study demonstrates that low aggregation propensity is in fact a natural property also of random "disordered" sequences.

This trend is less pronounced for natural proteins in the Uni dataset (Fig. 6C) and completely absent for the PDB proteins (the PDB dataset contains proteins that were successfully expressed and structurally studied and therefore represents a biased sample of all extant proteins) (Fig. 6B). To better understand these differences, we performed sequence analysis of each library based on the structural content (ordered, average, and disordered). The ordered and disordered subsets deviated from the mean amino acid composition in the same fashion for each library (Fig. S4). As expected, the control Dis dataset generally deviated in amino acid composition, strengthening the trend observed for the "disordered" subsets of the other datasets. These deviations are in agreement with those observed in previous studies[25]. In addition, for natural proteins, these deviations may reflect a functional purpose. For example, according to ontology analyses (not shown), the "ordered" shoulder shown in Fig. 6C (aggregation score >60) is occupied mostly by membrane proteins. While amino acid composition generally affects the secondary structure content, it determines the aggregation propensity only in unevolved sequences. Evolutionary pressure can work with a given amino acid composition to minimize aggregation and/or prepare the protein for specific conditions, such as the membranous environment. Hypotheses that aggregation-prone sequences are disfavored by evolutionary selection and possible mechanisms for this phenomenon have been described in previous reports[23,26,27].

In summary, random sequences are not significantly different from natural proteins in terms of secondary structure occurrence and overall aggregation properties. Random sequences with low structural content may actually represent advantageous origin points for further evolution into soluble functional proteins, as they are better tolerated *in vivo* and have lower aggregation scores than random sequences with structural content. This is consistent with recent studies reporting that random sequences are often bioactive and can even increase fitness *in vivo*, as well as work suggesting that non-coding DNA translation (one of the hypotheses about *de novo* gene birth) gives rise to highly disordered proteins[28,29]. It is therefore not surprising that structurally dynamic proteins are often encountered during protein-directed evolution experiments in which proteins are selected based on function (rather than structure) from sequence libraries, even if they are originally based on a structured scaffold[12,30]. If proto-proteins arise from random sequences with high structural content, they would likely be disfavored based on their natural physicochemical properties unless their aggregation properties are selected for. Our study provides rationale for this hypothesis on a protein-sequence-space scale.

Methods

Construction and bioinformatic analysis of in silico libraries. Using the composition statistics of the TOP8000 dataset, a library of 10^4 random sequences (100-amino-acid randomized sequences both with and without additional 9 amino acids incorporated for recombinant expression, including an N-terminal methionine and C-terminal hexahistidine tag) was generated *in silico*. Each amino acid at a randomized position was picked randomly from the complete amino acid set with frequencies corresponding to the TOP8000 dataset. All positions in the random sequences were treated independently and without any correlation and additional constraints imposed with respect to the sequential neighbors, position in the sequence and the total composition. In parallel, we constructed three control libraries of 100/109-residue protein fragments from (i) the TOP8000 dataset of structured proteins deposited in the PDB database, (ii) the Uniprot sequence database, and (iii) the DisProt database of IDPs[18–20].

The similarity of the random and the known proteins sequences was assessed by the BLAST method implemented in BLAST + 2.6.0 software package. The NCBI Protein Reference Sequences (Sep 18, 2017; 92 439 966 sequences) were employed as the reference database. The alignments were constructed using default parameters (BLOSUM62 similarity matrix, gap opening and extension penalty 11 and 1, respectively)[31,32].

The secondary structure content was predicted using several methods – GOR4, Jnet, Predator, Simpa, and Psipred[33–37]. In addition, the libraries were analyzed by different protein disorder predictors (Disopred, DisEmbl, VSL2 and IUpred) and empirical indices predicting solubility (CVsol and Gravy)[38–43].

To investigate the aggregation propensity of structurally distinct protein sequences, proteins of high (ordered), low (disordered), and average structure content were selected from the random, PDB, Uni, and Dis datasets. Secondary structure and protein disorder predictions for each sequence were combined with the intention of reducing the false positive rate of individual predictors. Sequence selection into the three groups was based on the following criteria:

	Ordered	Disordered	Average
Secondary structure	$SS_{cont} \geq \mu_{SS} + \frac{1}{2} \cdot \sigma_{SS}$	$SS_{cont} \leq \mu_{SS} - \frac{1}{2} \cdot \sigma_{SS}$	$\mu_{SS} + \frac{1}{2} \cdot \sigma_{SS} \geq \boldsymbol{SS_{cont}} \geq \mu_{SS} - \frac{1}{2} \cdot \sigma_{SS}$
Disorder	$Disorder \leq \mu_{dis} - \frac{1}{2} \cdot \sigma_{dis}$	$Disorder \geq \mu_{dis} + \frac{1}{2} \cdot \sigma_{dis}$	$\mu_{dis} + \frac{1}{2} \cdot \sigma_{dis} \geq \boldsymbol{Disorder} \geq \mu_{dis} - \frac{1}{2} \cdot \sigma_{dis}$

SS_{cont} is the total secondary structure content of the sequence; Disorder is the relative disorder score of the sequence; μ_{SS} and μ_{dis} are the mean values of total secondary structure content and relative disorder score distributions, respectively; and σ_{SS} and σ_{dis} are the standard deviations for those distributions.

Per-residue aggregation scores were generated with the ProA predictor. The final aggregation score for each sequence was obtained by summing all per-residue scores from the ProA output.

Experimental screening of the in silico library. *Protein expression and solubility analysis.* DNA sequences encoding the 3 × 15 selected proteins were codon-optimized for *E. coli* expression and synthesized by Thermo Fisher Scientific, USA. The DNA sequences were subcloned into the pET24a plasmid using *NdeI/XhoI* restriction sites, and the resulting proteins had an additional Met residue at the N-terminus and a Leu, Glu, and 6 × His-tag at the C-terminus (equivalent to the sequences used for bioinformatics analyses controls). The proteins were expressed in 5 mL cultures of *E. coli* BL21 (DE3) for 5 h with 0.5 mM IPTG at 30 °C. The cells were harvested, and pellets were resuspended in 0.5 mL B-PER reagent (Thermo Fisher, USA) supplemented with 5 U/mL benzonase and 100 μg/mL lysozyme. The lysate was centrifuged at 13,000 × g for 10 min at 4 °C, and the supernatant (soluble fraction) was separated from the pellet (insoluble fraction). The pellet was resuspended in 7.5 mL SDS-PAGE sample buffer, and 10 μL soluble fraction and 6 μL insoluble fraction were analyzed by 18% SDS-PAGE. As a control, bacterial pellet from 1 mL pre-induction culture was resuspended in 1 mL sample buffer, and 30 μL of this sample was analyzed in parallel. After electroblotting onto a nitrocellulose membrane, proteins of interest were specifically detected with an anti-His-tag iBody (a synthetic antibody mimetic, present at 5 nM concentration) overnight at 4 °C[44]. The conjugate carries the Cy7.5 fluorophore, which was detected using an Odyssey CLx Imager (LI-COR)[44].

The identities of all expressed proteins were verified using LC-MS following in-gel tryptic digest according to standard procedures[45].

Large-scale expression and purification of selected proteins. Larger-scale expression and purification were attempted for all proteins that expressed in a soluble form. Some proteins that were not soluble in the initial analysis were also overexpressed on a larger scale after further optimization of the expression conditions to solubilize them (such as decreasing the expression temperature). Briefly, proteins were expressed in 0.2–4 L of LB medium for 4–12 h, typically with 0.2 mM IPTG at 20–37 °C, depending on the individual optimal conditions. The bacterial pellets were resuspended in 50 mM phosphate buffer, 30 mM NaCl, 1 mM 2-mercaptoethanol, pH 8, and sonicated 5 × 30 s on ice prior to centrifugation at 20,000 × g for 30 min at 4 °C. The supernatant was subjected to purification on Talon matrix (Clontech, USA) using a gravity-flow arrangement. The eluted fractions (in 50 mM phosphate buffer, 30 mM NaCl, 250 mM imidazole, 1 mM 2-mercaptoethanol, pH 8) were dialyzed thoroughly into 10 mM Tris, 10 mM NaCl, 1 mM TCEP, pH 8, and concentrated to approximately 1 mg/mL before further characterization. Where preliminary DLS measurement suggested a mixture of aggregated and lower-oligomeric species, gel filtration chromatography was used to isolate the lower oligomeric form. Only 22 proteins were purified in sufficient quantity and stability to allow downstream characterization.

Biophysical characterization of selected proteins. Prior to analysis, the identities and molecular weights of purified proteins were confirmed by mass spectrometry. In addition, precise protein concentrations were determined by amino acid analysis using a Biochrom 30 + Series Amino Acid Analyser (Biochrom, UK).

The same protein preparations were used for ECD and DLS measurements. ECD spectra were collected using a Jasco 815 spectrometer (Japan) in the 185–300 nm spectral range using a 0.02 cm cylindrical quartz cell. The experimental setup was as follows: 0.1 nm step resolution, 5 nm/min scanning speed, 16 s response time, and 1 nm spectral band width. After baseline correction, the spectra were expressed as molar ellipticity per residue θ (deg·cm^2·dmol^{-1}). If needed, samples were diluted in 10 mM Tris, 10 mM NaCl, 1 mM TCEP, pH 8. To collect ECD spectra with co-solvents, the samples were diluted to reach the final concentrations of 0.4 M GuHCl and 50% (v/v) TFE.

The ECD spectra of individual proteins were subjected to hierarchical cluster analysis using the Euclidean distance and Ward linkage algorithms in the MATLAB environment (MathWorks, USA).

One dimensional hydrogen NMR spectra were acquired at 25 °C on 850 MHz Bruker Avance III spectrometer equipped with a triple-resonance (15 N/13 C/1 H) cryoprobe. The sample volume was 0.35 ml.

Prior to DLS measurement, the protein samples were centrifuged at 20,000 × g for 30 min at 4 °C. To completely remove dust particles, the samples were immediately filtered using 0.1 μm Ultrafree®-MC centrifugation filters (Millipore, USA). The measurements were performed at 20 °C using a Zetasizer Nano ZS instrument (Malvern Instruments, Great Britain) equipped with an internal 633 nm He-Ne laser. Proteins were measured in a 3 × 3 mm quartz cuvette (internal volume of 40 μL). The results were processed using the original Zetasizer 6.2 Malvern Instruments software (Great Britain).

References

1. Luisi, P. L. The bottle neck: macromolecular sequences in The Emergence of Life, From Chemical Origins to Synthetic Biology, 59–84 (Cambridge University Press, 2010).
2. LaBean, T. H., Butt, T. R., Kauffman, S. A. & Schultes, E. A. Protein folding absent selection. *Genes* **2**, 608–626 (2011).
3. Orengo, C. A. & Thornton, J. M. Protein families and their evolution-a structural perspective. *Annu. Rev. Biochem.* **74**, 867–900 (2005).
4. Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453**, 1262–1265 (2008).
5. Marsh, J. A. & Teichmann, S. A. How do proteins gain new domains? *Genome Biol.* **11**, 126, https://doi.org/10.1186/gb-2010-11-7-126 (2010).
6. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
7. Orengo, C. A. *et al.* CATH - a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
8. Levitt, M. Nature of the protein universe. *Proc. Natl. Acad. Sci. USA* **106**, 11079–11084 (2009).

9. Metpally, R. P. R. and Reddy, B. V. B. Protein structure evolution and the SCOP database in Structural Bioinformatics (ed. Gu, J. and Bourne, P.) 419–732 (Wiley-Blackwell, 2009).
10. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
11. Cossio, P. *et al*. Exploring the universe of protein structures beyond the Protein Data Bank. *PLoS Comput. Biol.* **6**, e1000957, https://doi.org/10.1371/journal.pcbi.1000957 (2010).
12. Chao, F.-A. *et al*. Structure and dynamics of a primordial catalytic fold generated by *in vitro* evolution. *Nat. Chem. Biol.* **9**, 81–83 (2013).
13. Minervini, G. *et al*. Massive non-natural proteins structure prediction using grid technologies. *BMC Bioinformatics* **10**(Suppl 6), S22, https://doi.org/10.1186/1471-2105-10-S6-S22 (2009).
14. Prymula, K. *et al*. *In silico* structural study of random amino acid sequence proteins not present in nature. *Chem. Biodivers.* **6**, 2311–2336 (2009).
15. Yu, J. F. *et al*. Natural protein sequences are more intrinsically disordered than random sequences. *Cell. Mol. Life Sci.* **73**, 2949–2957 (2016).
16. Davidson, A. R. & Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* **91**, 2146–2150 (1994).
17. Chiarabelli, C. *et al*. Investigation of de novo Totally Random Biosequences. *Chem. Biodivers.* **3**, 840–859 (2006).
18. Berman, H. M. *et al*. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
19. Apweiler, R. *et al*. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115–119 (2004).
20. Piovesan, D. *et al*. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).
21. Fang, Y., Gao, S., Tai, D., Middaugh, C. R. & Fang, J. Identification of properties important to protein aggregation using feature selection. *BMC Bioinformatics* **14**, 314, https://doi.org/10.1186/1471-2105-14-314 (2013).
22. Ángyán, A. F., Perczel, A. & Gáspári, Z. Estimating intrinsic structural preferences of *de novo* emerging random-sequence proteins: Is aggregation the main bottleneck? *FEBS Lett.* **586**, 2468–2472 (2012).
23. Naranjo, Y., Pons, M. & Konrat, R. Meta-structure correlation in protein space unveils different selection rules for folded and intrinsically disordered proteins. *Mol. Biosyst.* **8**, 411–416 (2012).
24. de Groot, N. S. *et al*. Evolutionary selection for protein aggregation. *Biochem. Soc. Trans.* **40**, 1032–7 (2012).
25. Uversky, V. N. Paradoxes and wonders of intrinsic disorder: Prevalence of exceptionality. *Intrinsically Disordered Proteins* **3**, e1065029, https://doi.org/10.1080/21690707.2015.1065029 (2015).
26. Chen, Y. & Dokholyan, N. V. Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol. Biol. Evol.* **25**, 1530–3 (2008).
27. Monsellier, E. & Chiti, F. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.* **8**, 737–42 (2007).
28. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 0217, https://doi.org/10.1038/s41559-017-0127 (2017).
29. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146, https://doi.org/10.1038/s41559-017-0146 (2017).
30. Murphy, G. S., Greisman, J. B. & Hecht, M. H. De Novo Proteins with Life-Sustaining Functions Are Structurally Dynamic. *J. Mol. Biol.* **428**, 399–411 (2016).
31. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleaic Acids Res.* **25**, 3389–3402 (1997).
32. Schaffer, A. A. *et al*. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).
33. Levin, J. M., Pascarella, S., Argos, P. & Garnier, J. Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* **6**, 849–854 (1993).
34. Garnier, J., Gibrat, J. F. & Robson, B. GOR secondary structure prediction method version IV. *Methods Enzymol.* **266**, 540–553 (1996).
35. Frishman, D. & Argos, P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* **27**, 329–335 (1997).
36. Cuff, J. A. & Barton, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34**, 508–519 (1999).
37. Jones, T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
38. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
39. Linding, R. *et al*. Protein disorder prediction: Implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).
40. Wilkinson, D. L. & Harrison, R. G. Predicting the solubility of recombinant proteins in Escherichia coli. *Biotechnology* **9**, 443–448 (1991).
41. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
42. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**, 827–839 (2005).
43. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208, https://doi.org/10.1186/1471-2105-7-208 (2006).
44. Šácha, P. *et al*. IBodies: Modular synthetic antibody mimetics based on hydrophilic polymers decorated with functional moieties. *Angew. Chem. Int. Ed. Engl.* **55**, 2356–2360 (2016).
45. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2006).

Acknowledgements

We thank Prof. Stephen Freeland (Univeristy of Maryland, USA) and Michal Svoboda (IOCB, Czech Republic) for their helpful suggestions and discussions. We also thank Dr. Václav Veverka and Dr. Rozálie Hexnerová for the NMR spectra collection and analysis. This work was supported by the Center of Excellence of the Czech Science Foundation (No. P208-12-G016) and by the project "BIOCEV" (CZ.1.05/1.1.00/02.0109).

Author Contributions

K.Hl. and J.Vo. conceived the approach; V.T., J.Vo., J.K. and K.Hl. designed the experiments; J.Vy. and V.T. performed the bioinformatic analyses; V.T. and H.J. performed the biochemical experiments; L.B., V.K., K.Ho., M.H. and R.S. performed the biophysical characterization; and V.T. and K.Hl. wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-15635-8.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bioinformatics, 2020, 1–8
doi: 10.1093/bioinformatics/btaa804
Advance Access Publication Date: 21 September 2020
Original Paper

Structural bioinformatics

CoLiDe: Combinatorial Library Design tool for probing protein sequence space

Vyacheslav Tretyachenko[1,2], Václav Voráček[3,*], Radko Souček[4], Kosuke Fujishima[5] and Klára Hlouchová [1,4,*]

[1]Department of Cell Biology, Faculty of Science, Charles University, Biocev, Prague, Czech Republic, [2]Department of Biochemistry, Faculty of Science, Charles University, 128 00 Prague 2, Czech Republic, [3]Department of Cybernetics, Center for Machine Perception, Faculty of Electrical Engineering, Czech Technical University in Prague, 166 27 Prague, Czech Republic, [4]Institute of Organic Chemistry and Biochemistry IOCB Research Centre & Gilead Sciences, Academy of Sciences of the Czech Republic, 166 10 Prague, Czech Republic and [5]Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo 1528550, Japan

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on June 2, 2020; revised on July 28, 2020; editorial decision on August 14, 2020; accepted on September 7, 2020

Abstract

Motivation: Current techniques of protein engineering focus mostly on re-designing small targeted regions or defined structural scaffolds rather than constructing combinatorial libraries of versatile compositions and lengths. This is a missed opportunity because combinatorial libraries are emerging as a vital source of novel functional proteins and are of interest in diverse research areas.

Results: Here, we present a computational tool for Combinatorial Library Design (CoLiDe) offering precise control over protein sequence composition, length and diversity. The algorithm uses evolutionary approach to provide solutions to combinatorial libraries of degenerate DNA templates. We demonstrate its performance and precision using four different input alphabet distribution on different sequence lengths. In addition, a model design and experimental pipeline for protein library expression and purification is presented, providing a proof-of-concept that our protocol can be used to prepare purified protein library samples of up to 10^{11}–10^{12} unique sequences. CoLiDe presents a composition-centric approach to protein design towards different functional phenomena.

Availabilityand implementation: CoLiDe is implemented in Python and freely available at https://github.com/voracva1/CoLiDe.

Contact: klara.hlouchova@natur.cuni.cz or voracva1@fel.cvut.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Considering the vastness of the potential protein sequence space, naturally occurring proteins are constructed from a small number of coding sequences that arrange into a limited number of structural folds. While there are 20^{100} possible combinations for the design of a 100-amino-acid protein within the canonical amino acid alphabet, only $\sim 10^{15}$ sequences encode all proteins on Earth (Luisi, 2006). Furthermore, these sequences are estimated to fold into only ~2000 distinct topologies (Govindarajan *et al.*, 1999). These observations raise numerous questions in the fields of biotechnology, synthetic biology and evolutionary biology: How easily can a useful sequence be encountered in the unexplored sequence space? Are there protein folds and functions outside those formed by the natural sequence pool?

Several recent studies have started providing answers to these questions. Both secondary and tertiary structures seem to be abundant in completely random sequences (Chiarabelli *et al.*, 2006; Davidson and Sauer, 1994; LaBean *et al.*, 2011; Tretyachenko *et al.*, 2017). Novel folds and functions have been encountered in random and semi-random sequence libraries, and some researchers argue that protein function may be discovered by entirely stochastic means (Chao *et al.*, 2013; Donnelly *et al.*, 2018; Fisher *et al.*, 2011; Keefe and Szostak, 2001; Ravarani *et al.*, 2018). In addition, the bioactivity of and cellular response to random sequences has been actively discussed in association with de novo gene birth (Bornberg-Bauer and Heames, 2019; Neme *et al.*, 2017). While it seems that protein structure and function can be encountered in random sequence space, different biological functions have been associated with specific amino acid composition and hence physicochemical properties. For example, positively charged and aromatic amino acids are known to promote protein–RNA interaction, evolutionary early amino acids promote solubility and trends in amino acid composition have been related to phenomena such as protein disorder and liquid–liquid phase separation (Blanco *et al.*, 2018; Doi *et al.*, 2005; Newton *et al.*, 2019; Wang *et al.*, 2018; Vymětal *et al.*, 2019). Local residue composition is apparently what makes natural sequences stand out from randomness (Weidmann *et al.*, 2019). Overall, these studies highlight the importance of developing tools to probe the protein sequence space in a rational way.

Several approaches to constructing synthetic protein sequence libraries have been developed. The simplest is direct chemical synthesis of a peptide from amino acid precursors but has major restrictions in sequence length and conformational biases [reviewed in Jaradat (2018)]. Another approach is based on construction of a degenerate DNA template with subsequent expression. The template can be designed either using triplet codon as the minimal unit, where pre-synthesized triplets are linked together, or at the single nucleotide level. Although the former method can provide a library with unbiased amino acid distribution at each template position, the cost of the trinucleotide phosphoramidite precursors limits its widespread adoption in laboratory practice (Virnekas *et al.*, 1994). On the other hand, template synthesis at the nucleotide level is economically feasible and is offered by multiple commercial oligonucleotide synthesis companies. Using this approach, random libraries have been constructed from simple repeat of frequently used degenerate codons, such as NNN and NNK. The major drawback of NNN/NNK method for protein engineering is its high level of degeneracy (NNK codes 20 amino acids via 32 different codons). An elegant solution to reduce the degeneracy introduced by Kille *et al.* combines three degenerate codons in a vertical way to cover all 20 amino acids using 22 codons (so-called '22c-trick') without an introduction of STOP codons (Kille *et al.*, 2013). Nevertheless, this solution is effective only when screening a few positions because of an increased cost of oligonucleotide synthesis (mere three mutagenized positions would demand 3^3=27 separate oligonucleotides) and the experimental effort during template assembly. Both of these methods are focused on producing the highest mutational coverage without any attention to amino acid distribution of the mutant library.

While several computational algorithms for library design exist, they have been optimized to introduce as few degenerate codons as possible (Jacobs *et al.*, 2015; Shimko *et al.*, 2020; Tang *et al.*, 2012). An optimal solution to amino acid distribution approximation by combinations of degenerate codons was recently introduced in SwiftLib and DeCoDe algorithms (Jacobs *et al.*, 2015; Shimko *et al.*, 2020). Both produce compact combinatorial libraries by as few degenerate codons as possible while DeCoDe implements complex patterns of covariation into the library design (Shimko *et al.*, 2020). Degenerate codon positions consist of nucleotide mixtures at equimolar ratios where more than one nucleotide is found at a single position. An alternative approach is represented by use of spiked codons where nucleotides can be represented by variable ratios. Mapping of amino acid distribution into a single spiked codon was implemented by Wolf *et al.* and Craig *et al.* via numerical optimization and genetic algorithms. Unfortunately neither of these algorithms is publicly available (Craig *et al.*, 2009; Wolf and Kim, 2008). Although these tools are particularly useful for site-specific randomization strategies, there remains a missed opportunity for the overall design of protein libraries. Specifically, the formation of combinatorial segments of versatile length with a desired amino acid composition would benefit synthetic biology practitioners.

Here, we present a combinatorial library design tool (CoLiDe) for the DNA template design of versatile protein libraries. CoLiDe aids in construction of libraries with specific amino acid distributions and lengths, i.e. optimization of the overall amino acid composition. Such libraries are notably in demand for investigating phenomena that are principally related to amino acid composition—protein liquid–liquid phase separation (Wang *et al.*, 2018), intrinsic protein disorder (Vymětal *et al.*, 2019), spatial protein localization in vivo (Cedano *et al.*, 1997), protein degradation half-life in the cellular milieu and chain elongation rate during ribosomal synthesis (Guruprasad *et al.*, 1990; Riba *et al.*, 2018). In addition, our algorithm allows for incorporation of spiked trinucleotides (i.e. with variable nucleotide composition for single position) and removal of specific codons, such as for codon reassignment and incorporation of unnatural amino acids (Liu and Schultz, 2010).

As a proof-of-concept, we demonstrate the use of CoLiDe by construction of a combinatorial protein library of 33 amino acids in length and composed of a 10 amino acid alphabet (A, S, D, G, L, E, T, I, P and V). Total amino acid composition of the library and therefore each protein sequence was specified using the CoLiDe input option. Moreover, CoLiDe can be used to upgrade currently available DNA block shuffling methods to prepare combinatorial libraries that are hundreds of amino acids in length.

2 Materials and methods

2.1 CoLiDe algorithm

2.1.1 Basic definitions

The following procedure addresses problem-solving with spiked codons (degenerate codons with variable nucleotide composition). If the domain is restricted to degenerate codons, the procedure differs slightly, as noted below. We considered spiked codon to be a 12-tuple concatenated from 4-tuples representing each degenerated position of the triplet:

$$(T_1, C_1, A_1, G_1, T_2, C_2, A_2, G_2, T_3, C_3, A_3, G_3)$$

satisfying

$$\forall i \in \{1, 2, 3\} : T_i + C_i + A_i + G_i = 1$$
$$\forall i \in \{1, 2, 3\} \forall S \in \{T, C, A, G\} : S_i \geq 0.$$

We also introduced a 12-tuple base-codon term:

$$(T_1, C_1, A_1, G_1, T_2, C_2, A_2, G_2, T_3, C_3, A_3, G_3)$$

satisfying

$$\forall i \in \{1, 2, 3\} : T_i + C_i + A_i + G_i \geq 1$$
$$\forall i \in \{1, 2, 3\} \forall S \in \{T, C, A, G\} : S_i \in \{0, 1\}.$$

Base-codons serve as templates for codons. For example, the codon NNS can be represented by the 12-tuple (1,1,1,1,1,1,1,1,0,1,0,1), meaning that the first two positions can include all four bases and the last position is restricted to C or G only. By defining base-codon $\boldsymbol{b}$, a spiked codon can be obtained by replacing 1's in $\boldsymbol{b}$ with non-zero numbers. Note that in cases of restriction to degenerate codons, there is one-to-one mapping between degenerate codons and base-codons.

The optimization problem can be formulated as follows: given amino acid sequence length $\boldsymbol{l}$; desired amino acid distribution $\boldsymbol{D}$, which is a vector of 21 non-negative numbers summing up to 1, one number for each amino acid; a set of forbidden codons $\boldsymbol{F}$; and a distance function dist, find a multiset $\boldsymbol{M}$ cardinality $\boldsymbol{l}$ of codons, minimizing $dist(\boldsymbol{D}, \boldsymbol{M})$, subject to $\forall \boldsymbol{m} \in \boldsymbol{M} \forall \boldsymbol{f} \in \boldsymbol{F} \exists p : \boldsymbol{f}_p \neq 0 \Rightarrow \boldsymbol{m}_p = 0$, where $\boldsymbol{f}_p$ is an element of $\boldsymbol{f}$ on position $\boldsymbol{p}$. This condition guarantees that there are no forbidden codons in $\boldsymbol{M}$.

Every codon encodes a distribution of amino acids. Hence, $\boldsymbol{M}$ representing a multiset of degenerate codons, can be considered as a mixture distribution of amino acids encoded by its codons. The closer the mixture distribution encoded by $\boldsymbol{M}$ is to $\boldsymbol{D}$, the smaller $dist(\boldsymbol{D}, \boldsymbol{M})$ should be. We defined $\boldsymbol{D}$ as a vector in $\mathbb{R}^{21}$, so that we could use a norm to measure the distance between two distributions. Common norms include the $\boldsymbol{L}^1$ norm, which is a sum of absolute values of elements, and the $\boldsymbol{L}^2$ norm, which is a square root of the sum of squares of elements. As square root is a strictly increasing function, minimizing the square root of a sum of squares and minimizing a sum of squares yield the same optimal argument. The third common norm is the $\boldsymbol{L}^\infty$ norm, which is the greatest absolute value of elements. We used the $\boldsymbol{L}^2$ norm in our implementation, as it penalizes large differences considerably but is permissive for slight deviations.In the first step, valid base-codons are generated. There

Algorithm

We present the base implementation of the CoLiDe algorithm as a pseudocode:

```
1. $BC \leftarrow$ generate valid base-codons
2. $M \leftarrow \varnothing$
3. For $i = 1$ to $l$:
   a. $bc \leftarrow$ random element from $BC$
   b. $c \leftarrow$ make random codon from $bc$
   c. $M \leftarrow M \cup \{c\}$
4. $rejected \leftarrow 0$
5. While $rejected < 1000 \cdot \mathrm{l}$:
   a. $bc \leftarrow$ random element from $BC$
   b. $c \leftarrow$ make random codon from $bc$
   c. $\mathrm{d}_{\mathrm{old}} \leftarrow dist(D, M)$
   d. $M2 \leftarrow M \cup \{c\} \setminus$ (random element from $M$)
   e. $d_{\mathrm{new}} \leftarrow dist(D, M2)$
   f. If $d_{\mathrm{new}} < d_{\mathrm{old}}$
      i. $M \leftarrow M2$
      ii. $rejected \leftarrow 0$
      Else
      i. $rejected \leftarrow rejected + 1$
6. Output $M$
```

are 3 independent sequences in base-codon ($T_i, C_i, A_i, G_i, i \in \{1, 2, 3\}$), and every sequence is an arbitrary binary string of length 4, excluding string 0000. There are 16 – 1 such strings, so the number of base codons is $(16 - 1)^3$=3375. Along the fact that there are at most 64 forbidden codons, the time needed to execute this step is negligible with any reasonable implementation.

In the third step, filling multiset $\boldsymbol{M}$ with random codons yields an initial result.

In the fifth step, the optimization is performed. Once per loop, a random codon is generated, and an attempt is made to replace a random codon in $\boldsymbol{M}$ with this codon. If the objective improves, the change is accepted; otherwise, it is rejected. The algorithm works reasonably well and reasonably quickly (visualization of results is many times slower than the algorithm itself). The base algorithm can be easily modified, because *dist* can be chosen arbitrarily. In our implementation, *dist* is chosen as the L^2 norm of the vector of differences between $\boldsymbol{D}$ and the distribution of amino acids encoded by codons of $\boldsymbol{M}$. This problem also could be formulated as a quadratic programming task, but it would be difficult or even impossible to add new requirements to the result. The ability of the algorithm to be easily extended to new problems offers flexibility.

2.2 Library construction

2.2.1 Preparation of DNA and RNA templates

A degenerate ssDNA of 197 bases was synthesized by Integrated DNA Technologies (Supplementary Material Sequences, library). The oligonucleotide was converted to dsDNA by Klenow extension with a 5′ complementary reverse primer (Supplementary Material Sequences, reverse). Annealing of the primer was performed by cooling down a mixture of 2 μM oligonucleotide and primer in the presence of 200 μM dNTPs in buffer NEB1 from 90 to 25°C at a rate of 1°C/min. Total 10 U Klenow polymerase was added to the annealed mixture, and extension step was carried out for 1 h at 37°C followed by polymerase deactivation at 50°C for 15 min. The dsDNA library product was purified with the Monarch® PCR & DNA Cleanup Kit (New England Biolabs) and used for the downstream in vitro transcription, carried out with the Ampliscribe T7-Flash kit (Lucigen) according to the manufacturer's recommendations. The resulting mRNA was purified by ammonium acetate precipitation and dissolved in RNase free water to a final concentration of 3 µg/µl.

2.2.2 cDNA preparation for high-throughput sequencing (HTS)

Complementary DNA (cDNA) was prepared from 1 µg transcribed mRNA. cDNA was synthesized according to the SuperScript IV (Thermo Fisher Scientific) instruction manual using reverse primer (Supplementary Material Sequences, reverse) and 20 μl reverse transcribed product was further amplified with Q5 DNA polymerase (New England Biolabs) in a 100-µl reaction volume for 11 amplification cycles with a primer annealing temperature of 68°C.

2.2.3 Protein expression and purification for amino acid analysis and mass spectrometry

The protein library was prepared in a PUREfrex 2.0 (GeneFrontier Corporation) cell-free protein expression system. The reaction was prepared according to the manufacturer's recommendations, supplemented with 0.05% Triton X-100 (v/v), and initiated by addition of 3 µg library mRNA. Protein expression was conducted for 4 h at 30°C. The reaction was diluted 10 times with guanidine denaturation buffer (6 M guanidine hydrochloride, 100 mM sodium phosphate, 500 mM NaCl, 0.05% Triton X-100, pH 8) and incubated with 4 µl TALON affinity chromatography resin (Clontech) for 12 h at 25°C. The resin was washed twice with urea denaturation buffer (8 M urea, 100 mM sodium phosphate, 500 mM NaCl, 0.05% Triton X-100, pH 8) and twice with distilled water supplemented with 0.05% Triton X-100. The library was eluted by boiling the affinity matrix in 50 µl of 2% (w/v) aqueous SDS. Eluted fractions were purified from SDS by addition of 5× volumes of ice-cold acetone. The precipitates were centrifuged, washed with 100% acetone and air-dried.

2.2.4 Preparation of libraries for HTS and data analysis

The dsDNA library template was analyzed by HTS with an Illumina MiSeq. Prior to sequencing the library preparation, quantification was carried out on a Quantus$^{\mathrm{TM}}$ Fluorometer (Promega). A total of 100 ng of DNA sample was used as an input for library preparation with the NEBNext Ultra II DNA Library Prep kit (New England Biolabs) with AMPure XP purification beads (Beckman Coulter). The length of the prepared library was determined with an Agilent 2100 Bioanalyzer (Agilent Technologies) and quantified with a Quantus Fluorometer (Promega). Samples were sequenced on a MiSeq Illumina platform using the Miseq Reagent Kit v2 for 500 cycles (2 × 250) in paired-end mode. Raw data was processed with Galaxy platform. Sequence analysis of assembled and filtered paired reads was performed with MatLab scripts developed by the Heinis lab (Afgan *et al.*, 2018; Rebollo *et al.*, 2014).

2.2.5 Amino acid analysis and mass spectrometry

The purified and precipitated library samples were hydrolyzed in 6 M hydrochloric acid at 110°C for 20 h, the hydrolysate was evaporated, and reconstituted with 0.1 M hydrochloric acid containing the internal standard. Amino acid analysis was performed on an Agilent 1260 HPLC (Agilent Technologies) equipped with a fluorescence detector using automated o-phtalaldehyde/2-mercaptopropionic acid (OPA/MPA) derivatization. For mass spectrometry, the purified protein library sample was resuspended in water. The spectrum was collected after addition of 2,5-dihydroxybezoic acid matrix substance (Merck) using an UltrafleXtreme$^{\mathrm{TM}}$ MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Germany) in linear mode.

3 Results and discussion

In this work, we present a computational tool for automated design of combinatorial libraries. CoLiDe uses evolutionary approach to find a satisfactory solution. The algorithm provides a set of degenerate codons which approximate the total amino acid distribution of

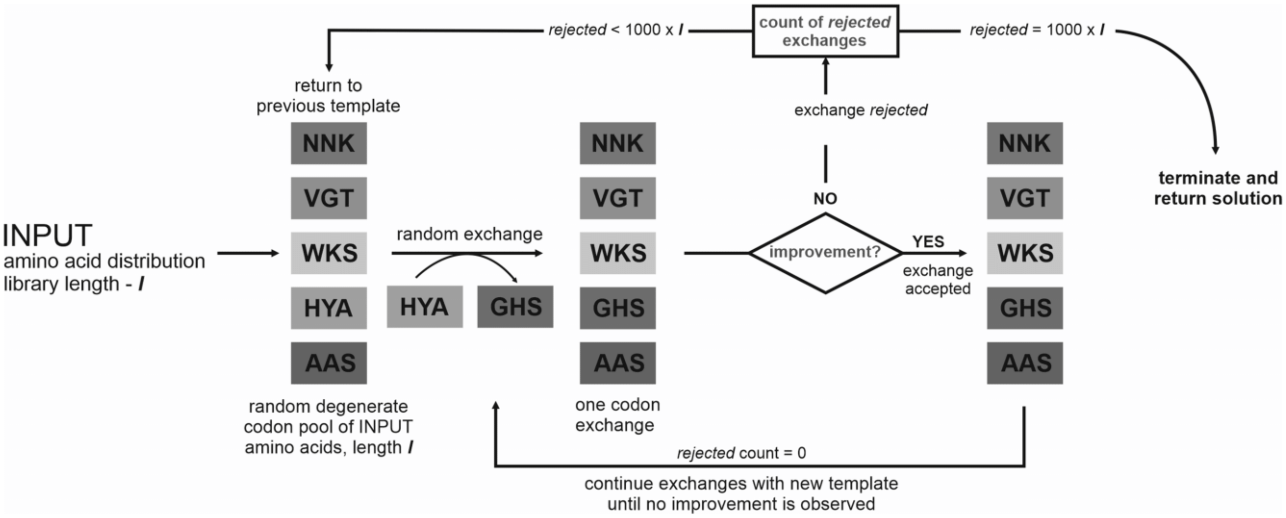


Fig. 1. Outline of the CoLiDe algorithm. Based on the input amino acid distribution and length of the randomized library, at first an unoptimized vector of degenerate codons of given length is generated. Then the vector is optimized by single exchanges of codons until a vector of degenerate codons with minimal distance from the input distribution is obtained

protein without regard to individual degenerate positions in the coding template. The principle of the algorithm is summarized inFigure 1.

Mandatory inputs include library length, amino acid distribution and degenerate codon type (standard or spiked, Supplementary Fig. S1). Other parameters, such as organism-specific codon preference, extent of degeneracy or codon removal/reassignment, also can be specified (Supplementary Fig. S1). Once the input parameters are defined, codons are pre-selected based on the amino acid input from a total pool of 3375 degenerate codons. The codon pre-selection removes undesired amino acid and STOP codons. This step guarantees that the combinatorial library is composed only of input amino acids and will not contain prematurely terminated templates. On the other hand, depending on input distribution, most highly degenerate codons are removed which reduces degeneracy of individual library positions.

Only the pre-selected degenerate codons serve in the subsequent library construction pipeline. The pipeline starts with random sets of degenerate codons of desired library length and follows with random codon exchanges (standard codons) or a shift in nucleotide ratios (spiked codons). Exchanges and shifts are kept within the optimized codon set if the amino acid product comes closer to input distribution (evaluated by mean squared error) and rejected if not. Optimization is finished when repeated changes do not further improve the solution (specifically, after $n = 1000 \times$ [library length] rejected mutations) This threshold was selected after test runs of the optimization path which recorded the rejection rate of mutations and provided satisfactory deviation on all tested distributions (Supplementary Fig. S2A–D). The output of the algorithm is a vector of degenerate codons of given library length. In other words, CoLiDe provides a list of degenerate codons combined randomly into a single oligonucleotide template.

CoLiDe offers a graphical user interface (Supplementary Fig. S1) that aids input of all variables, displays statistics of the optimized solution and allows the user to generate a report as a PDF document. CoLiDe is implemented in Python 3, and the source code is available as open source under MIT license at https://github.com/voracva1/CoLiDe.

3.1 CoLiDe performance analysis

We tested CoLiDe's precision and reproducibility on the following four amino acid distributions: (i) a reduced alphabet used in protein evolution studies to approximate an early version of the genetic code (Solis, 2019), (ii) a functional distribution derived from an analysis of RNA-binding proteins (Blanco *et al.*, 2018), (iii) a natural amino acid distribution from the UniProt database (UniProtKB/Swiss-Prot UniProt release 2019_11) and (iv) a rational selection of a reduced set of amino acids for protein engineering (Murphy *et al.*, 2000) (Fig. 2A–D and Supplementary Table S1). For each amino acid distribution, optimization was performed 10 independent times for library lengths of 5, 10, 15, 20, 40, 60, 80 and 100 amino acids (Fig. 2E–H). CoLiDe was able to reliably spread all the tested distributions on a DNA template of given length.

Mean squared errors in the shortest amino acid libraries ranged from 0.11 to 0.17 between individual alphabets and converged with increasing template length to values around 0.005. Variance in precision between solutions—measured as a coefficient of variation was highest in short libraries, ranging between 10^{-2} and 10^{-3}, and decreased to values around 10^{-5} in longer templates (Supplementary Table S2).

Our results confirmed that the algorithm consistently finds precise solutions to selected input amino acid distributions. The precision of the solution increases and the variance between solutions within each group decreases along with the increase in library template length. With reduced template length, error became dependent on the specific amino acid alphabet. Solutions using spiked codons showed better precision with similar variance within each group (Supplementary Table S2). CoLiDe runtimes were tested on four library templates (Fig. 2A–D) with the template sizes ranging from 5 to 400 degenerate codons. Reported runtimes range from ∼3 to 600 s on Intel i5-8250U laptop (Supplementary Fig. S3).

Diverse degenerate libraries can be produced with other available tools, even though they are designed for construction of different library types. CoLiDe, in contrast to alternative design tools (SwiftLib, DeCoDe), focuses on combinatorial library design without position-specific restraints. Designed libraries are suitable for probing the constrained sequence space rather than for screening small, rationally designed library of protein variants (Jacobs *et al.*, 2015; Shimko *et al.*, 2020). As an example, we compare the solutions for combinatorial libraries provided by degenerate codon optimization algorithm SwiftLib (Jacobs *et al.*, 2015). SwiftLib outputs an optimized set of degenerate codons which cover the provided amino acid variability with as few degenerate codons as possible. Such approach faces difficulty to assure the precision of the distribution when targeting longer regions, whereas that is not the case for CoLiDe (Supplementary Fig. S5). On the other hand, SwiftLib outperforms CoLiDe when very short randomized regions (of 2–3 codons) are calculated (Supplementary Fig. S4). Deviations of ratios of single amino acids are reported in Supplementary Tables S4 and S5. CoLiDe provides a better choice for combinatorial design of longer protein templates provided that overall amino acid distribution

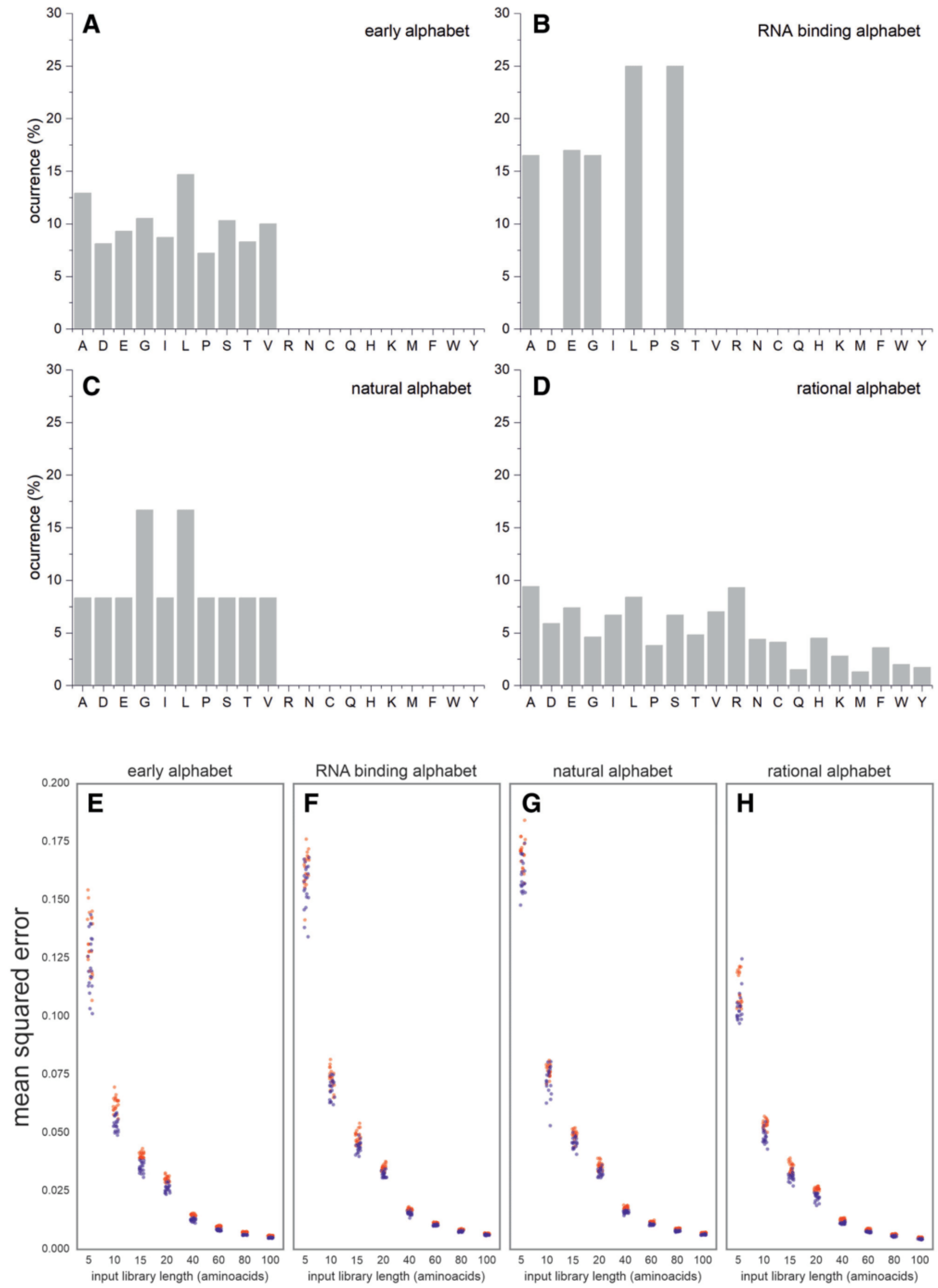


Fig. 2. CoLiDe performance analysis. Amino acid distributions used to benchmark CoLiDe performance (**A–D**) and comparison of solutions generated from each (**E–H**). Each distribution was approximated *via* degenerate (red) and spiked (blue) codons. Solutions were produced in 10 replicates for various library lengths ranging from 5 to 100 amino acids

of sequence is preferred over the specific amino acid variations on predefined positions. Furthermore CoLiDe can be used in protein engineering applications for coarse grained yet computationally efficient vertical design (multiple degenerate oligonucleotides per one tube) of degenerate codons to approximate amino acid distributions in single protein positions, similarly to established deterministic approaches described by Jacobs and coworkers (Jacobs *et al.*, 2015).

3.2 Proof-of-concept experimental library design

To identify general pitfalls and experimental bottlenecks of library preparation, we experimentally evaluated one specific CoLiDe solution from DNA to protein level. A 45 amino acid protein library was prepared with a randomized region of 33 amino acids, following the early alphabet distribution (Fig. 2A). The mean squared error of the randomized region with CoLiDe solution was 0.0022 with an error variance of 0.00011 (Fig. 3). The random 33 codon region was tagged with an 8×H+QH (i.e. octa-His + Gln-His) coding sequence (separated by a two amino acid linker, KS) on the C-terminus for subsequent purification (Supplementary Material, Sequence). The protein coding sequence was embedded into a linear expression cassette, and the library was transcribed as described in Section 2 (Supplementary Fig. S6).

The length of the protein library was selected so that a single commercially synthesized oligonucleotide could be used for the downstream procedure. However, a larger construct could be prepared by DNA shuffling methods as previously described (Cho *et al.*, 2000). Thus, CoLiDe algorithm can also be utilized for the

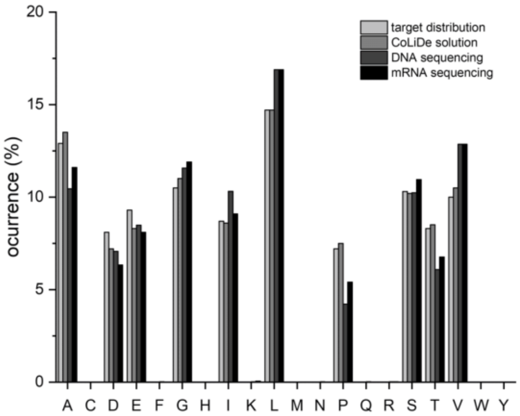


Fig. 3. Comparison of the amino acid distribution of the CoLiDe solution of 33 amino acid long library to its target distribution and the DNA and mRNA templates obtained from the high-throughput sequencing (HTS) data (upon *in silico* translation)

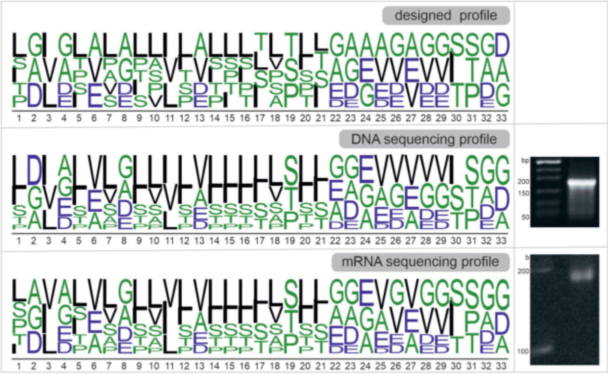


Fig. 4. Preparation and analysis of DNA and RNA libraries. (left) Sequence logos generated *in silico* from the designed template (top), sequenced DNA template (middle) and sequenced reverse-transcribed mRNA (bottom). (right) Agarose gel electrophoresis of dsDNA library template (middle) and urea PAGE analysis of single stranded random library mRNA and (bottom). Polar and small amino acids (G, S, T, P, A) are green, hydrophobic and large amino acids are black (L, V, I) and negatively charged residues (D, E) are blue

construction of random protein libraries with amino acids residues up to several hundreds.

3.3 Construction and characterization of the oligonucleotide library

Nucleotide sequences for degenerate libraries were analyzed on the DNA and mRNA template levels by high-throughput sequencing (HTS). The *in silico* translated amino acid composition (from both the DNA and mRNA templates) showed good agreement with the designed construct (Figs 3 and 4, Supplementary Table S6). While deviations of whole distributions are listed here as mean squared error calculated on (0,1) scale, we plot single amino acid occurrence as percentage of input distribution on (0,100) scale. Deviations between the CoLiDe solution and the in silico translated DNA template were observed in enrichment of valine, leucine and isoleucine (2.9, 2.2 and 1.6%) and depletion of proline, threonine and alanine (3, 2.2 and 2.4%) (Figs 3 and 4, Supplementary Table S6).

Upon analysis of nucleotide frequencies at each position, we found that deviation can be explained by the nucleotide composition bias during the oligonucleotide synthesis and have been confirmed as the current bottleneck by the provider (Supplementary Fig. S7). Statistical analysis of the sequencing data provides a confirmation of library diversity and shows that vast majority (99.9%) of all sequences are unique (Supplementary Table S7). Overall, mean squared error of amino acid distribution of DNA and RNA templates remained to be around ∼0.02 (Supplementary Table S6). Hence, we found that while CoLiDe algorithm can provide low mean squared error for the library design, one should be aware of the nucleotide bias that will be introduced during the oligonucleotide synthesis of highly degenerate DNA oligonucleotides. Such nucleotide composition bias of DNA library depends on each oligonucleotide provider (unpublished observation).

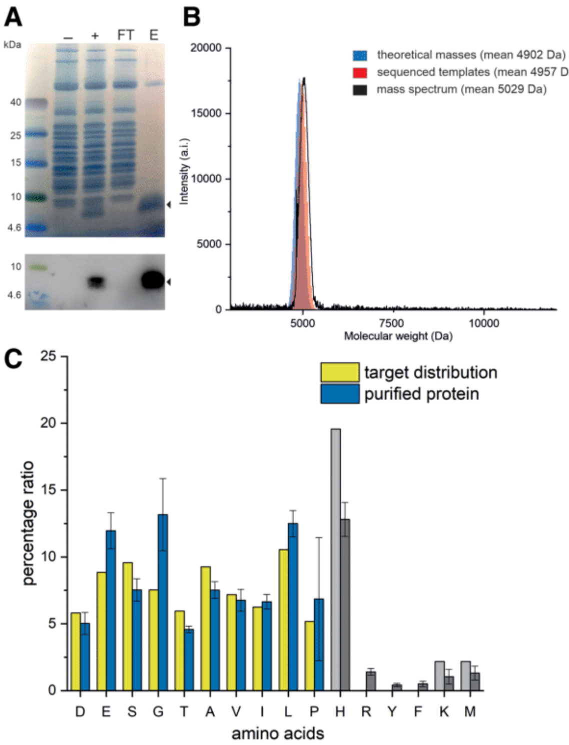


Fig. 5. Preparation and analysis of the protein library. (**A**) SDS-PAGE and Western blot analysis of library expression and purification. The library was expressed in a recombinant cell-free system PUREfrex 2.0. ± stands for cell free fraction without and with expressed library, FT is affinity purification flow through and E is eluted fraction. (**B**) MALDI-TOF MS analysis of the purified library (black) compared with the theoretical mass distribution (blue) and mass distribution calculated from sequenced DNA templates (red). (**C**) Results of amino acid analysis deviations of variable (colored) and constant sequence regions/contaminations (grey) of the expressed and purified protein library in percentage units

3.4 Construction and characterization of the protein library

The combinatorial protein library was expressed using an in vitro translation system and His-tag purified for downstream analysis (Fig. 5A). Expressed proteins were assessed by mass spectrometry (Fig. 5B) and amino acid analysis (Fig. 5C, Supplementary Table S6).

MALDI-TOF mass spectrometry revealed good agreement with expected values. The expected mass distribution was produced by analysis of 600 000 random sequences corresponding to the degenerate DNA template and by in silico translation of 600 000 sequences obtained by HTS of DNA and mRNA templates. The experimental spectrum is represented by normal weight distribution with a mean value of 5029 Da and a standard deviation of 120.6 (Fig. 5B). This is slightly shifted from the mean value of the molecular weight distribution expected from the design (4902 Da), partly as a result of sequence bias during the solid-state oligonucleotide synthesis. However, in silico translation of sequences obtained by HTS (producing a mean molecular weight of 4957 Da) confirms that this explains only part of the shift. This result indicates that the translation and purification steps have introduced additional compositional shift into the protein library. Most notably, the purified protein library is under-represented in alanine, aspartic acid and threonine (by 2–4% from the desired amount) and enriched in glutamic acid and glycine (by $\sim$5% from the input) as assessed by amino acid analysis (Fig. 5C), likely due to their impact on protein solubility and contamination by carry over protein components from the cell-free expression system in the purified library sample (Fig. 5A). While these deviations do not represent a major difference in the overall amino acid ratio profile [amino acid analysis shows an overall of 0.05 mean squared error (Supplementary Table S6)], it is important to be aware of the sequence biases that may be introduced into designed libraries during oligonucleotide synthesis and downstream procedures as a result of the translation and purification process or the physicochemical properties of the expressed proteins themselves.

Currently, there is no satisfactory methodology to analyze the variability of the large protein sequence pool directly. One translation reaction (in a 20 µl volume) is typically primed with 10^{11}–10^{12} different template molecules. Even with the genotype-phenotype linked display methods (i.e. mRNA-display, ribosome display, etc.) number of characterized sequences is limited to the performance of HTS. Because neither DNA library preparation, RNA transcription nor the *in vitro* translation involve sequence amplification, a similar variability of protein sequences is expected after translation. The computational protocol therefore presents a tool for truly effective exploration of the protein sequence space.

4 Conclusions

Here, we present CoLiDe, a novel tool for precise design of combinatorial protein libraries of flexible length and desired amino acid composition. We provide evidence that it performs with minimal error and variance across several different amino acid distributions and lengths. It significantly outperforms SwiftLib (that have been developed for other applications) especially when designing combinatorial libraries longer than $\sim$10 amino acids.

In addition, we present a model protocol for combinatorial library (composed of a 10 amino acid alphabet) preparation by cell-free expression. By monitoring the DNA and mRNA sequence pool during library preparation using HTS, we confirmed the desired variability (99.9% of the sequences representing unique species). While negligible error is detected between the input sequence and the CoLiDe solution, up to 3% deviations of individual amino acid ratios were detected upon in silico translation of the mRNA sequence pool. The error was primarily attributable to nucleotide compositional bias from the synthesis of the starting material.

Using the template mRNA, we expressed and purified a highly variable protein library (represented by a normal weight distribution). To our knowledge, this is the first report of purification of a combinatorial protein library in an amount sufficient for biophysical characterization. The experimental procedure introduced additional detectable shifts among several amino acid compositions (up to 5% deviation), likely occurred during translation and purification steps of the library. Such an error is to be expected and may vary depending on the nature of individual amino acid alphabets. We estimate that 1011–1012 unique protein sequences can be produced in a 20-µl cell-free translation reaction using our protocol.

The design and experimental strategy presented here can be used in combination with vertical library design strategies (i.e. mixing multiple degenerate templates) and DNA shuffling synthesis. This represents a powerful tool for the synthesis of combinatorial protein libraries composed of hundreds of amino acids.

Acknowledgements

The authors thank Prof. Mirko Navara for helpful discussions, Dr. Martin Hubálek for MS analyses, Dr. Hillary Hoffman for language editing and Shota Nishikawa and Hidenori Watanabe for providing the library for comparison.

Funding

This work was supported by the Czech Science Foundation (GA ČR) [17-10438Y]; Human Frontiers Science Program [RGY0074/2019]; Charles University Grant Agency [260572/2020 to V.T.]; Ministry of Education, Youth and Sports of CR [National Sustainability Program II, BIOCEVFAR, LQ1604 to K.H.]; and ELSI-FirstLogic Astrobiology Donation Program to K.F.

Conflict of Interest: none declared.

References

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

Blanco,C. *et al.* (2018) Analysis of evolutionarily independent protein–RNA complexes yields a criterion to evaluate the relevance of prebiotic scenarios. *Curr. Biol.*, **28**, 526–537.e5.

Bornberg-Bauer,E. and Heames,B. (2019) Becoming a de novo gene. *Nat. Ecol. Evol.*, **3**, 524–525.

Cedano,J. *et al.* (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.*, **266**, 594–600.

Chao,F.-A. *et al.* (2013) Structure and dynamics of a primordial catalytic fold generated by in vitro evolution. *Nat. Chem. Biol.*, **9**, 81–83.

Chiarabelli,C. *et al.* (2006) Investigation of de novo Totally Random Biosequences. *Chem. Biodivers.*, **3**, 827–839.

Cho,G. *et al.* (2000) Constructing high complexity synthetic libraries of long ORFs using in vitro selection. *J. Mol. Biol.*, **297**, 309–319.

Craig,R.A. *et al.* (2009) Optimizing nucleotide sequence ensembles for combinatorial protein libraries using a genetic algorithm. *Nucleic Acids Res.*, **38**, 1–9.

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

Doi,N. *et al.* (2005) High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Eng. Des. Sel.*, **18**, 279–284.

Donnelly,A.E. *et al.* (2018) A de novo enzyme catalyzes a life-sustaining reaction in *Escherichia coli*. *Nat. Chem. Biol.*, **14**, 253–255.

Fisher,M. a. *et al.* (2011) De novo designed proteins from a library of artificial sequences function in *Escherichia Coli* and enable cell growth. *PLoS One*, **6**, e15364.

Govindarajan,S. *et al.* (1999) Estimating the total number of protein folds. *Proteins Struct. Funct. Genet.*, **35**, 408–414.

Guruprasad,K. *et al.* (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng. Des. Sel.*, **4**, 155–161.

Jacobs,T.M. *et al.* (2015) SwiftLib: rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res.*, **43**, 1–10.

Jaradat,D.M.M. (2018) Thirteen decades of peptide synthesis: key developments in solid phase peptide synthesis and amide bond formation utilized in peptide ligation. *Amino Acids*, **50**, 39–68.

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

Kille,S. *et al.* (2013) Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol.*, **2**, 83–92.

Labean,T.H. *et al.* (2011) Protein folding absent selection. *Genes (Basel)*, **2**, 608–626.

Liu,C.C. and Schultz,P.G. (2010) Adding new chemistries to the genetic code. *Annu. Rev. Biochem.*, **79**, 413–444.

Luisi,P.L. (2006) *The Emergence of Life: From Chemical Origins to Synthetic Biology*, 1st edn. Cambridge University Press, Cambridge, UK.

Murphy,L.R. *et al.* (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng. Des. Sel.*, **13**, 149–152.

Neme,R. *et al.* (2017) Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.*, **1**, 1–7.

Newton,M.S. *et al.* (2019) Genetic code evolution investigated through the synthesis and characterisation of proteins from reduced-alphabet libraries. *ChemBioChem*, **20**, 846–856.

Ravarani,C.N. *et al.* (2018) High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol. Syst. Biol.*, **14, e8190**.

Rebollo,I.R. *et al.* (2014) Identification of target-binding peptide motifs by high-throughput sequencing of phage-selected peptides. *Nucleic Acids Res.*, **42**, e169.

Riba,A. *et al.* (2018) Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. Proceedings of the National Academy of Sciences. 116, 15023–15032.

Shimko,T.C. *et al.* (2020) DeCoDe: degenerate codon design for complete protein-coding DNA libraries. *Bioinformatics*, **36**, 3357–3357.

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

Tang,L. *et al.* (2012) Construction of 'small-intelligent' focused mutagenesis libraries using well-designed combinatorial degenerate primers. *Biotechniques*, **52**, 149–158.

Tretyachenko,V. *et al.* (2017) Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci. Rep.*, **7, 1–9**.

Virnekas,B. *et al.* (1994) Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res.*, **22**, 5600–5607.

Vymětal,J. *et al.* (2019) Sequence versus composition: what prescribes IDP biophysical properties? *Entropy*, **21**, 654–658.

Wang,J. *et al.* (2018) A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*, **174**, 688–699.e16.

Weidmann,L. *et al.* (2019) Where Natural Protein Sequences Stand out from Randomness. bioRxiv, 706119.

Wolf,E. and Kim,P.S. (2008) Combinatorial codons: a computer program to approximate amino acid probabilities with biased nucleotide usage. *Protein Sci.*, **8**, 680–688.

Received: 4 January 2021 | Revised: 12 March 2021 | Accepted: 13 March 2021

DOI: 10.1002/pro.4068

ARTICLE

Enzyme catalysis prior to aromatic residues: Reverse engineering of a dephospho-CoA kinase

Mikhail Makarov[1,2] | **Jingwei Meng**[3] | **Vyacheslav Tretyachenko**[1,2] | **Pavel Srb**[4] | **Anna Březinová**[5] | **Valerio Guido Giacobelli**[1] | **Lucie Bednárová**[4] | **Jiří Vondrášek**[4] | **A. Keith Dunker**[3] | **Klára Hlouchová**[1,4]

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

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

[3]Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana

[4]Institute of Organic Chemistry and Biochemistry, IOCB Research Centre & Gilead Sciences, Academy of Sciences of the Czech Republic, Prague, Czech Republic

[5]Proteomics Core Facility, BIOCEV, Faculty of Science, Charles University, Prague, Czech Republic

Correspondence

A. Keith Dunker, Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA.
Email: kedunker@iu.edu

Klára Hlouchová, Department of Cell Biology, Faculty of Science, Charles University, Biocev, Prague 2, Czech Republic.
Email: klara.hlouchova@natur.cuni.cz

Funding information

European Regional Development Fund, Grant/Award Number: CZ.1.05/1.1.00/02.0109; Grantová Agentura České Republiky, Grant/Award Number: 17-10438Y; Ministerstvo Školství, Mládeže a Tělovýchovy, Grant/Award Number: LM2018127; Univerzita Karlova v Praze, Grant/Award Number: SVV260572/2020

Abstract

The wide variety of protein structures and functions results from the diverse properties of the 20 canonical amino acids. The generally accepted hypothesis is that early protein evolution was associated with enrichment of a primordial alphabet, thereby enabling increased protein catalytic efficiencies and functional diversification. Aromatic amino acids were likely among the last additions to genetic code. The main objective of this study was to test whether enzyme catalysis can occur without the aromatic residues (aromatics) by studying the structure and function of dephospho-CoA kinase (DPCK) following aromatic residue depletion. We designed two variants of a putative DPCK from *Aquifex aeolicus* by substituting (a) Tyr, Phe and Trp or (b) all aromatics (including His). Their structural characterization indicates that substituting the aromatics does not markedly alter their secondary structures but does significantly loosen their side chain packing and increase their sizes. Both variants still possess ATPase activity, although with 150–300 times lower efficiency in comparison with the wild-type phosphotransferase activity. The transfer of the phosphate group to the dephospho-CoA substrate becomes heavily uncoupled and only the His-containing variant is still able to perform the phosphotransferase reaction. These data support the hypothesis that proteins in the early stages of life could support catalytic activities, albeit with low efficiencies. An observed significant contraction upon ligand binding is likely important for appropriate organization of the active site. Formation of firm hydrophobic cores, which enable the assembly of stably structured active sites, is suggested to provide a selective advantage for adding the aromatic residues.

Mikhail Makarov and Jingwei Meng shares authorship to this study.

KEYWORDS

aromatic amino acids, catalysis evolution, genetic code evolution, protein disorder, protein structure evolution

1 | INTRODUCTION

The extant alphabet of canonical amino acids was apparently selected in the first 10–15% of Earth history from a plethora of amino acids (a) available on primordial Earth and (b) synthesized through gradually developing metabolic pathways.[1] Recent analyses reveal that, compared to alternatives, the extant alphabet comprises an unusually good repertoire of physical properties.[2–4] Even entirely random sequences built from the canonical alphabet give rise to secondary structure-rich proteins.[5] Nevertheless, soluble and well-expressing proteins have been successfully recovered from random libraries of simpler alphabet of evolutionary early amino acids.[6,7] However, the stage of the amino acid alphabet evolution at which proteins could have gained dominance in binding and catalysis (i.e., functionally support early metabolism) remains unclear.

Aromatic amino acids are considered among the last additions to the genetic coding system, that is, to the canonical amino acid alphabet.[8,9] Because of their relatively high redox reactivity, their fixation in the genetic code could be driven by the biospheric oxygen.[10] There is recent support that for some of the aromatics (Tyr and Trp) this possibly happened even in the post-last universal common ancestor (LUCA) period.[10–12] These proposals suggest that there was a time when living cells existed without aromatic amino acids.

Even though different reduced sets (of 7–13) of the amino acid alphabet have been shown or predicted to be sufficient for protein folding and catalysis, to our knowledge, none of the experimental studies recovered enzyme activity in complete absence of aromatics.[13–18] Computational inquiry indicates that the aromatics are the strongest structure promoters among the 20 amino acid alphabet.[19] This conclusion is consistent with observation that aromatics are mostly clustered within the hydrophobic cores of structured proteins and with quantum chemistry calculations showing the interactions between aromatics to be stronger and more specific than aliphatic side chains interactions.[20] A comparison of the structure/disorder propensities of the 20 amino acids with the chronology of amino acid inclusion into the genetic code indicates that the earliest amino acids are strongly disorder-promoting while the last to be added, for example, the aromatics, are among the most strongly structure-promoting.[9,19,21] Indeed, aromatics are heavily underrepresented in intrinsically disordered proteins and regions (IDPs and IDRs), that is, proteins that lack stable 3D structure and yet frequently carry out crucial biological functions, associated with signaling and regulation in particular.[22,23] While some functions can thus be delivered even in lack of tertiary structure, it remains unclear if and how early enzymes could achieve specific catalysis without a stable hydrophobic core supported by the aromatic residues.

Here, we perform an analysis of structure/function consequences of amino acid reduction by aromatic amino acids. As an exemplary target, we choose a highly conserved metabolic enzyme from a hyperthermophilic bacteria (and hence of potential relevance to early life)—an enzyme that catalyzes the final step of coenzyme A biosynthesis, which is known to be essential for all life and considered among the most ancient cofactors. At the same time, the dephospho-CoA kinase (DPCK) enzyme belongs to the family of P-Loop NTPases that have been argued to be one of the oldest protein architectures, widely preserved.[24] We present evidence that enzyme catalysis can occur in the absence of aromatic amino acids and a firm hydrophobic core, formation of which evidently becomes induced upon ligand binding.

2 | RESULTS

2.1 | Target selection by analysis of LUCA proteins

In order to identify conserved structured protein families, we applied our VSL2B disorder predictor[25] to a collection of LUCA assigned proteins identified by their ubiquity across all kingdoms of cellular life.[26] The non-enzymes, mostly ribosomal or other RNA binding proteins, were all predicted to be massively disordered while the enzymes were predicted to be structured. These modern-day versions of the ancient enzymes all contain multiple aromatic residues and we have been unable to identify a single efficient modern enzyme that lacks multiple aromatic residues. Among the LUCA assigned enzymes identified by Brooks and Fresco, we selected DPCK for further study because it has the lowest number of aromatic amino acids.[26] Other advantages of this choice are that there are multiple 3D structures of different DPCK family members and that the DPCK proteins have relatively small sizes.

2.2 | Sequence design, expression and purification of DPCK variants

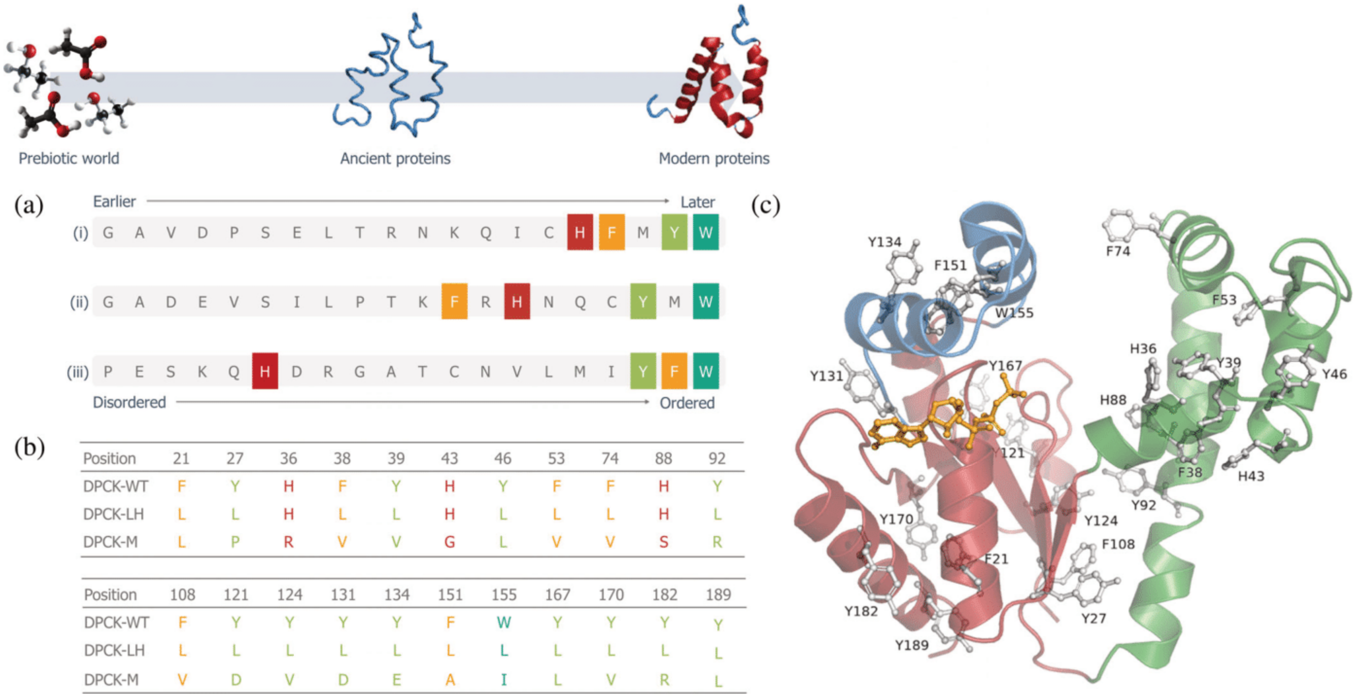
To evaluate the significance of aromatic amino acids for the structure and function of DPCK, the PDB database was first searched for solved structures of confirmed and putative DPCKs from different thermophilic bacterial species (Table S1). An initial test of expression, solubility, ease of large-scale purification and DPCK activity led to selection of a putative DPCK from *Aquifex aeolicus* (PDB ID: 2IF2) for this study (Table S1).

Mutant variants of DPCK were designed as follows. First, all Phe, Tyr and Trp residues were substituted by (a) Leu residues (DPCK-LH) and (b) non-aromatic amino acids based on the best preservation of thermodynamic stability (DPCK-MH) using the Hot Spot Wizard server.[27] Second, all of the above amino acids plus His were substituted using the same logic, producing DPCK-L and DPCK-M variants respectively. DPCK-LH/MH and DPCK-L/M variants thus have 10 and 11% of the total protein sequence substituted, respectively. Synthetic genes of all these variants were subcloned and expressed in *Escherichia coli* with a C-terminal polyhistidine tag using standard protocols (see Section 4 for details). Upon preliminary purification and DPCK activity characterization, only DPCK-LH and DPCK-M variants were selected for detailed characterization (Figure 1). Intriguingly, DPCK-L mutant had a very poor expression (even after optimization attempts) in *E. coli* and both DPCK-L and DPCK-MH mutants did not have any measurable phosphotransferase/ATPase activity (Table S2).

DPCK-WT, -LH and -M variants were purified to homogeneity using a three-step purification protocol (Figure S1). Prior to further experiments, the identity, molecular weight, and oligomeric status of the protein variants were tested by mass spectrometry and analytical size exclusion chromatography (Figure S1). All protein variants were of expected molecular weight. DPCK-WT and -LH eluted as monomers while the -M variant resembles either a dimeric or disordered monomeric form in the elution profile.

2.3 | Enzyme activity characterization

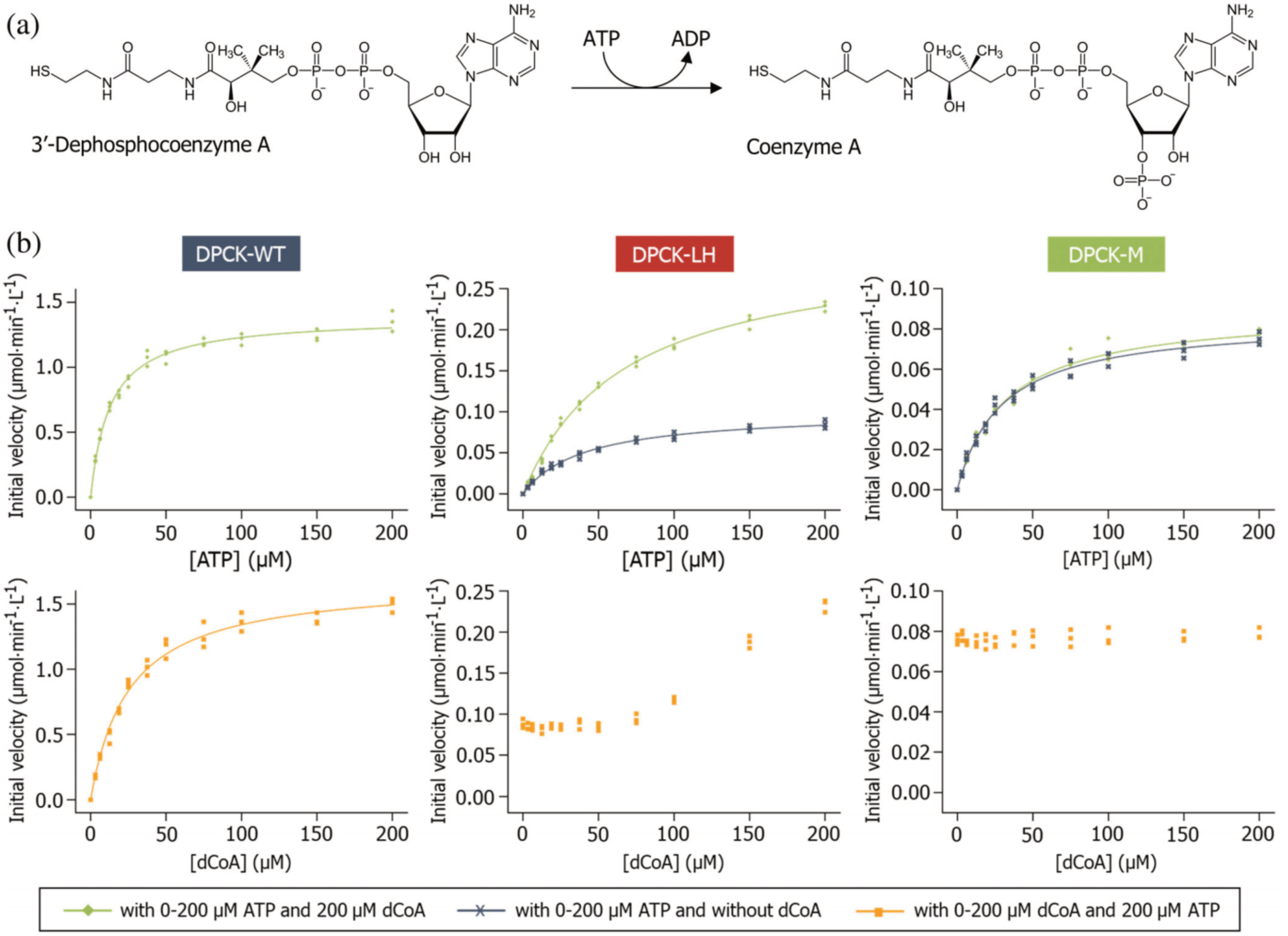
The specificity and rates of enzyme reactions of the DPCK variants were initially characterized using a commercial kit relying on a coupling detection of ADP, one of the reaction products (Figure 2a). In the assay, ADP is converted to pyruvate which is then quantified by a



Position	21	27	36	38	39	43	46	53	74	88	92
DPCK-WT	F	Y	H	F	Y	H	Y	F	F	H	Y
DPCK-LH	L	L	H	L	L	H	L	L	L	H	L
DPCK-M	L	P	R	V	V	G	L	V	V	S	R

Position	108	121	124	131	134	151	155	167	170	182	189
DPCK-WT	F	Y	Y	Y	Y	F	W	Y	Y	Y	Y
DPCK-LH	L	L	L	L	L	L	L	L	L	L	L
DPCK-M	V	D	V	D	E	A	I	L	V	R	L

FIGURE 1 Sequence design of dephospho-CoA kinase (DPCK) variants lacking aromatic amino acids. (a) The chronological order and ranking of 20 amino acids: (i) order of appearance in the genetic code derived by meta-analyses by Trifonov (9); (ii) order of appearance based on the prebiotic availability and thermodynamic stability by Higgs and Pudritz (8); (iii) ranking based on their increasing propensity to promote structure (19). (b) Aromatic amino acid content of DPCK-WT, -LH and -M variants. (c) Aromatic residues highlighted in the structure of DPCK from *Aquifex aeolicus* (PDB ID:2IF2), with ATP molecule positioned based on structural alignment with the *Haemophilus influenzae* DPCK complex with ATP (PDB ID:1JJV)



(c)

Enzyme	Substrate	K_M (μM)	V_{max} (μM · min^{-1})	k_{cat} (s^{-1})	k_{cat}/K_M ($M^{-1} \cdot s^{-1}$)
DPCK-WT	dCoA (with 200 μM ATP)	24.3 ± 1.7	1.57 ± 0.17	0.817 ± 0.088	33621.4
	ATP (with 200 μM dCoA)	12.7 ± 0.3	1.41 ± 0.05	0.730 ± 0.026	57480.3
	ATP (without dCoA)	n.d.	n.d.	n.d.	n.d.
DPCK-LH	dCoA (with 200 μM ATP)	>200	n.d.	n.d.	n.d.
	ATP (with 200 μM dCoA)	65.9 ± 4.5	0.30 ± 0.010	0.0234 ± 0.0008	355.1
	ATP (without dCoA)	40.2 ± 3.8	0.11 ± 0.013	0.0086 ± 0.0010	213.9
DPCK-M	dCoA (with 200 μM ATP)	n.d.	n.d.	n.d.	n.d.
	ATP (with 200 μM dCoA)	32.3 ± 3.4	0.088 ± 0.003	0.0038 ± 0.0001	117.6
	ATP (without dCoA)	29.2 ± 2.5	0.085 ± 0.002	0.0037 ± 0.0001	126.7

FIGURE 2 Kinetic characterization of dephospho-CoA kinase (DPCK) variants. (a) DPCK reaction scheme. (b) Michaelis–Menten plots of DPCK proteins for initial velocity versus (TOP) ATP concentration, monitoring production of ADP; reactions were performed without and with 200 μM dCoA to estimate ATPase and phosphotransferase activities of enzymes. (BOTTOM) dCoA concentration, monitoring production of ADP. Reactions were performed in 15 mM Hepes (pH 7.4), 20 mM NaCl, 1 mM EGTA, 0.02% Tween-20, 10 mM $MgCl_2$, and 0.1% bovine gamma globulin, and was initiated with ATP. The lines represent nonlinear least squares fits. (c) Summary of catalytic efficiencies

fluorometric method. Any basal ATP hydrolysis (in the absence of enzyme) was appropriately subtracted (Figure S2a). Because the assay was performed at two regimes (varying the concentration of ATP or dCoA), it was possible to observe significant differences in the reaction specificity of the variants (Figure 2b,c).

DPCK-WT has similar catalytic efficiency for both ATP and dCoA as substrates while the ATP hydrolysis activity is dependent on dCoA binding (Figure 2). The herein measured catalytic efficiency of the reaction (3.4×10^4 and 5.7×10^4 M^{-1} s^{-1} for dCoA and ATP, respectively) is similar to previously reported efficiency of DPCK from *Entamoeba histolytica*.[28] In contrast, the catalytic efficiency of DPCK-LH and DPCK-M are significantly lower (355 and 118 M^{-1} s^{-1} for ATP, respectively), resulting in a decreased turnover number (Figures 2 and S2b). In the case of DPCK-M, the reaction rates are independent of varying concentrations of dCoA implying an impaired efficiency of the phosphate transfer, that is, only ATPase activity is observed (Figure 2c). While both DPCK-LH and -M variants have the ability to hydrolyze ATP in the absence of dCoA (unlike DPCK-WT), DPCK-LH has also the dCoA-dependent phosphotransferase activity (above ∼80 μM dCoA) with K_M greater than 200 μM. This activity has been difficult to measure using the commercial kit due to the ATP concentration range limitation. In order to confirm the identity of the reaction products and reaction specificity, the DPCK reactions were performed at a fixed substrate concentration above the DPCK-WT K_M value (where the reaction rate is less dependent or independent of substrate concentration) in order to reach sufficient substrate conversion for detection of the products using HPLC-MS analysis. This analysis detected significant CoA formation only in the reaction catalyzed by DPCK-WT and 100x lower CoA formation was detected in the reactions catalyzed by DPCK-LH (Figure S2b).

2.4 | Secondary and tertiary structure characterization

Using the purified proteins, their structural properties were investigated using electronic circular dichroism (ECD), NMR and limited proteolysis.

ECD spectrum of DPCK-WT (Figure 3a) with comparable intensity of negative maxima at 209 and 225 nm and with intense positive maximum at 195 nm indicates relatively high partition of α-helical structure (∼45%). This is confirmed by the numerical data analysis and agrees with the secondary structure assignment of the X-ray structure (PDB code: 2IF2) (Table S3). In the case of DPCK-LH, the first negative maximum is blue-shifted to 207 nm and its intensity is comparable to that of the

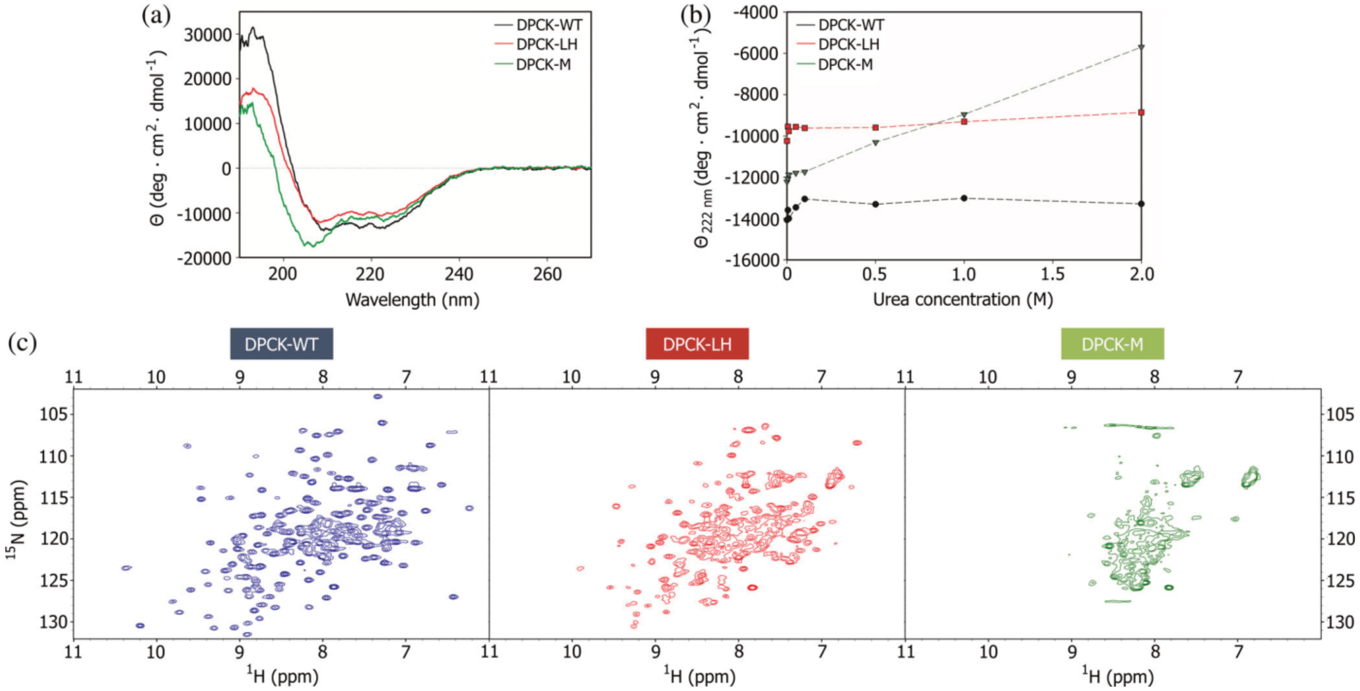


FIGURE 3 Secondary and tertiary structure characterization of dephospho-CoA kinase (DPCK) variants. (a) Far-UV CD spectra of DPCK proteins. The spectra were collected in PBS buffer (11.8 mM phosphate (pH 7.6), 137 mM NaCl, 5 mM $MgCl_2$, 2.7 mM KCl and 0.5 mM DTT). (b) Change in ellipticity at 222 nm upon 0–2 M urea titration of DPCK proteins. (c) 2D NMR of DPCK proteins. The spectra were collected in 50 mM phosphate (pH 7.6), 280 mM NaCl, 20 mM KCl, 10 mM $MgCl_2$, and 0.5 mM TCEP

second negative maximum at 225 nm. This together with the positive maximum at 195 nm (almost half intensity compared to ECD spectrum of DPCK-WT) reveals a significant content of α-helical structure (∼40%) together with more pronounced partition of β-sheet structure, confirmed also by the numerical data analysis (Table S3). ECD spectrum of DPCK-M has the first negative maximum also blue-shifted up to 205 nm but this spectral band is more intense compared to the second negative maximum at 222 nm, which could imply possible formation of 3_{10}-helical structure as well as enrichment of unordered structure. The overall spectral shape and mainly spectral intensity of a positive spectral band at 192 nm could be due to a relatively high portion of β-sheet structure (Table S3).

To estimate the influence of aromatic amino acid substitution on overall protein structure stability, the proteins were unfolded with urea in concentration ranging from 0 to 2 M and were further studied using CD spectroscopy. While DPCK-WT and -LH ECD spectra remain relatively constant upon mild urea titration (up to 2 M), the urea titration spectra indicate loss of structural stability in the DPCK-M variant, starting already in very low urea concentrations (Figure 3b).

Similarity of structural resemblance of DPCK-LH and DPCK-WT was further confirmed by 1D and 2D HN NMR spectra. DPCK-WT spectrum has a good signal dispersion in the -NH- region (6–9 ppm) and clear signals near 1 ppm indicative of methyl groups in the hydrophobic core, all features corresponding to a well-folded protein. While the signal of the methyl groups in the hydrophobic core is absent in the DPCK-LH variant spectrum (as expected from the removal of aromatic residues), the signal dispersion in the -NH- region implies that the -LH variant is at least partially folded, in contrast with that of the -M variant where the signal in the same region is less dispersed, implying lack of specific tertiary structure (Figures 3c and S3). Based on the analyses of N-edited 3D NOESY spectra, the following counts of α-helical peaks at 131, 57 and 14 were estimated for DPCK-WT, -LH and -M variants, respectively (Table S3).

The tertiary structure of the proteins was additionally characterized by limited proteolysis using endoproteinase Lys-C as its cleavage site map is conserved among all studied variants (Figure S4). DPCK-WT is highly resistant to proteolytic digestion during the whole-time scale of the limited proteolysis experiment, reflecting its globular structure. In contrast, both mutant variants are gradually digested by Lys-C over time, with the amounts of the intact DPCK-LH and DPCK-M decreasing exponentially over time. While relatively large cleavage fragments with the approximate size of 15 kDa can be observed during proteolysis of DPCK-LH, no large cleavage fragments are detected for DPCK-M, an indication of its loose or absent tertiary structure (Figure 4).

In summary, DPCK-LH variant (which has all the aromatic amino acids substituted by leucine) shows relatively high conservation of secondary structure but a loose tertiary structure (probably of molten globular nature) when compared with DPCK-WT. On the other hand, both secondary and tertiary structures of DPCK-M variant are severely impaired, in which all histidines were substituted in addition to aromatics.

2.5 | Structural characterization of ATP binding

For an efficient phosphorylation reaction, the γ-phosphate of ATP must be protected from a nucleophilic attack by water molecules. DPCK active site must therefore be shielded from water once the ATP molecule is bound. For several kinases this shielding is accomplished by an induced-fit conformational change upon ATP binding. Such a conformational change has also been observed for DPCK.[29]

To study the structural changes of DPCK variants upon ATP binding, 2D HN NMR spectra were collected in response to ATP titration (see Figure S5). While the NMR spectra of the DPCK-M variant are of generally low quality, which is probably caused by complex dynamics on the millisecond time scale making the protein signals invisible for NMR spectroscopy, the spectra of both DPCK-WT and -LH variants show expected perturbations upon ATP titration. For DPCK-WT we observe typical examples of slow exchange behavior, where only free and bound forms are observed with peak intensity proportional to the population. Interestingly, for the DPCK-LH variant, we typically observed examples of fast exchange with only a single peak visible at a given protein:ATP ratio, although examples of slow exchange are observed as well (Figure 5a). This suggests that compared to the DPCK-WT:ATP interaction, an additional process occurs during DPCK-LH titration with ATP.

To further investigate this intriguing observation, DPCK-WT and -LH (i.e., those variants that are capable of phosphotransferase activity that requires a hydrophobic core) structural response to substrate binding was tested using dynamic light scattering and 8-anilinonaphthalene-1-sulfonic acid (ANS) titration. The steady-state fluorescence measurements lend support to the molten globule nature of DPCK-LH variant since it shows higher fluorescence intensity values in comparison with DPCK-WT, resulting from the high affinity of ANS to the exposed hydrophobic core of molten globular intermediates.[30] While the fluorescence intensity decreases for

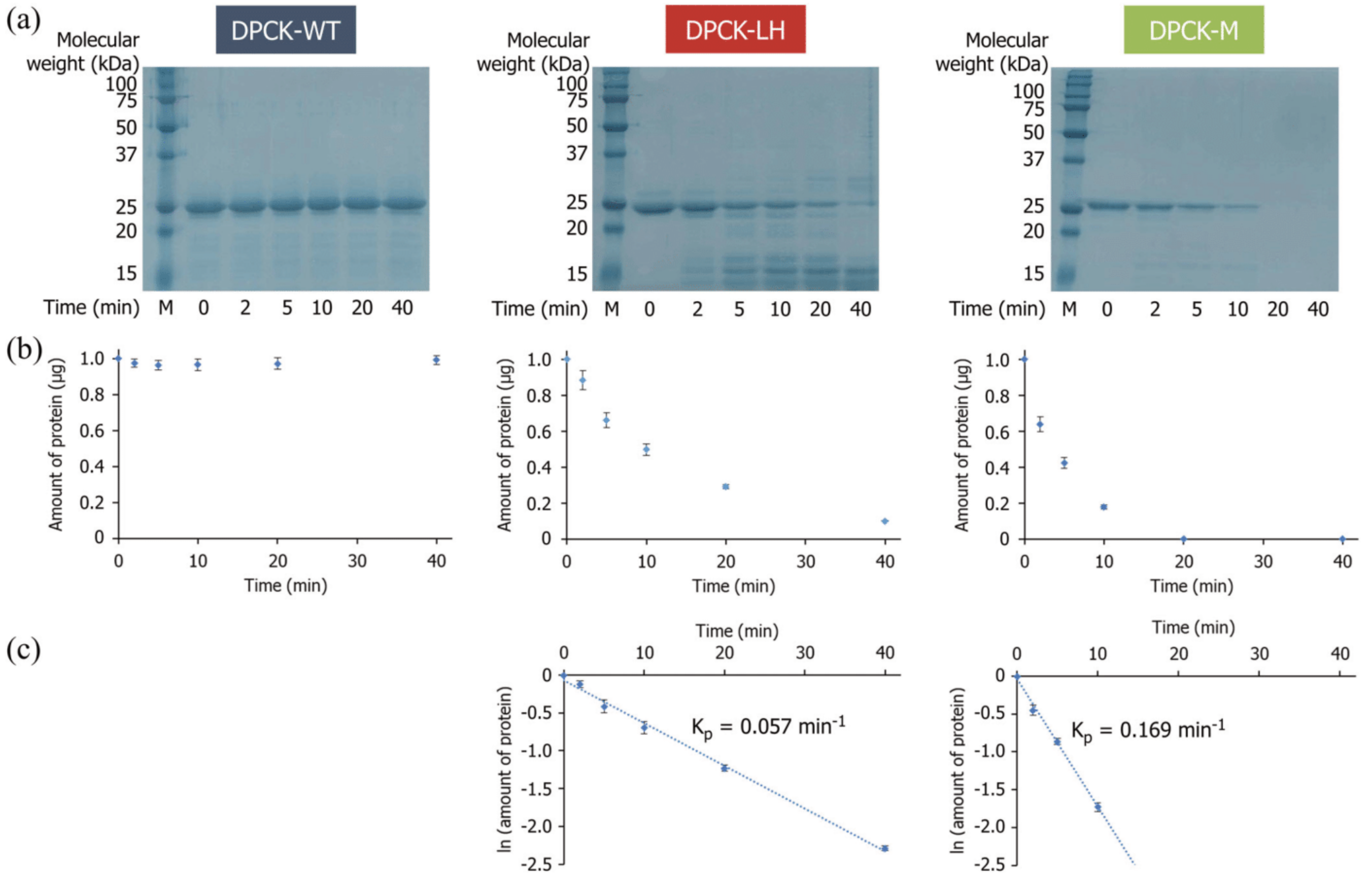


FIGURE 4 Limited proteolysis of dephospho-CoA kinase (DPCK) proteins. (a) 14% SDS-polyacrylamide gels visualized by imidazole-zinc staining after SDS-PAGE with the protein samples exposed to Lys-C endoproteinase for different times. (b) Graphs representing the amount of the proteins remaining at each time point. (c) Determination of proteolysis rate constants (Kp) assuming pseudo-first order of proteolytic reactions

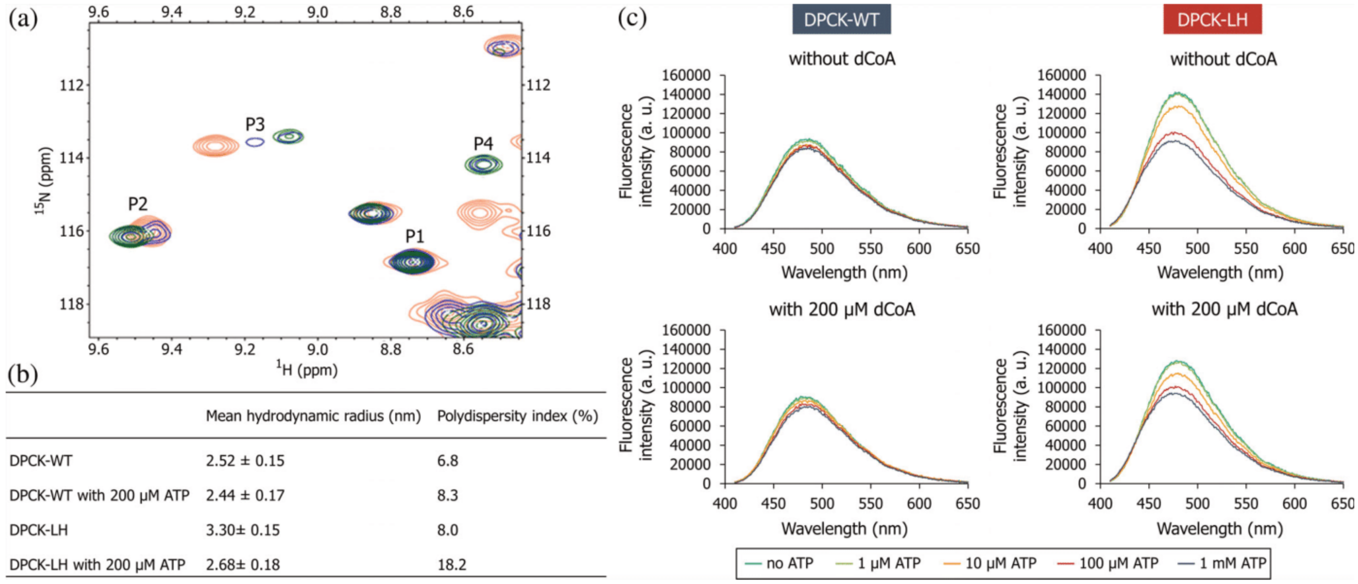
both variants upon substrate binding, this change is significantly more dramatic for DPCK-LH (Figure 5c). ATP (out of the two substrates) has a remarkable effect on additional folding of DPCK-LH protein, explaining its ability to perform the phosphotransferase activity despite its molten globular nature in the free state. Both 2D HN NMR and ANS titration observations were further supported by DLS measurements where the mean hydrodynamic radius of DPCK-LH was recorded to be reduced by ~20% and reached that of DPCK-WT value upon ATP addition (Figure 5b).

3 | DISCUSSION

Aromatic residues are essential for formation of a stable hydrophobic core of extant proteins.[20] At the same time, tight protein folding is frequently required for enzyme catalysis even though most enzymes undergo dynamic structural changes during the reaction. With aromatics being apparently the latest addition to the amino acid alphabet, how specific protein catalysis could be achieved in their absence remains unclear. The work reported here sheds some light on this problem.

To examine the contribution of the aromatic amino acids to enzyme catalysis, we performed a detailed analysis of two aromatics-less mutants of the *Aquifex aeolicus* DPCK where (a) all Phe, Tyr and Trp residues were substituted by Leu residues (DPCK-LH), and (b) all Phe, Tyr, Trp and His were substituted by non-aromatic amino acids based on predicted preservation of thermodynamic stability (DPCK-M).

DPCK catalyzes the transfer of phosphate group from ATP to dCoA, where dCoA acts as the leading substrate.[31] It belongs to the ancient family of P-loop NTPases with the preserved three-layer αβα sandwich architecture.[24] The P-loop motif has been detected among the primordial peptide fragments and is known to underlie hundreds of essential enzyme families.[32,33] Besides mononucleotide binding, polypeptides constructed around this scaffold have been shown to bind polynucleotides/RNA/ssDNA and even unwind dsDNA, pointing to the functional plasticity of the P-loop motif.[34,35] The specific function of an NTPase relies on the topology and



	Mean hydrodynamic radius (nm)	Polydispersity index (%)
DPCK-WT	2.52 ± 0.15	6.8
DPCK-WT with 200 μM ATP	2.44 ± 0.17	8.3
DPCK-LH	3.30± 0.15	8.0
DPCK-LH with 200 μM ATP	2.68± 0.18	18.2

FIGURE 5 Structural characterization of dephospho-CoA kinase (DPCK)-WT and -LH upon substrate binding. (a) An exemplary close-up of DPCK-LH 2D NH NMR spectra induced by ATP binding (red—free protein [100 μM], blue—300 μM ATP, green—1000 μM ATP); labeled peaks: (P1) N-H signal not influenced by protein-ATP interaction. (P2, P3) N-H signal undergoing medium-slow to slow exchange on NMR chemical shift time scale (μs-ms). (P4) N-H signal documenting a slow exchange process. (b) Mean hydrodynamic radius of DPCK-WT and -LH variants with and without 200 μM ATP measured by dynamic light scattering. (c) The steady-state fluorescence spectra of ANS binding at excitation wavelength 380 nm. The spectra were measured at different concentrations of ATP (with and without 200 μM dCoA), and each spectrum is the average of three individual scans. The fluorescence was recorded between 410 and 650 nm after exciting the protein solution at 380 nm

overall structural context, including many additional active site residues. DPCK is a well folded protein with high α-helical content and its domain movements upon ATP binding play a crucial role during catalysis.[29] None of the aromatic amino acid residues has been reported essential for the ligand binding and catalysis in DPCK.[31]

Both aromatics-less mutants and wild type protein were characterized in terms of their structure and activity. Interestingly, the DPCK-M variant (selected for best predicted preservation of thermodynamic stability) had a more impaired structural integrity than DPCK-LH. This may be either the consequence of the specific substitutions or the indispensability of the DPCK's His residues. From an evolutionary perspective, His was among the last amino acids incorporated into genetic coding.[9] On the other hand, according to the order–disorder propensity scale, His is among the most disorder-promoting amino acids, likely due to its significant positive charge and the two hydrogen-bonding nitrogen atoms that could promote structural instability by hydrogen bond switching.[19] Despite its high disorder-promoting tendency, His often plays an important role in inducing protein structure formation in the presence of divalent cations, especially zinc, due to its metal ion coordination. Given the speculative role of His for hydrogen bond switching, it would be interesting to determine whether His plays a role in facilitating the domain movements needed for catalysis.

CD and NMR measurements of the DPCK-LH variant showed a similar content of secondary structure to the wild type protein but limited proteolysis and 2D NMR all imply its molten globule tertiary conformation. DPCK-M variant has no measurable phosphotransferase activity while both of the mutant variants are able to hydrolyze ATP even in the absence of dCoA. This is likely due to the loss of structural orchestration of the catalytic events and demonstrates that some activities can be performed even in the absence of a firm hydrophobic core. However, this is probably untrue for the phosphotransferase activity where the gamma-phosphate has to be protected from a nucleophilic attack by water molecules in order to be efficiently transferred to the desired substrate. Interestingly, DPCK-LH variant is still able to perform this activity although with significantly lower efficiency (∼100×) in comparison with the wild type protein. Both DPCK-WT and -LH variants share slow-exchange behavior in the NMR spectra upon ATP titration, suggesting the ATP-induced change in their structural conformation. However, the DPCK-LH variant undergoes significant additional folding upon ATP binding, explaining its ability to perform the phosphotransferase reaction. This ligand-induced folding scenario is in agreement with previously reported behavior of engineered molten globular enzymes.[36–39]

The study of the engineered molten globular enzyme[37] includes the hypothesis that modern enzymes evolved from molten globular precursors. If the earliest cells indeed existed without aromatics, then the data in this paper adds weight to this evolutionary scenario of molten globular polypeptide → molten globular enzyme → modern enzyme, where the last step is enabled by the expansion of the genetic code to include the aromatic residues. While small aromatics-less peptides have been reported previously to have catalytic properties,[40,41] aromatic amino acids have been considered essential for formation of tight structured proteins to support high-performance catalysis. Association of protein fold stabilization with genetic code evolution has been addressed by several recent studies.[16,18,42,43] Most significantly for protein folding, basic and aromatic amino acids (at least the canonical ones) were most probably absent in the prebiotic set.[8,9] Early protein foldability thus would not be supported by salt bridges and aromatic core packing interactions that make up extant protein cores. Earlier studies suggested that this hindrance could be compensated by a halophilic environment because high salt stabilizes proteins structure and supported halophilic origins of life scenarios.[44,45] Using a small designed β-trefoil protein highly enriched in the prebiotic amino acids (and completely devoid of aromatics), Longo et al. demonstrated that incorporation of a single aromatic amino acid can convert a foldable halophilic protein to a stable mesophile.[42] However, two other studies referenced here concluded that robust protein folds can be built with prebiotically plausible subset of the current 20 amino acids while the other amino acids (i.e., evolutionary late) contribute mainly to efficient catalysis.[16,18] These conclusions were drawn from stability and catalytic characterization of multiple variants of nucleoside diphosphate kinase reduced to 13 and 10 amino acid alphabets, respectively. While we cannot directly deny their conclusions by our study, it is important to notice that only ~80% of the proteins′ sequence was occupied by prebiotically available amino acids in the two studies by the Akanuma group. None of the successfully expressed variants was completely rid of aromatic and other amino acids that are not regarded as prebiotically plausible (such as positively charged Lys/Arg). Therefore, these studies support (or at least do not rule out) key importance of aromatics in protein fold stability and their role in the transition from molten globule to stable globular proteins.

If the transition from molten globular to stable folded enzymes was mediated by the evolutionary later amino acids, this transition was also likely accompanied by evolution of functionality and substrate specificity. The specific aim of our study was not to resurrect an early version of DPCK per se but rather to explore the specific effect of the aromatic amino acid replacements on its structure–function relationship. However, future reverse evolution studies of this enzyme class should bear in mind that the early function could be altered or less specific.

To further test the role of aromatics in protein fold evolution, work in progress is to use bioinformatics tools to carry out disorder prediction with VSL2B on DPCK and the other identified ancient enzymes with their modern sequences and with their aromatics replaced by Leu. All of the ancient enzymes so far tested are predicted to be structured with the aromatics and disordered without these residues. The next step will be to apply additional bioinformatic tools that distinguish molten globules from other types of disorder.[46]

In summary, we report an enzyme without aromatic amino acids that is still capable of a specific, hydrophobic core dependent catalysis. This enzyme is rich in secondary structure but exhibits a molten globule conformation in an unliganded form. Our study provides evidence that a tightly packed protein environment can be formed upon its ligand binding. This phenomenon could be relevant in the early stages of enzyme catalysis before the fixation of the contemporary amino acid alphabet.

4 | METHODS

4.1 | Plasmid preparation

DPCK genes for DPCK-WT, DPCK-LH, -L, -MH and -M were amplified by PCR using Pfu-X DNA polymerase (Jena Bioscience, Germany) according to the following program: an initial denaturation at 95°C for 2 min; followed by the 32 cycles of denaturation at 95°C for 30 s; annealing at 56°C for 30 s; elongation at 68°C for 30 s; and a final extension at 68°C for 2 min. The PCR amplification for all genes was performed with the same set of primers: forward, 5′-AAAAACATATGAAACGTATCGG TCTGACC-3′, and reverse, 5′-AAAAACTCGAGTTCCAG CGGGTCACGG-3′. The PCR fragments were digested with *XhoI* (New England BioLabs) and *NdeI* (New England BioLabs), purified with Monarch PCR & DNA Cleanup Kit (New England BioLabs) and cloned into PET-24a (+) C-terminal polyhistidine-tag vector (Novagen, Germany), which was digested by *XhoI* and *NdeI* and dephosphorylated by Antarctic Phosphatase (New England BioLabs) prior to ligation.

The plasmids were introduced into One Shot TOP10 Chemically Competent *E. coli* cells (Thermo Fisher

Scientific) by heat shock protocol at 42°C for 60 s, and the cells were grown overnight at 37°C on LB agar plates containing 50 μg/ml of kanamycin (Sigma Aldrich). A single colony was selected, cells were grown overnight at 37°C in 5 ml of LB Broth (Sigma Aldrich) supplemented with 50 μg/ml of kanamycin (Sigma Aldrich) and plasmid DNA was isolated and analyzed by Sanger sequencing.

4.2 | Protein expression and purification

Isolated plasmids were introduced into BL21 (DE3) Chemically Competent *E. coli* cells (Thermo Fisher Scientific), and the cells were grown overnight at 37°C in 5 ml of LB Broth (Sigma Aldrich) in the presence of 50 μg/ml of kanamycin. The overnight cultures were used to inoculate 500 ml of fresh LB medium, and the culture was propagated at 37°C at 220 rpm shaking. When OD_{600} reached 0.7-0.8, isopropyl β-D-thiogalactopyranoside (IPTG, Sigma Aldrich) was added to final concentration of 0.5 mM and the cultivation was continued for 4 hr at 37°C. The cells were harvested by centrifugation at ×3000*g* for 20 min at 4°C. The cell pellets were resuspended in 15 ml of lysis buffer (20 mM Tris (pH 8.0), 20 mM NaCl, and 1 mM β-mercaptoethanol) with one tablet of EASYpack protease inhibitor cocktail (Sigma Aldrich), incubated with 50 μg/ml of Lysozyme (Sigma Aldrich) and 6 U of RNase-free DNase I (Jena Bioscience, Germany) at room temperature for 30 min, sonicated on ice at 1.5 W (18 cycles, 10 s on, 20 s off) and centrifuged at ×35000*g* for 30 min at 4°C. After, Tween-20 (Sigma Aldrich) was added to supernatants to the final concentration of 0.1% (vol/vol), and the crude lysates were applied to 5 ml HiTrap Capto Q column (GE Healthcare Life Sciences) equilibrated with 5 volumes of buffer A (20 mM Tris (pH 8.0), 20 mM NaCl, 1 mM beta-mercaptoethanol and 0.1% (vol/vol) Tween-20). Then, the DPCK proteins were eluted with 0–50% gradient of buffer B (20 mM Tris (pH 8.0), 1 M NaCl, 1 mM beta-mercaptoethanol and 0.1% (vol/vol) Tween-20), and fractions from 15 to 35% of buffer B were collected and applied to 5 ml HisTrap HP column (GE Healthcare Life Sciences) equilibrated with 5 volumes of buffer C (20 mM Tris (pH 7.6), 500 mM NaCl, 10 mM imidazole, 1 mM beta-mercaptoethanol and 0.1% (vol/vol) Tween-20). The column was washed with 3% of buffer D (20 mM Tris (pH 7.6), 500 mM NaCl, 500 mM imidazole, 1 mM beta-mercaptoethanol and 0.1% (vol/vol) Tween-20) to remove unbound proteins, and the DPCK proteins were eluted with 0–50% gradient of buffer D. Fractions from 20 to 30% of buffer D were collected, concentrated up to 0.5 ml by centrifugation using 4 ml Amicon Ultra centrifugal unit (MWCO 10000, Millipore) and applied to Superdex 75 10/300 GL column (GE Healthcare Life Sciences) equilibrated with 2 column volumes of buffer E (50 mM Tris (pH 7.6), 500 mM NaCl, 20 mM KCl, 10 mM $MgCl_2$ and 0.5 mM DTT). The DPCK variants were eluted as single peaks with approximate sizes of 29 kDa (DPCK-WT), 33 kDa (DPCK-LH) and 55 kDa (DPCK-M). Molecular weights were estimated using Gel filtration low molecular weight calibration kit (GE Healthcare Life Sciences). After the confirmation of proteins integrity and purity by SDS-PAGE analysis on 14% SDS-polyacrylamide gel, the purified proteins were concentrated up to 10 mg/ml concentration and aliquoted. The aliquots were flash frozen in liquid nitrogen and stored at −80°C.

4.3 | Basic biophysical characterization

The identities and molecular weights of purified proteins were confirmed by mass spectrometry using UltrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker, Germany) according to the standard procedure. Protein concentrations were determined by amino acid analysis using a Biochrom 30+ Series Amino Acid Analyser (Biochrom, United Kingdom).

The size distribution of protein samples was characterized using dynamic light scattering (DLS) technique. Protein samples were diluted in PBS buffer (11.8 mM phosphate buffer (pH 7.6), 137 mM NaCl, 5 mM $MgCl_2$, 2.7 mM KCl and 0.5 mM DTT) to the final concentration of 0.5 mg/ml and centrifuged at ×25000*g* for 30 min at 4°C. In order to remove dust particles, samples were filtered using 0.22 μm Ultrafree-MC centrifugation filter (Millipore). The DLS measurements were performed in a quartz glass cuvette (light path 10 mm) at 18°C using a laser spectroscatter-201 system (RiNA GmbH Berlin, Germany). A series of 35 measurements with a sampling time of 30 s and a wait time of 1 s was conducted for each sample. A diode laser of wavelength 685 nm and an optical power of 30 mW was used as the source. The scattered light was collected at a fixed scattering angle of 90°, and the autocorrelation functions were analyzed with the program CONTIN to obtain hydrodynamic radius distributions. DLS measurements were performed for protein samples in the presence of 200 μM ATP to estimate the effect of ATP binding on the hydrodynamic radius of proteins.

4.4 | Enzyme assays

DPCK activities of recombinant proteins were measured by a coupling assay using ADP Quest Assay kit (Eurofins

DiscoverX) according to the manufacturer's instructions. Enzyme assays were carried out using 80 ng (32 nM) of DPCK-WT, 500 ng (214 nM) of DPCK-LH and 900 ng (386 nM) of DPCK-M and two kind of substrates, 0–200 μM for dephospho-CoA (dCoA) at 200 μM ATP and 0–200 μM for ATP without and with 200 μM dCoA to estimate ATPase and phosphotransferase activities of enzymes. All reactions were performed in assay buffer containing 15 mM Hepes (pH 7.4), 20 mM NaCl, 1 mM EGTA, 0.02% Tween-20, 10 mM $MgCl_2$, and 0.1% bovine gamma globulin in 96-well black microplate with 40 μl total volume. After 20 μl of reagent A and 40 μl of reagent B were added, the plates were heated at 37°C for 10 min, and the reactions were started by adding ATP. The fluorescent intensity signal was measured at 37°C in kinetic mode with 2 min intervals using CLARIO star microplate reader (BMG LABTECH, Germany) at excitation/emission wavelengths of 530/590 nm. The kinetic parameters were calculated using the non-linear regression function using the single saturating concentrations of substrates. Substrate conversion did not exceed 10%. The experiments were repeated three times, and kinetic values are presented as the means ± *SE*.

HPLC-MS analysis was used for comparative detection of the reaction analytes. For this purpose, 100 μl of reaction mixtures were prepared by mixing 1 μg (0.42 μM) of protein, 100 μM dCoA and 100 μM ATP in 25 mM NH_4HCO_3 (pH 7.6), 300 mM NaCl, 20 mM KCl and 10 mM $MgCl_2$. The reaction mixture was incubated at 37°C for 1 hr, then, reaction was stopped by adding 100 μl of acetonitrile (Sigma Aldrich). Precipitated recombinant protein was separated by centrifugation at ×20000*g* at 4°C for 20 min.

The reaction samples were analyzed using the Dionex Ultimate 3000RS HPLC equipped with TSQ Quantiva MS detector (Thermo Fisher Scientific). The ESI source was used for ionization in a positive mode. The HPLC solvent system consisted of 10 mM $(NH_4)_2CO_3$ (pH 9.3) (A) and 97% acetonitrile (B). One microliter sample was injected in 50% B and the analysis was performed using the gradient of 15% A and 85% B for 3.5 min followed by an increase to 75% A and 25% B over 11.5 min and its continuation for 10 min with the SeQuant® ZIC®-*p*HILIC column (5 μm, 150 mm × 2.1 mm, Merck), at a flow rate of 0.13 ml/min.

4.5 | Circular dichroism spectroscopy

ECD spectra were collected using a Jasco 1500 spectrometer (JASCO, Japan) in the 195–280 nm spectral range using a 0.01 cm cylindrical quartz cell. The experimental setup was as follows: 0.05 nm step resolution, 5 nm/min scanning speed, 16 s response time, 1 nm spectral band width and 2 accumulations. After baseline correction, the spectra were expressed as molar ellipticity per residue θ ($\deg{\cdot}cm^2{\cdot}dmol^{-1}$). The protein samples were diluted in PBS buffer (11.8 mM phosphate (pH 7.6), 137 mM NaCl, 5 mM $MgCl_2$, 2.7 mM KCl and 0.5 mM DTT) with addition of 0–2 M urea (specifically 5, 10, 50, 100, 500, 1000, and 2000 mM urea concentrations). The blank spectrum of an aqueous buffer (with or without urea in a corresponding concentration) was used to correct the observed spectrum of the sample. The numerical analysis of secondary structures was performed using the CDPro software package.[47]

4.6 | Limited proteolysis

Kinetic studies on specific proteolytic cleavage by Lys-C endoproteinase were performed as follows. First, recombinant proteins were diluted in Lys-C cleavage buffer (25 mM Tris (pH 8.0), 300 mM NaCl, 1 mM EDTA, and 0.5 mM TCEP) to the final concentration of 1 mg/ml, and then reaction mixtures for proteolytic digestion were prepared by mixing 7 μl of 1 mg/ml recombinant protein and 56 μl of Lys-C cleavage buffer. After incubation at 37°C for 10 min proteolytic cleavage was initiated by adding 7 μl of 5 ng/μl Lys-C endoproteinase. After 0, 2, 5, 10, 20 and 40 min of incubation at 37°C 10 μl of the reaction mixture was taken out, and Lys-C was inactivated by adding 2 μl of 6× SDS-PAGE sample buffer (375 mM Tris–HCl (pH 6.8), 9% SDS, 50% glycerol, 9% beta-mercaptoethanol and 0.03% bromophenol blue) followed by heating at 95°C for 10 min. All samples then were subjected to SDS-PAGE.

For quantitative evaluation of limited proteolysis, the rate constants of proteolysis were determined by monitoring the disappearance of an intact protein in a proteolysis reaction by SDS-PAGE. The areas of the bands corresponding to the intact proteins were estimated from the gels using the ImageJ program and then expressed as the amount of protein remaining after each time point. Assuming the pseudo-first order kinetics, the natural logarithms of the intact protein amounts were plotted against the time, and the plots were fitted with a first-order rate equation.

4.7 | Steady-state ANS fluorescence

Steady-state fluorescence measurements were performed using CLARIO star microplate reader (BMG LABTECH, Germany). Protein samples were diluted to 2 μM with ANS buffer (100 mM Tris (pH 7.6), 300 mM NaCl, 20 mM KCl, 10 mM $MgCl_2$ and 0.5 mM DTT) and incubated with 0; 1; 10; 100 and 1000 μM of adenosine 5′-[γ-thio]triphosphate tetralithium salt (ATP-γ-S, Sigma Aldrich) at room temperature for 30 min. After incubation, 8-anilino-1-naphthalenesulfonic acid ammonium salt (ANS, Sigma Aldrich) was added to the reaction mixtures to the final concentration of 400 μM, and the reaction mixtures were incubated for additional 5 min. The final volume of each reaction mixture was 50 μl. The ANS fluorescence was excited at 380 nm, and emission spectra were recorded between 410 and 650 nm. To estimate the conformational changes induced upon dCoA binding the fluorescence intensity measurements were performed for protein samples in the presence of 200 μM dCoA. All measurements were performed in triplicates and then averaged to yield steady-state fluorescence spectra of ANS binding.

4.8 | NMR spectroscopy

NMR spectra were obtained using the Bruker© Avance HD III 850 MHz instrument, equipped with triple-resonance cryo-probe. Sample volume was 0.16 ml in 3 mm NMR tubes, in 50 mM phosphate (pH 7.6), 280 mM NaCl, 10 mM $MgCl_2$, 20 mM KCl and 0.5 mM TCEP. Protein concentration was 150 μM for 3D $^{15}N/^{1}H$ NOESY-HSQC spectra, 30 μM for DPCK-WT ATP titration, and 100 μM for both aromatic amino acid-lacking mutants ATP titration. All proteins used in the study were ^{15}N labeled. ATP titrations were followed using a series of standard 1D and 2D HN correlation spectra.

ACKNOWLEDGMENTS

We would like to thank Dr Radko Souček for help with the amino acid analysis and Dr Tereza Ormsby, Dr Rozálie Hexnerová and Dr Václav Veverka for helpful suggestions. This work was supported by the Czech Science Foundation (GAČR) grant number 17-10438Y, project SVV260572/2020 (Vyacheslav Tretyachenko and Mikhail Makarov) and by the project BIOCEV (CZ.1.05/1.1.00/02.0109), from the European Regional Development Fund. We would also like to acknowledge the facility and support of the CMS-Biocev ("Biophysical techniques") supported by MEYS ČR (LM2018127).

AUTHOR CONTRIBUTIONS

Mikhail Makarov: Formal analysis; investigation; methodology; writing-original draft; writing-review & editing. **Jingwei Meng:** Investigation; methodology. **Vyacheslav Tretyachenko:** Formal analysis; methodology; supervision. **Pavel Srb:** Formal analysis; investigation; methodology; validation; visualization. **Anna Březinová:** Formal analysis; methodology; validation; visualization. **Valerio Giacobelli:** Formal analysis; methodology; supervision; validation. **Lucie Bednárová:** Formal analysis; methodology; validation; visualization; writing-original draft. **Jiri Vondrasek:** Conceptualization; methodology; software. **Keith Dunker:** Conceptualization; formal analysis; methodology; project administration; resources; supervision; writing-review & editing. **Klara Hlouchova:** Conceptualization; formal analysis; funding acquisition; methodology; project administration; supervision; writing-original draft; writing-review & editing.

ORCID

A. Keith Dunker https://orcid.org/0000-0002-0744-5243
Klára Hlouchová https://orcid.org/0000-0002-5651-4874

REFERENCES

1. Cleaves HJ II. The origin of the biologically coded amino acids. J Theor Biol. 2010;263:490–498.
2. Philip GK, Freeland SJ. Did evolution select a nonrandom "alphabet" of amino acids? Astrobiology. 2011;11:235–240.
3. Ilardo M, Meringer M, Freeland SJ, Rasulev B, Cleaves HJ II. Extraordinarily adaptive properties of the genetically encoded amino acids. Sci Rep. 2015;5:9414.
4. Ilardo M, Bose R, Meringer M, et al. Adaptive properties of the genetically encoded amino acid alphabet are inherited from its subsets. Sci Rep. 2019;9:12468.
5. Tretyachenko V, Vymětal J, Bednárová L, et al. Random protein sequences can form defined secondary structures and are well-tolerated *in vivo*. Sci Rep. 2017;7:15449.
6. 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;19:786–795.
7. 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;20:846–856.
8. Higgs PG, Pudritz RE. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology. 2009;9:483–490.
9. Trifonov EN. Consensus temporal order of amino acids and evolution of the triplet code. Gene. 2000;261:139–151.
10. Granold M, Hajieva P, Toşa MI, Irimie FD, Moosmann B. Modern diversification of the amino acid repertoire driven by oxygen. Proc Natl Acad Sci USA. 2018;115:41–46.
11. Fournier GP, Alm EJ. Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. J Mol Evol. 2015;80:171–185.
12. Yang XL, Otero FJ, Skene RJ, McRee DE, Schimmel P, Ribas de Pouplana L. Crystal structures that suggest late development of genetic code components for differentiating aromatic side chains. Proc Natl Acad Sci USA. 2003;100:15376–15380.
13. Riddle DS, Santiago JV, Bray-Hall ST, et al. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol. 1997;4:805–809.
14. Akanuma S, Kigawa T, Yokoyama S. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci USA. 2002;99:13549–13553.

15. Longo LM, Lee J, Blaber M. Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. Proc Natl Acad Sci USA. 2013;110:2135–2139.
16. Shibue R, Sasamoto T, Shimada M, Zhang B, Yamagishi A, Akanuma S. Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. Sci Rep. 2018;8:1227.
17. Solis AD. Reduced alphabet of prebiotic amino acids optimally encodes the conformational space of diverse extant protein folds. BMC Evol Biol. 2019;19:158.
18. Kimura M, Akanuma S. Reconstruction and characterization of thermally stable and catalytically active proteins comprising an alphabet of ~13 amino acids. J Mol Evol. 2020;88:372–381.
19. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett. 2008;15: 956–963.
20. Burley SK, Petsko GA. Aromatic-aromatic interaction: A mechanism of protein structure stabilization. Science. 1985;229:23–28.
21. Di Mauro E, Dunker AK, Trifonov EN. Disorder to order, nonlife to life: In the beginning there was a mistake. In: Seckbach J, editor. Genesis—In the beginning. New York: Springer, 2012; p. 415–435.
22. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins. 2001;42:38–48.
23. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 2014;83:553–584.
24. Bukhari SA, Caetano-Anollés G. Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. PLoS Comput Biol. 2013;9: e1003009.
25. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinform. 2006;7:208.
26. Brooks DJ, Fresco JR. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol Cell Proteomics. 2002;1:125–131.
27. Sumbalova L, Stourac J, Martinek T, Bednar D, Damborsky J. HotSpot wizard 3.0: Web server for automated design of mutations and smart libraries based on sequence input information. Nucleic Acids Res. 2018;46:W356–W362.
28. Nurkanto A, Jeelani G, Yamamoto T, et al. Biochemical, metabolomic, and genetic analyses of dephospho coenzyme a kinase involved in coenzyme a biosynthesis in the human enteric parasite *Entamoeba histolytica*. Front Microbiol. 2018;9:2902.
29. Seto A, Murayama K, Toyama M, et al. ATP-induced structural change of dephosphocoenzyme a kinase from Thermus thermophilus HB8. Proteins. 2005;58:235–242.
30. Semisotnov GV, Rodionova NA, Kutyshenko VP, Ebert B, Blanck J, Ptitsyn OB. Sequential mechanism of refolding of carbonic anhydrase B. FEBS Lett. 1987;224:9–13.
31. Walia G, Surolia A. Insights into the regulatory characteristics of the mycobacterial dephosphocoenzyme a kinase: Implications for the universal CoA biosynthesis pathway. PLoS One. 2011;6:e21390.
32. Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. Elife. 2015;4:e09410.
33. Longo LM, Jabłońska J, Vyas P, et al. On the emergence of P-loop NTPase and Rossmann enzymes from a Beta-alpha-Beta ancestral fragment. Elife. 2020;9:e64415.
34. Romero Romero ML, Yang F, Lin YR, et al. Simple yet functional phosphate-loop proteins. Proc Natl Acad Sci USA. 2018; 115:E11943–E11950.
35. Vyas P, Trofimyuk O, Longo LM, Deshmukh FK, Sharon M, Tawfik DS. Helicase-like functions in phosphate loop containing beta-alpha polypeptides. *BioRxiv*. https://doi.org/10.1101/2020.07.30.228619
36. Vamvaca K, Vögeli B, Kast P, Pervushin K, Hilvert D. An enzymatic molten globule: Efficient coupling of folding and catalysis. Proc Natl Acad Sci USA. 2004;101:12860–12864.
37. Pervushin K, Vamvaca K, Vögeli B, Hilvert D. Structure and dynamics of a molten globular enzyme. Nat Struct Mol Biol. 2007;14:1202–1206.
38. Walter KU, Vamvaca K, Hilvert D. An active enzyme constructed from a 9-amino acid alphabet. J Biol Chem. 2005;280: 37742–37746.
39. Sapienza PJ, Li L, Williams T, Lee AL, Carter CW Jr. An ancestral tryptophanyl-tRNA synthetase precursor achieves high catalytic rate enhancement without ordered ground-state tertiary structures. ACS Chem Biol. 2016;11:1661–1668.
40. Bonfio C, Valer L, Scintilla S, et al. UV-light-driven prebiotic synthesis of iron-sulfur clusters. Nat Chem. 2017;9: 1229–1234.
41. Weber AL, Pizzarello S. The peptide-catalyzed stereospecific synthesis of tetroses: A possible model for prebiotic molecular evolution. Proc Natl Acad Sci USA. 2006;103:12713–12717.
42. Longo LM, Tenorio CA, Kumru OS, Middaugh CR, Blaber M. A single aromatic core mutation converts a designed "primitive" protein from halophile to mesophile folding. Protein Sci. 2015;24:27–37.
43. Longo LM, Despotović D, Weil-Ktorza O, et al. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. Proc Natl Acad Sci USA. 2020;117:15731–15739.
44. Longo LM, Blaber M. Protein design at the interface of the prebiotic and biotic worlds. Arch Biochem Biophys. 2012;526: 16–21.
45. Longo LM, Blaber M. Prebiotic protein design supports a halophile origin of foldable proteins. Front Microbiol. 2014;4:418.
46. Huang F, Oldfield C, Meng J, et al. Subclassifying disordered proteins by the CH-CDF plot method. Pac Symp Biocomput. 2012;17:128–139.
47. Sreerama N, Woody, RW. Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. Anal Biochem. 2000;287:252–260.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Makarov M, Meng J, Tretyachenko V, et al. Enzyme catalysis prior to aromatic residues: Reverse engineering of a dephospho-CoA kinase. *Protein Science*. 2021;1–13. https://doi.org/10.1002/pro.4068