Charles University, Faculty of Science
Department of Biochemistry

Ph.D. study program: Biochemistry

Summary of the Ph.D. thesis

The effect of amino acid repertoire on protein structure evolution

Mgr. Vjačeslav Treťjačenko

Supervisor: Mgr. Klára Hlouchová, Ph.D.

Prague 2021

Abstract

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.

Introduction

Evolution of proteins and their amino acid alphabet

Contemporary proteins are the most versatile molecules of life. The evolution of their structure dates back to the simplest, short oligopeptides able to support basic catalytic functions and perhaps to interact with the primordial RNA. First proteins probably consisted of limited amino acid set representing the prebiotic chemical variety [1,2]. These amino acids could have come from endogenous and exogenous sources. Two independent meta-analyses on temporal order of amino acid incorporation into the proteins were carried out by compiling the results from numerous reports on amino acid formation in prebiotic chemistry experiments, analyses of meteorite composition and hydrothermal chemistry dynamics [3,4]. These meta-analyses agree on approximately ten amino acids being the prebiotically available to the first protein formation, with the rest being gradually incorporated via biosynthetic pathways. These analytical conclusions were further confirmed by the observations of increased occurrence of prebiotically plausible amino acids in the vicinity of enzyme's active centers and with the non-essentiality of some evolutionary younger amino acids in ancestrally reconstructed proteins and pathways [5–7].

Evolution of sequence, structure and function from the simple prebiotic oligopeptides presents a thought-provoking conundrum. Even a short 100 amino acid protein with the 10 amino acid alphabet can be constructed in overwhelming 10^{100} possible ways. Several studies focused on identification of the earliest protein sequences and structures. Alva *et al.* analyzed sets of non-homologous protein domains and identified 40 short sequences shared in otherwise unrelated proteins [8]. Moreover, 14 of these conserved short sequences establish folds by repetition. Alternatively, Caetano-Annoles and coworkers investigated the fold usage across all three domains in life and derived the temporal order of protein fold emergence based on phylogenetic analysis. The authors discovered that most ancient protein structures included P-loop NTPases (SCOP fold c.37), TIM β/α 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. 1**) [9–12]

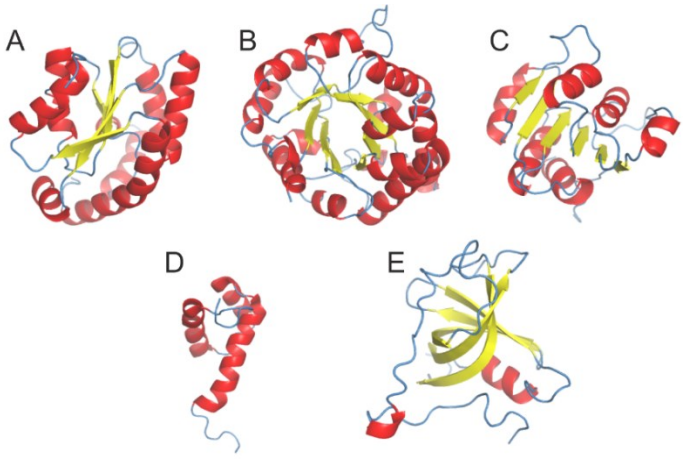


Figure 1. Most ancient protein folds by Caetano-Annoles and coworkers. (A) P-loop NTPase, (B) TIM β/α barrel, (C) NAD(P)-binding Rossman fold, (D) DNA/RNA binding 3-helical bundle, (E) ololigonucleotide/oligosaccharide binding fold

Functions of these folds are notably connected with carbohydrate and nucleotide metabolism. Indeed, another study of Goldman *et al.* identified that 9 of 10 of the most ancient folds described by Caetano-Annoles *et al.* are widespread in translation-related proteins and were involved in functions such as RNA modification, binding, and phosphoryl transfer [13].

In summary, protein evolution can be traced back in time by the analysis of contemporary proteins and reconstruction of most widely shared fragments. However, the origin and identity of the first sequences which gave rise to the modern protein world remains obscured.

Chaperones and cofactors in protein evolution

The complexity of the contemporary folding pathways has grown over the past 4 billion years and gave rise to the sophisticated folding assistance machinery represented by chaperones [14]. These molecules, however, are also the products of protein structure and function evolution. It was hypothesized that organic and metallic cofactors could and still play a crucial stabilizing role in protein structure and activity [15]. Interestingly, the 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) [16]. Thus, it was proposed that protein function and its conformation could have been selected by ligand binding from the early pool of disordered proteins [17]. Furthermore, it was recently demonstrated that ancient organic cofactor 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 [18].

Moreover, modern protein chaperones were shown to be evolutionary capacitors allowing novel and still evolving proteins to reach otherwise impossible folding trajectories and avoid detrimental aggregation. The interaction of protein with chaperones was shown to be directly correlated with protein age and evolutionary rate [19–21]. Houben *et al.* hypothesised that chaperone emergence is connected with the amino acid alphabet evolution and suggested that introduction of basic amino acids coincided with the first chaperoning activity [22]. Indeed, the expansion of the world of protein sequences and structures is related to higher intracellular chaperone concentrations and with the establishment of complex interaction networks of chaperones and co-chaperones [14].

Here we investigate an effect of chaperones on random unevolved sequences as a model of novel and prebiotic proteins. In addition, we provide an insight into the interaction between ancient organic cofactor and a simplified variant of a contemporary protein.

Characteristics of unevolved sequence space

Natural protein sequences altogether represent only a minute fraction of the possible sequence space. Do alternative structures and functions exist beyond the known? Several computational and experimental studies approached the unevolved sequence space exploration. It has been suggested that

structured molecules frequently occur in random sequence space and that their structural repertoire is comparable to the natural protein universe [23,24]. Experimental investigations of random proteins with different amino acid alphabets confirmed frequent occurrence of secondary and tertiary structures in unevolved sequences. Several studies on random proteins with reduced amino acid alphabets pointed out that proteins consisting of prebiotically plausible amino acids have higher solubility than their 20 amino acid counterparts [25,26]. Along with the investigation of structure and solubility of random proteins, an attempt to screen for a simple protein function – ATP-binding was made. Using mRNA display technique, Keefe and Szostak detected 4 ATP binders in the library of 6×10^{12} random sequences. Interestingly, the structure of one ATP-binder revealed a completely unknown, flexible, and a metal-cofactor dependent protein fold [27,28].

In addition, exploration of unevolved sequence space has a direct relevance to novel proteins generated by contemporary organisms *de novo*. The genes for these proteins emerge from previously untranscribed and untranslated parts of the genomes, thus novel sequences do not have any homology to the known proteins and their levels of evolutionary optimization are on par with random proteins (reviewed in [29]). Moreover, a recent study has shown that random or pseudo-random sequences could affect the organismal fitness both in negative and positive ways suggesting the importance of *de novo* gene formation in contemporary evolutionary processes [30].

Combinatorial protein library design

The most efficient path for investigation of collective protein properties is via the combinatorial library approach. Construction of protein mutant libraries is a traditional methodology in protein engineering, design, and selection. Contemporary computational methods facilitate construction of such libraries based on various experimentalist demands including cost efficiency and thermodynamic stability or even functional promiscuity [31–35]. However, the common feature of these algorithms consists in a rational approach to the library variability minimalization in order to screen only pre-selected mutagenized positions. For effective random protein screening an alternative approach is required.

Here we designed an algorithm for combinatorial DNA template construction coding the vast randomized protein libraries constrained only by their amino acid composition. This tool fills the gap in the modern repertoire of computational techniques and may serve for studies of amino-acid content dependent protein properties and dynamics of *de novo* gene formation from organisms with distinct genomic nucleotide compositions.

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.

Results and discussion

Scarce sampling of random sequence space

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 any 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.

We generated 4 *in silico* datasets with 10 000 protein sequences of 100 amino acids to investigate properties of random proteins and compare them to their natural counterparts. We used 5 secondary structure predictors, 3 protein disorder predictors and one 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).

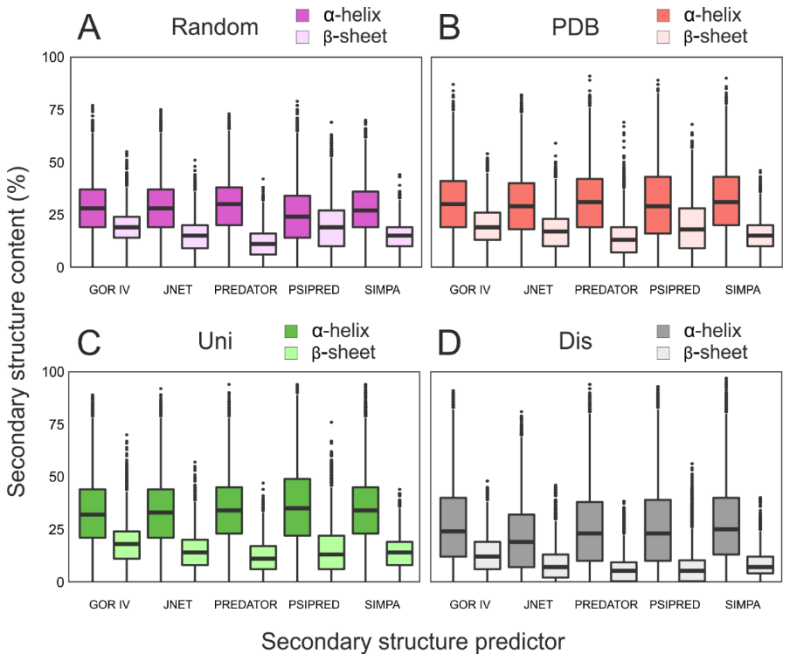


Figure 2. Predictions of secondary structure occurrence in the (A) Random, (B) PDB, (C) Uni, and (D) Dis datasets. α-helical and β-sheet content was determined by five different predictors. The center of the box represents the median; upper, and lower borders are 3rd and 1st quartile respectively. The solid lines illustrate maximum and minimum contents, which are shown as dots. The Dis dataset is included as a negative reference

The results of the analysis showed that although secondary structure content in random sequence space is not significantly different from the natural proteins, optimization of protein aggregation is higher in natural proteins in comparison with random sequences (**Fig. 2**). Based on structural and solubility prediction, 45 proteins from random sequence pool were selected for the following experimental characterization. The expression and solubility analysis demonstrated that the most tolerable and soluble proteins are of a disordered nature (**Fig. 3**). The structural characterization of the purified

random proteins showed a good agreement with structure and aggregation predictions. Moreover, dynamic light scattering of purified random proteins confirmed the computationally derived conclusion of direct correlation between random protein structure content and its aggregation tendency.

In summary, this study emphasized the importance of disordered proteins on either prebiotic or contemporary *de novo* protein formation and thoroughly validated the efficacy of computational secondary structure/aggregation prediction beyond the natural protein space.

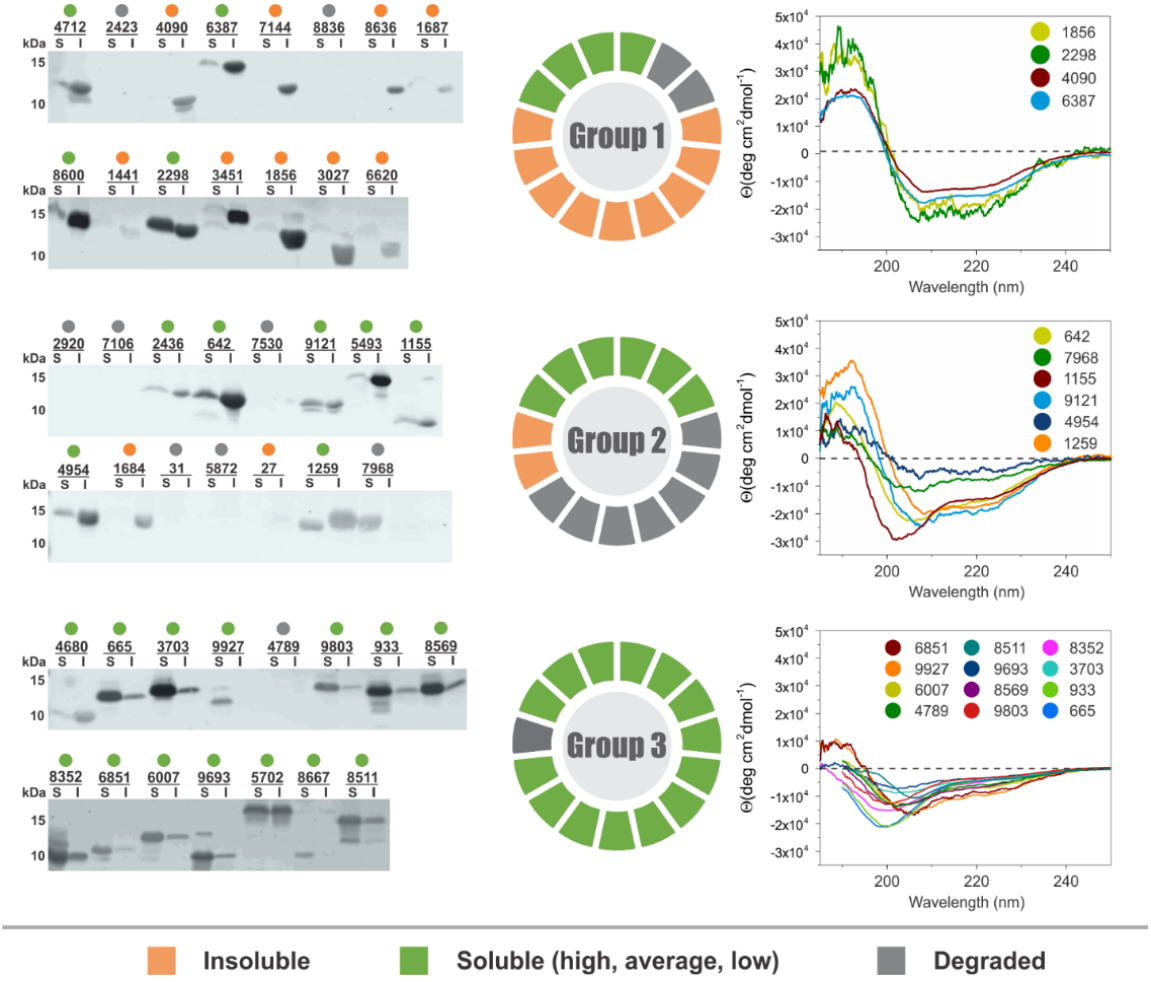


Figure 3. Summary of expression/solubility analyses and circular dichroism spectra of random proteins from Group 1, 2 and 3. **(Left)** western blot expression analysis of NBP's in *E. coli*. S – soluble fraction of the lysate, I - insoluble fraction; **(Middle)** a pie graph summarizing the solubility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

Development of combinatorial library design tool (CoLiDe)

Following the scarce characterization of random protein space, we decided to undertake an investigation of collective protein structural features via the library approach. Unfortunately, existing 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 the sequence. For that reason, we implemented the combinatorial library design (CoLiDe) tool which is optimized for a computationally efficient and accessible diverse library construction.

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 a 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 the user 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 a defined amino acid distribution.

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 $1000 \times l$ 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. 4**.

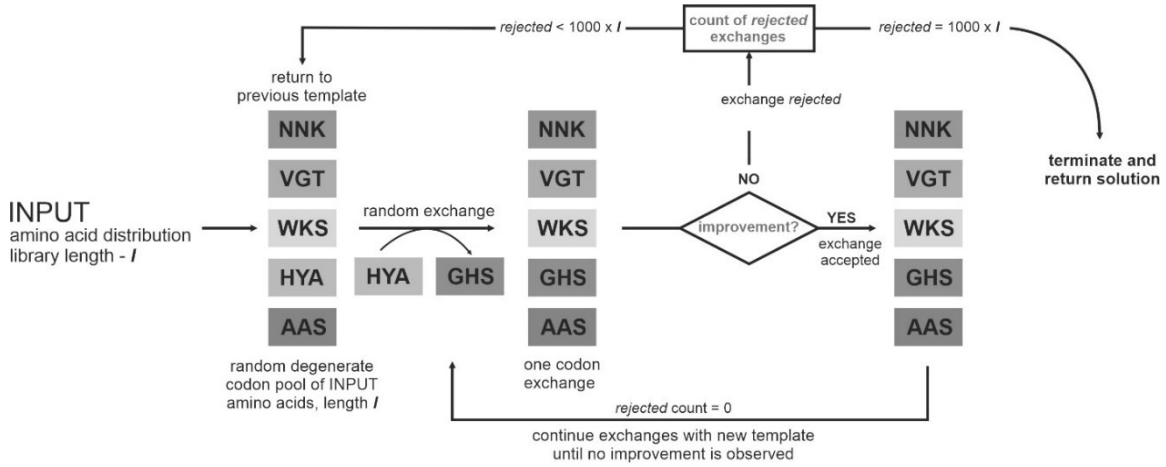


Figure 4. Schematic representation of CoLiDe computational pipeline. Input is user defined library length and amino acid distribution of protein library. Program filters degenerate codon pool and excludes all codons which code non defined amino acids. Subsequently algorithm randomly generates a string of degenerate codons of the library length and introduces codon exchanges until the input amino acid distribution is approximated

The algorithm performance was tested on a 45 amino acid library with a variable 33 amino acid part. The library was synthesized as a semi-degenerate oligonucleotide string, converted to double stranded DNA template by Klenow reaction and translated via cell free protein expression system. The library was purified by His-tag affinity chromatography, its molecular weight distribution was evaluated by MALDI-TOF mass spectrometry and its amino acid composition verified by amino acid analysis (**Fig. 5**). The resulting degenerate protein library showed a good agreement in the mass distribution and amino acid composition in comparison with the designed template.

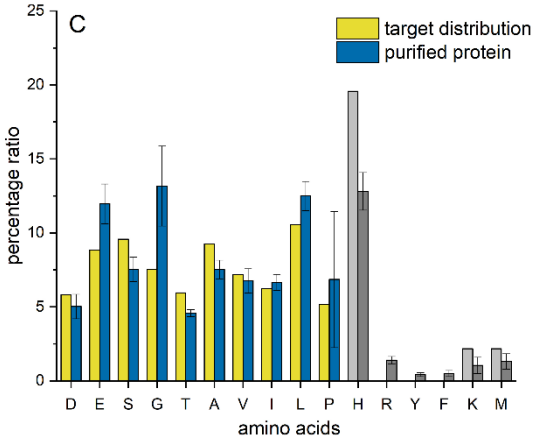


Figure 5. Amino acid analysi of the purified protein library and its comparison to the designed distribution

In summary, we implemented and experimentally validated a computational tool for diverse degenerate library construction and provided the executable code for the broad scientific community.

Characterization of combinatorial protein libraries with distinct amino acid alphabets

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, 20F) and 10 (early set A, S, D, E, G, T, I, L, P and V, 10E) 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.

We designed 106 amino acid long libraries with 85 amino acid variable part, thrombin cleavage site in the middle of the protein sequence and two tag sequences for the affinity purification and detection by a chemiluminiscent antibody. The libraries were ordered as two overlapping single stranded DNA oligonucleotides, converted to double stranded DNA template, and produced by cell free expression system. We tested the effects of DnaK chaperone system on library solubility and potential structure formation. Moreover, we used selective proteolysis to assess the total folding state of the library proteins by unspecific Lon protease and specific thrombin cleavage on the central part of the molecule.

The preliminary results from solubility and proteolysis assays suggest the different interaction modes of library 20F and 10E (**Fig. 6, Fig. 7**). Library 20F shows significant solubilization and protease resistance upon the chaperone treatment in all tested temperatures of 25, 30 and 37 °C (**Fig. 6B**).

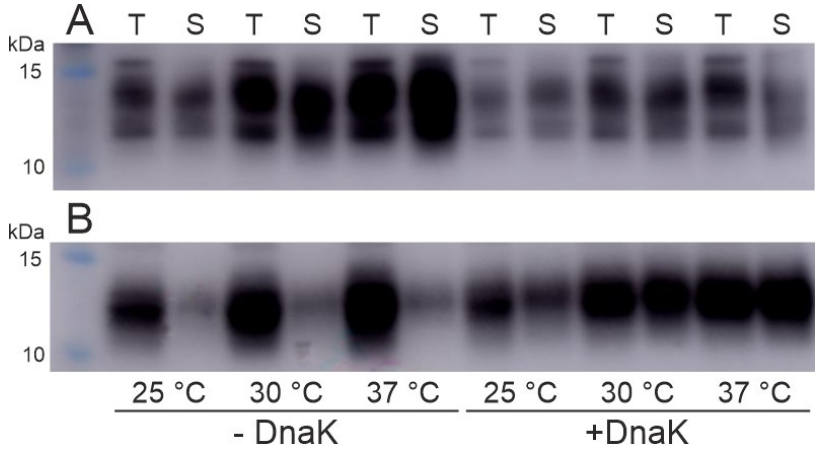


Figure 6. Western blot analysis of 10E (A) and 20F (B) solubilities in 25, 30 and 37 °C. Reactions were performed in absence (-DnaK) and presence (+DnaK) of chaperones. Equal volumes of total (T) and soluble (S) reaction products were analysed

Moreover, in the absence of chaperones soluble fraction of library 20F appears to be effectively cleaved by the Lon protease which specifically cleaves unfolded proteins (**Fig. 7B**). This signifies that

without chaperone assistance, most of the 20F proteins appear to be insoluble or in an unfolded state. With addition of chaperones, 20F random proteins become completely solubilized. However, most of this soluble protein fraction is unfolded as is indicated by the analysis of chaperone and Lon supplemented reaction (**Fig. 7B**). Nevertheless, chaperones do provide a significant protection against Lon protease which is manifested in higher fraction of soluble and uncleaved proteins in Lon/chaperone supplemented reaction (**Fig. 7B**). In contrast, the library 10E is highly soluble independently on chaperone supplementation (**Fig 6A, Fig 7A**). Moreover, chaperones do not provide any protection against proteolysis by Lon as was concluded from the comparison of Lon supplemented and both chaperone absent and supplemented reactions (**Fig. 7A**). Interestingly, chaperone supplementation appears to suppress expression rates of the 10E library (**Fig. 6A, Fig. 7A**). This initial screening suggests that although the occurrence of folded or compact structures is present in both libraries under some circumstances, the lower digestion rates of 10E proteins might indicate higher occurrence of the early-alphabet protein folding.

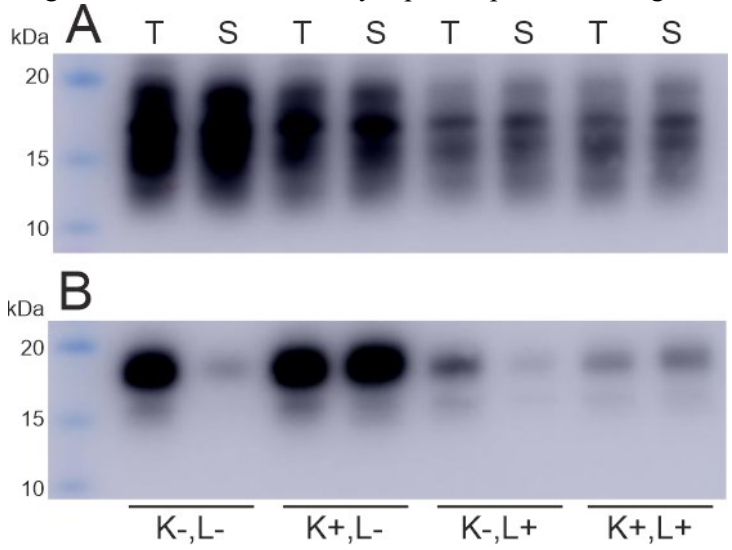


Figure 7. Western blot analysis of 10E (A) and 20F (B) libraries solubilities in co-translational presence/absence of chaperones (K+/K-) and Lon protease (L+/L-). Same volumes of total (T) and soluble (S) reaction products were analysed

These preliminary results will be further verified by quantitative western blotting of protein digestions in different conditions and by analytical gel chromatography.

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

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. 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.

We successfully expressed the wild type DPCK from *Aquifex aeolicus* (DPCK-WT) and two mutant variants with either all aromatic amino acids except histidine substituted with leucine (DPCK-LH) or rationally designed variant where all aromatics including histidine were substituted by prebiotically plausible hydrophobic amino acids (DPCK-M) (**Fig. 8A**)

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

B

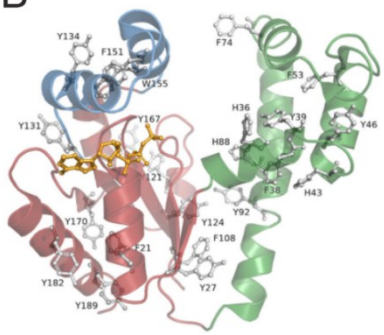


Figure 8. Summary of mutated aromatic residues and their positions in comparison to the DPCK-WT sequence (A) and (B) structure of the DPCK enzyme with highlighted aromatic amino acids (grey) and bound ATP substrate (yellow)

All proteins were expressed in *E. coli* BL21-DE3, purified to homogeneity via a three step purification protocol and their catalytic activity and efficiency were compared. DPCK-WT showed similar activities to the previously reported DPCK from *E. histolytica*. While DPCK-WT and -LH had both ATP hydrolytic and dCoA-dependent phosphotransferase activity, DPCK-M variant exhibited only ATPase activity. Furthermore, upon DPCK-LH catalysis 100× less CoA was detected in comparison with the DPCK-WT.

Structural characterization of all DPCK variants by electronic circular dichroism and 1D/2D HN NMR spectroscopies showed higher disorder content in mutant variants in comparison to DPCK-WT. 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 a lack of tertiary structure formation.

Furthermore, limited proteolysis of DPCKs by LysC protease indicated different proteolysis dynamics of mutant variants in comparison to protease resistant DPCK-WT enzyme. 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. 9**).

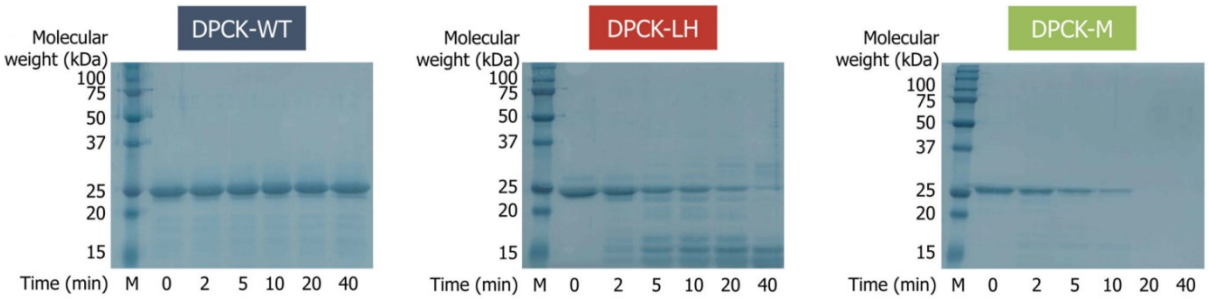


Figure 9. 14% SDS-polyacrylamide gels of limited proteolysis of dephospho CoA kinase (DPCK) proteins visualized by imidazole-zinc staining after SDS-PAGE with the protein samples exposed to LysC endoproteinase for different times

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 reached that of DPCK-WT value upon ATP addition (**Table 1**). This observation supports the previously stated hypotheses that cofactors might play a crucial role in early protein structure stabilization.

Table 1. Summary of DLS measurement of DPCK-WT and DPCK-LH variants with and without ATP

	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

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. 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).

2. 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).

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

4. 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).

5. 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).

6. 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).

7. 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).

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

9. Caetano-Anollés, G. & Caetano-Anollés, D. An evolutionarily structured universe of protein architecture. *Genome Res.* **13**, 1563–1571 (2003).

10. 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).

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

12. 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).

13. 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).

14. 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.

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

16. 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).

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

18. Longo, L. M., Despotovi, D., Weil-ktorza, O., Walker, M. 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. **117**, (2020).

19. 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).

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

21. 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).

22. 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).

23. 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 Suppl 6**, S22 (2009).

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

25. 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).

26. 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).

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

28. 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).

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

30. 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).

31. 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).

32. 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).

33. 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).

34. 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).

35. 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).

Univerzita Karlova, Přírodovědecká fakulta
Katedra biochemie

Doktorský studijní program: Biochemie

Autoreferát disertační práce

Vliv repertoáru aminokyselin na strukturu a funkci bílkovin

Mgr. Vjačeslav Treťjačenko

Školitelka: Mgr. Klára Hlouchová, Ph.D.

Praha 2021

Abstrakt

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. Charakterizovali jsme *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 prebioticky dostupných 10 aminokyselin. Navrhli jsme 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 v časech prvotního vývoje biosféry.

Úvod

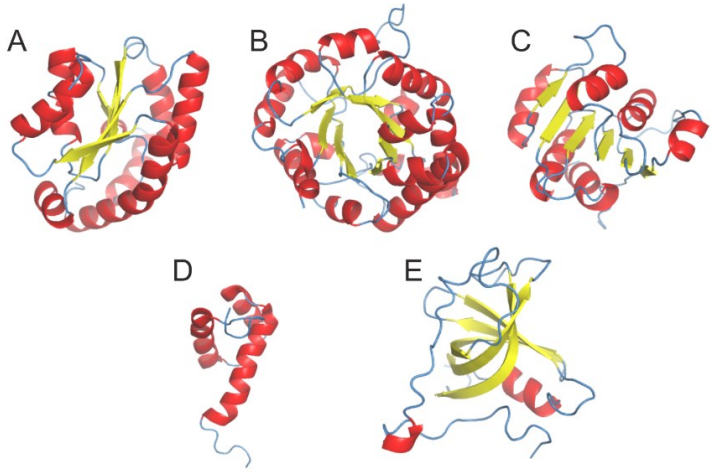
Evoluce proteinů a jejich aminokyselinového repertoáru

Současné proteiny jsou nejvšestrannějšími molekulami života. Počátky jejich strukturní evoluce pramení z krátkých oligopeptidů schopných jednoduché katalýzy a pravděpodobně také interakcí s prvotními molekulami RNA. Nejranější proteiny nejspíše disponovaly omezeným aminokyselinovým repertoárem který reprezentoval chemické poměry prebiotického prostředí [1,2]. Tyto aminokyseliny mohly pocházet jak z endogenních tak z exogenních zdrojů. Časové zařazení aminokyselin do strukturního repertoáru proteinů bylo odvozeno ve dvou na sobě nezávislých metaanalýzách na základě zpracování experimentálních dat ze simulací prebiotického prostředí, analýz složení meteoritů a chemických pochodů v okolí oceánských hydrotermálních pramenů [3,4]. Tyto metaanalýzy se shodují v přiřazení přibližně deseti aminokyselin k prebioticky dostupným zdrojům a v odvození dalších deseti prostřednictvím biosyntetických pochodů. Tyto závěry byly dale utvrzeny analýzou struktur enzymů, jež vykazují vysokou míru obohacení prebioticky dostupnými aminokyselinami v okolí jejich aktivních center. Dodatečně bylo ukázáno, že některé pozdější aminokyseliny jsou postradatelné pro tvorbu jejich vlastních biosyntetických proteinů [5-7].

Evoluce proteinové sekvence, struktury a funkce z jednoduchých oligopeptidů představuje jeden z mnoha dosud neobjasněných přírodních hlavolamů. Pro přiblížení, krátký protein složený z pouhých 100 aminokyselin s 10 aminokyselinovým repertoárem může být poskládán 10^{100} různými kombinacemi. Identifikací prvotních proteinových sekvencí a struktur se zabývalo několik studií. V jedné z nich, Alva *et al.* vyhodnotil soubor nehomologních proteinových domén a odhalil 40 krátkých sekvencí jež byly sdíleny mezi nepříbuznými proteiny [8]. Kromě toho, 14 z těchto sekvencí vykázaly schopnost tvorby nativní proteinové struktury prostřednictvím repetice. V jiné práci, Caetano-Annoles se spolupracovníky se zaměřil na využití proteinových strukturních podjednotek (angl. folds) napříč všemi říšemi života. Autorům se pomocí fylogenetické analýzy podařilo odvodit časovou osu vzniku proteinových podjednotek, mezi nejstaršími strukturami byly ATPázy obsahující Walkerův motiv (SCOP fold c.37), TIM β/α barel (c.1), NAD(P) vazebné domény s Rossmanovou strukturou (c.2), DNA/RNA vazebné trihelikální svazky (a.4) a oligonukleotid/oligosacharid vazebné motivy (b.40) (**Obr. 1**) [9-12].

Funkce těchto proteinů jsou nejčastěji spojeny s metabolismem uhlovodíků a nukleotidů. Goldman *et al.* ukázal, že 9 z 10 nejstarších proteinových struktur ze studie Caetana-Annollese jsou hojně zastoupeny v proteinech souvisejících s translací a jejich funkce jsou často reprezentovány modifikací RNA, její vazbou a přenosem fosfátové skupiny [13].

Na závěr této kapitoly lze říci, že evoluce proteinů se dá vystopovat pomocí rekonstrukce nejvíce zastoupených a tedy i nejkonzerovanějších proteinů až ke krátkým sekvenčním a strukturním prekurzorům současných proteinů. Nicméně, identita prvních proteinových sekvencí jež stály u zrodů všech moderních proteinů je i nadále zahalena rouškou neznáma.



Obrázek 1. Nejstarší proteinové struktury dle Caetana-Annolese a spolupracovníků. (A) ATPáza obsahující Walkerův motiv (SCOP fold c.37), (B) TIM β/α barel (c.1), (C) NAD(P) vazebné domény s Rossmanovou strukturou (c.2), (D) DNA/RNA vazebné trihelikální svazky (a.4) a (E) oligonukleotid/oligosacharid vazebné motivy (b.40)

Role chaperonů a kofaktorů v proteinové evoluci

Mechanismus provázející dosažení nativní proteinové konformace je výsledkem 4 miliard dlouhé evoluce, která dala vznik sofistikovaným proteinovým asistentům – chaperonům [14]. Nicméně, vznik těchto molekul je také spojen s jejích samotnou evolucí struktury a funkce. Je předpokládáno, že organické a kovové kofaktory mohly hrát podobnou roli na proteinovou strukturu jak v časech dávno minulých tak i v současnosti [15]. Stojí za zmínku, že volná energie uvolněná vazbou kofaktorů k proteinům je v průměru podobná té, uvolněné při pochodech spojených s dosažením nativní proteinové konformace (~10-15 kcal/mol při vazbě kofaktoru a ~10-20 kcal/mol při skládání proteinů) [16]. Tyto předpoklady vedly k domněnce, že funkce a konformace proteinů mohla být selektována ze souborů prebiotických neuspořádaných sekvencí právě pomocí vazby kofaktorů [17]. V nedávné studii bylo také demonstrováno, že odvěký organický kofaktor ATP stimuluje koacervaci komplexu peptid/DNA jež je velmi diskutovaným mechanismem vzniku prvotních molekulárních kompartmentů [18].

Na druhou stranu, současné proteinové chaperony se projevují evolučními zprostředkovateli umožňujícími dosažení nativní proteinové konformace i stále se vyvíjejícím proteinům vystaveným riziku nespecifické agregace. Chaperon-proteinová interakce byla prokazatelně spojena s věkem a rychlosti evoluce proteinů [19-21]. Houben *et al.* poukázal na možnou souvislost vzniku proteinových chaperonů s

evolucí aminokyselinového repertoáru [22]. Podle této studie je inkorporace bazických aminokyselin přímo spojena s vývojem chaperonů. Tento vztah je dále podpořen analýzami globálního vývoje proteomů, jež zachycují evoluční rozvoj a expanzi proteomů v souvislosti s vyšší intracelulární abundancí chaperonů a rozvojem komplexních interakcí chaperonu a ko-chaperonů [14].

V této práci jsme se zaměřili na studium vlivů chaperonů na náhodné proteinové sekvence které zde slouží modelovým příkladem nových a původních prebiotických proteinů. Navíc je v této práci uveden názorný příklad jak interakce prastarého organického kofaktoru ovlivňuje strukturu zjednodušeného moderního proteinu.

Charakteristiky náhodného sekvenčního prostoru

Počet všech přírodních proteinových sekvencí představuje zanedbatelný podíl ze všech možných proteinů. Je kompaktní proteinová struktura a funkce dosažitelnou mimo tento přírodní sekvenční prostor? Průzkumem náhodného sekvenčního prostoru se zabývalo několik výpočetních a experimentálních prací. Teoretické analýzy naznačují vysoký výskyt strukturovaných proteinů v náhodném sekvenčním prostoru a jejich podobnost přírodnímu strukturnímu repertoáru [23,24]. Tyto anlýzy jsou v souladu s experimentálními charakterizacemi náhodných proteinů s různými aminokyselinovými repertoáry, které potvrzují vysoký výskyt sekundárních a terciárních struktur v náhodném sekvenčním prostoru. Několik studii náhodných proteinů složených z prebioticky dostupných aminokyselin poukazují na jejích vyšší rozpustnost ve srovnání s jejích 20 aminokyselinovými protějšky [25,26]. Paralelně s průzkumem strukturního potenciálu náhodných proteinů byl také studován výskyt funkčních molekul v náhodném sekvenčním prostoru. Keefe a Szostak izolovali 4 ATP vazebné proteiny z knihovny náhodných sekvencí čítající 6×10^{12} molekul. Jeden z těchto vazebných proteinů byl charakterizován strukturně, což vedlo k odhalení dosud neznámé flexibilní proteinové struktury obsahující kovový kofaktor [27,28].

Náhodné proteinové sekvence mají také přímou relevanci k přírodním proteinům generovaným *de novo* z dříve nekódujících části genomů. Tyto proteiny nejsou homologní s žádnou dosud známou rodinou proteinů a jejich stupeň evoluční optimalizace se dá srovnat s náhodnými proteinovými sekvencemi (shrnuto v [29]). V nedávné studií bylo ukázáno, že náhodné a pseudo-náhodné proteiny mohou ovlivnit životaschopnost organismu jak negativním tak i pozitivním směrem což poukazuje na důležitost formace *de novo* proteinů v současné evoluci [30].

Návrh kombinatoriálních proteinových knihoven

Nejefektivnějším způsobem průzkumu kolektivních vlastností proteinů je prostřednictvím kombinatoriálních proteinových knihoven. Příprava proteinových knihoven je nedílnou součásti racionálního návrhu proteinů a proteinového inženýrství. Současné výpočetní metody umožňují návrh na základě různorodých experimentálních požadavků jako je cenová dostupnost, termodynamická stabilita nebo dokonce funkční promiskuita proteinových produktů [31-35]. Nicméně společnou charakteristikou všech těchto přístupu je návrh co nejkompaktnější knihovny za účelem maximálního pokrytí všech variant s několika mála variabilními pozicemi. Pro účely charakterizace náhodných proteinů, u kterých není kladen důraz na specifickou sekvenci, je vyžadován alternativní přístup. V této práci jsme navrhli a implementovali výpočetní algoritmus pro návrh vysoce komplexních kombinatoriálních knihoven náhodných proteinů se specifikovatelnými poměry aminokyselin. Tento nástroj je doplňkem k stávajícímu arzenálu výpočetních metod a slouží ke studiu vlastnosti proteinů s charakteristikami danými spíše aminokyselinovým složením než jejich specifickou sekvencí.

Cíle práce

Cílem této práce bylo (i) studovat vlastnosti náhodného sekvenčního prostoru a vztah náhodných proteinových sekvencí k přirozeným proteinům a (ii) studovat vliv aminokyselinového repertoáru na strukturu a funkci bílkovin.

Konkrétními cili byly:

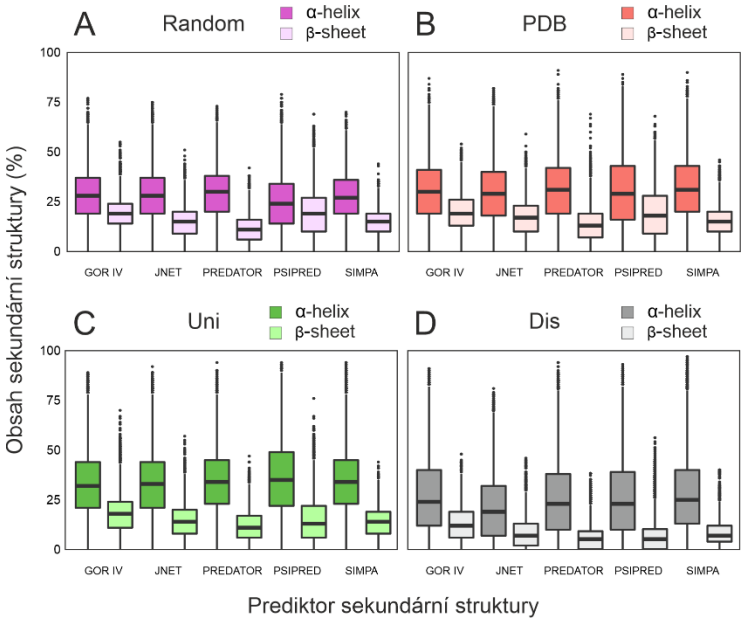
- Prozkoumat strukturní obsah náhodného sekvenčního prostoru a charakterizovat vybrané náhodné proteiny *in vitro.*
- Implementovat výpočetní nástroj, který by umožnil návrh degenerované proteinové knihovny se zadanými aminokyselinovými poměry.
- Experimentálně charakterizovat knihovny náhodných proteinových sekvencí s odlišnými aminokyselinovými poměry.
- Prostudovat vliv nejnovějších přírůstků do aminokyselinového repertoáru na strukturu a funkci vybraného proteinu.

Výsledky a diskuze

Selektivní charakterizace náhodného sekvenčního prostoru

Současné proteiny jsou výsledkem procesu evoluční optimalizace trvající 4 miliardy let. Naše znalosti proteinové struktury, funkce a evoluce vychází z experimentálních a teoreticky založených analýz přírodních proteinů. Nicméně, vlastnosti proteinů jež neprošly evolučním vývojem a optimalizací nejsou probádány. V této práci jsme provedli systematickou výpočetní a experimentální analýzu náhodných proteinů s kanonickým aminokyselinovým repertoárem a prozkoumali jejich vztah k přirozeným proteinům.

Za účelem srovnání náhodných proteinů s jejích přírodními protějšky jsme připravili 4 *in silico* datasety o 10 000 proteinových sekvencích o délce 100 aminokyselin. S pomoci 5 prediktorů sekundárních struktur, 3 prediktorů proteinové neuspořádanosti a jednoho prediktoru proteinové agregace jsme srovnali knihovny (A) náhodných proteinů s přirozenými výskyty aminokyselin (Random), (B) fragmentů přirozených proteinů z databáze TOP8000 neredundantních strukturně charakterizovaných proteinů z databáze PDB (PDB), (C) fragmentů přirozených proteinů z databáze UniProt (Uni) a (D) fragmentů přirozených vnitřně neuspořádaných proteinů z databáze Disprot (Dis).



Obrázek 2. Predikce výskytu sekundárních struktur v datasetech (A) Random, (B) PDB, (C) Uni, and (D) Dis. Obsah α-helixů a β-listů byl stanoven pěti různými prediktory. Střed krabicového diagramu představuje medián; horní a spodní okraj krabice jsou třetím a prvním kvartilem. Čára znázorňuje nejvyšší a nejnižší obsahy struktur které jsou vyznačeny tečkami. Dataset Dis byl zahrnut jako negativní kontrola

Výsledky této analýzy naznačují, že přestože obsah sekundárních struktur v náhodném sekvenčním prostoru není výrazně odlišný od přirozených proteinů, stupeň potlačení proteinové agregace je vyšší u přírodních sekvencí ve srovnání s proteiny náhodnými (**Obr. 2**).

Vybrali jsme 45 náhodných proteinových sekvencí pro další experimentální charakterizaci v návaznosti na předchozí analýzu obsahu sekundárních struktur a neuspořádanosti. Ukázali jsme, že nejvíce exprimovanými a rozpustnými náhodnými proteiny *in vivo* jsou sekvence postrádající stabilní obsah sekundárních struktur (**Obr. 3**). Strukturní charakterizace purifikovaných náhodných proteinů vykázala vysokou míru shody s výpočetními

predikcemi obsahu sekundárních struktur a míry agregace. Měření dynamického rozptylu světla purifikovaných vzorků náhodných proteinů potvrdilo dříve odvozenou přímou korelaci mezi obsahem sekundárních struktur a mírou agregace proteinu.

Obrázek 3. Souhrn z testování expresí/solubilit a spektra cirkulárního dichroismu náhodných proteinů ze Skupiny 1, 2 a 3. **(Vlevo)** analýza solubilit proteinů exprimovaných v *E. coli* pomocí metody western blot. S – solubilní frakce lyzátu, I – nerozpustná frakce; **(Střed)** koláčové grafy shrnující počty solubilních proteinů v každé skupině; **(Vpravo)** spektra cirkulárního dichroismu úspěšně připravených náhodných proteinů ze Skupin 1 až 3

Výsledky této práce vytyčili význam neuspořádaných sekvencí jak v prebiotické tak i současné *de novo* formaci proteinů a ověřila účinnost predikcí agregace a sekundárních struktur v kontextu nepřirozených aminokyselinových sekvencí.

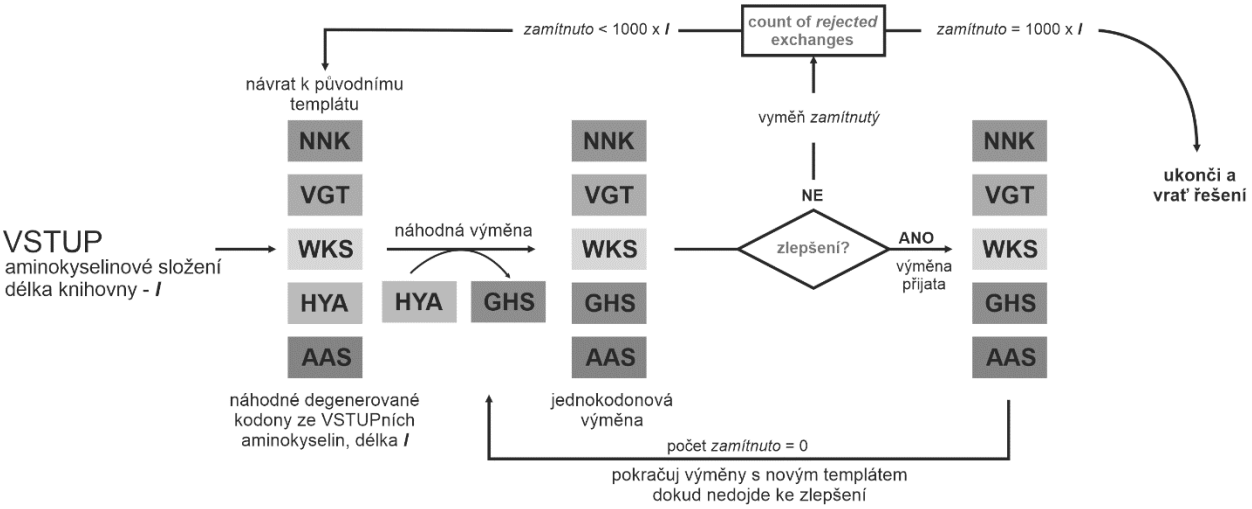
Vývoj nástroje pro návrh kombinatoriálních proteinových knihoven (CoLiDe)

Po selektivní charakterizaci náhodného sekvenčního prostoru jsme se rozhodli podniknout kroky k charakterizaci kolektivních vlastnosti proteinových knihoven. Bohužel, současné nástroje pro návrh kombinatoriálních knihoven nenabízí optimální prostředky pro návrh náhodných proteinů. Tyto algoritmy umožňují efektivní přípravu malých cílených knihoven v proteinovém inženýrství zatímco naším cílem byla knihovna o co nejvyšší variabilitě s každou proteinovou sekvencí omezenou pouze jejím aminokyselinovým složením. Za tímto účelem jsme implementovali nástroj CoLiDe pro návrh kombinatoriálních proteinových knihoven, nástroj je optimalizován pro efektivní a uživatelsky přístupný návrh experimentálních knihoven náhodných sekvencí.

Účelem nástroje CoLiDe je nalézt takový soubor degenerovaných kodonů, který při kombinaci do jednoho kódujícího DNA templátu a jeho následné translaci poskytne proteinovou knihovnu o daném aminokyselinovém složení. Vstupy algoritmu jsou délka požadované knihovny, její aminokyselinové složení, stupeň variability (maximální počet aminokyselin kódovaných jedním kodonem) a organismální kodonové preference. Program navíc umožňuje vyřadit specifický nedegenerovaný kodon ze zařazení do knihovny nebo překódovat některé kodony ke specifickým, uživatelem definovaným aminokyselinám, které budou následně vloženy do kombinatoriální knihovny. Hlavním výstupem CoLiDe je řetězec degenerovaných kodonů jež kóduje kombinatoriální knihovnu proteinů s definovaným obsahem aminokyselin.

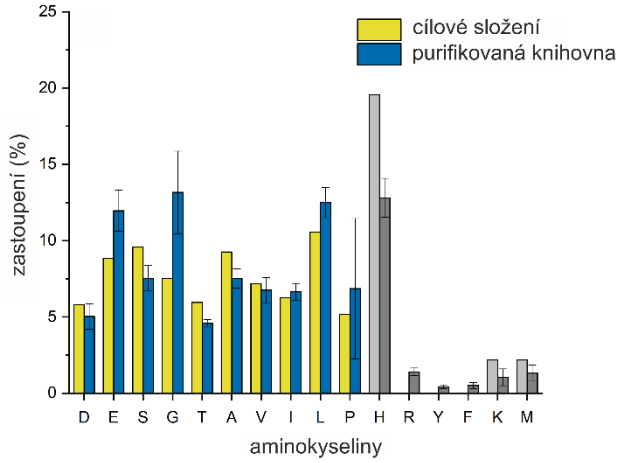
Principem CoLiDe je zjednodušená verze evolučního algoritmu – jedinci jsou selektováni prostřednictvím náhodných mutací, přičemž se současně optimalizuje pouze jeden jedinec a ne celá populace jak je tomu zvykem u evolučních algoritmů. Po načtení vstupních parametrů jsou ze souboru všech dostupných 3375 degenerovaných kodonů odstraněny ty, jež nekódují aminokyseliny ze zadání. Následovně je z tohoto souboru vybrán náhodný set degenerovaných kodonů o délce zadané proteinové knihovny a je z něj vypočtena odchylka aminokyselinového složení od zadané distribuce. Odchylka je spočítána jako součet čtverců odchylek pro každou aminokyselinu. Poté je jeden degenerovaný kodon v počátečním setu vyměněn za náhodně vybraný kodon z celkového souboru kodonů a odchylka je přepočtena. Pokud je výsledná distribuce touto výměnou odchýlena od zadání více než před substitucí, výměna je zamítnuta. V opačném případě je přijata. Tento cyklus výměna-přepočet-rozhodnutí je opakován dokud není dosaženo $1000 \times l$ konsekutivních zamítnutí. Výsledný řetězec kodonů, ve kterém žádná další

náhodná výměna nepřináší lepší aproximaci zadání, je vrácena jako řešení. Výpočetní postup je ilustrován na **Obr. 4**.



Obrázek 4. Schématická reprezentace výpočetního algoritmu CoLiDe. Vstupem je úživatelem definovaná délka degenerované knihovny a její aminokyselinové složení. Program nejdříve odstraní ty degenerované kodony jež nekódují zadané aminokyseliny. Poté je vygenerován náhodný řetězec degenerovaných kodonů o délce knihovny a kodony jsou vyměňovány dokud není dosaženo požadované aminokyselinové distribuce

Efektivita algoritmu byla otestována na 45 aminokyselinové knihovně s 33 aminokyselinovou variabilní části. Knihovna byla připravena z jednovláknového degenerovaného DNA templátu jež byl konvertován na dvouvláknouvou DNA pomocí Klenowovy reakce a translatován v bezbuňěčném expresním systému. Proteinová knihovna byla purifikována pomocí afinitní chromatografie specifické pro poly histidinovou kotvu, její hmotnostní distribuce byla evaluována MALDI-TOF hmotnostní spektrometrií a její aminokyselinové složení pomocí aminokyselinové analýzy. Výsledná směs proteinů vykázala dobrou shodu ve své hmotnostní a obsahové distribuci s teoretickými hodnotami (**Obr. 5**).



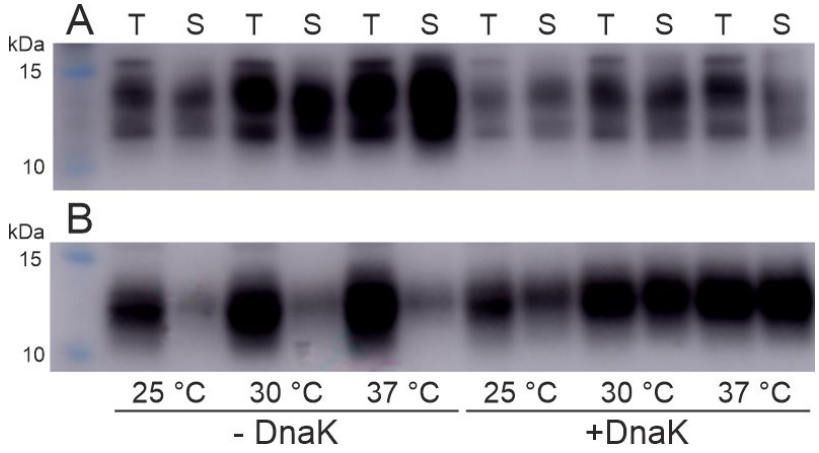
Obrázek 5. Aminokyselinová analýza purifikované knihovny a srovnání jejího složení se zadanou distribucí

Ve výsledku jsme implementovali a experimentálně ověřili výpočetní nástroj pro návrh komplexních degenerovaných proteinových knihoven a učinili tento nástroj přístupným široké vědecké veřejnosti.

Charakterizace kombinatoriálních proteinových knihoven se specifickými aminokyselinovými poměry

Selektivní charakterizace 45 náhodných proteinů zprostředkovala detailní popis jednotlivých náhodných proteinů. Nicméně k odvození obecných charakteristik náhodného sekvenčního prostoru a rovněž vlivu aminokyselinového repertoáru na strukturu proteinů je zapotřebí hromadné charakterizace statisticky významného vzorku sekvencí. V této práci jsme využili nástroje CoLiDe k návrhu dvou knihoven s 20 (knihovna 20F) a 10 (knihovna 10E; A, S, D, E, G, T, I, L, P a V) aminokyselinovým repertoárem a prostudovali efekty chaperonů DnaK, DnaJ a GrpE na náhodné proteiny. Zaměřili jsme se na rozpustnost, agregační tendence a senzitivitu proteinových knihoven vůči proteolýze dvěmi odlišnými proteázami.

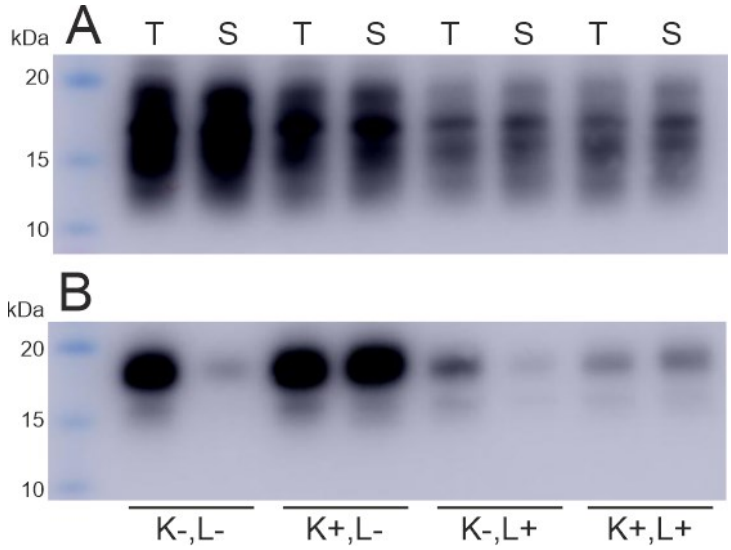
Byly navrženy dvě proteinové knihovny o délce 106 aminokyselin s 85 aminokyselinovou variabilní části, štěpným místem pro proteázu trombin uprostřed proteinových sekvencí a dvěmi afinitními kotvami pro afinitní purifikaci a chemiluminscenční detekci. Knihovny byly syntetizovány v podobě dvou překrývajících se jednořetězcových DNA oligonukleotidů, převedeny na dvouřetězcový DNA templát a proteiny byly připraveny pomocí bezbuněčného expresního systému. Otestovali jsme vliv DnaK chaperonového systemu v kontextu modulací proteinové rozpustnosti a potenciálně formace strukturovaných molekul. Dodatečně jsme využili selektivní proteolýzy k estimaci celkové strukturalizace knihovny pomocí proteáz Lon a trombin. Předběžná charakterizace solubilizace a proteolytické experimenty proteinových knihoven naznačují odlišné chaperon/proteinové interakce v knihovnách 20F a 10E (**Obr. 6, Obr. 7**). Knihovna 20F je v přítomnosti chaperonů významně solubilizovaná ve všech testovaných teplotách 25, 30 a 37 °C a zároveň prokázala vyšší rezistenci vůči proteolýze ve srovnání s



Obrázek 6. Western blot analýza solubilit proteinových knihoven 10E (A) a 20F (B) v přítomnosti (+DnaK) a absenci (-DnaK) chaperonů při teplotách 25, 30 a 37 °C. Analyzována byla celková frakce exprimované knihovny (T) a její rozpustná frakce (S)

knihovnou exprimovanou v nepřítomnosti chaperonů, jejíž rozpustná frakce byla účinně štěpena protezou Lon specifickou pro neuspořádané proteiny (**Obr. 6B, Obr. 7B**).

Toto pozorování naznačuje, že v nepřítomnosti chaperonů se většina proteinů 20F nachází v neuspořádané konformaci. Suplementace 20F knihovny chaperonem DnaK solubilizuje většinu původně nerozpustných proteinů, které jsou však posléze také aktivně degradovány Lon proteázou (**Obr. 7B**). Nicméně chaperony poskytují významnou ochranu před degradací což je zřejmé z porovnání poměru rozpustných frakci v chaperony suplementované a nesuplementované reakci v přítomnosti Lon proteázy (**Obr. 7B**). Na druhou stranu, knihovna 10E je vysoce rozpustná i bez chaperonové asistence (**Obr. 6A, Obr. 7A**). Navíc, chaperony neposkytují žádnou ochranu proti Lon dependentní degradaci jak je zřejmé ze srovnání reakcí s a bez chaperonů v přítomnosti Lon proteázy (**Obr. 7A**). Stojí za zmínku, že suplementace chaperony dokonce snižuje celkovou míru exprese knihovny 10E (**Obr. 6A, Obr. 7A**). Tento počáteční screening naznačuje, že přestože kompaktní struktury jsou přítomny v obou knihovnách 20F i 10E, nižší míra proteolýzy proteinů knihovny 10E může být vysvětlená vyšší frekvencí formace nativních struktur.



Obrázek 7. Western blot analýza solubilit a štěpení knihoven 10E (A) a 20F (B) proteázou Lon v přítomnosti (K+) a absenci (K-) chaperonů. Přítomnos Lon proteázy je vyznačena L+, její absence L-. Analyzována byla celková frakce exprimované knihovny (T) a její rozpustná frakce (S)

Tyto předběžné výsledky budou ověřeny pomocí kvantitativní analýzy proteolytického štěpení z PVDF membrán a analytické gelové chromatografie.

Charakterizace variant defosfokoenzym A kinázy (DPCK) bez aromatických aminokyselin

Aminokyselinám obsahujicím aromatický kruh je přisuzován největší vliv na stabilizaci proteinové struktury a zároveň jsou považovány za nejnovější přírůstky do aminokyselinového repertoáru proteinů. Nicméně, prvotní proteiny pravděpodobně disponovaly strukturou a funkcí i v nepřítomnosti těchto reziduí. V této práci jsme ověřili hypotézu neesenciality aromatických aminokyselin na příkladu varianty enzymu defosfo koenzym A kinázy (DPCK) neobsahující aminokyseliny s aromatickým kruhem. Žádná z aromatických aminokyselin není vyžadováná k enzymatické aktivitě DPCK, proto je tento protein ideálním kandidátem pro studium vztahu aminokyselinového repertoáru na proteinovou strukturu.

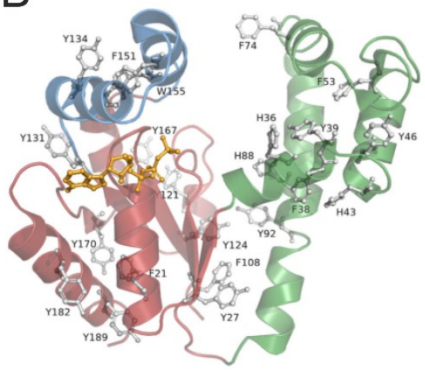
Podařilo se nám úspěšně připravit DPCK z hypertermofilní bakterie *Aquifex aeolicus* (DPCK-WT) a jeho dvě mutantní varianty buď se všemi aromatickými aminokyselinami s výjimkou histidinu substituovanými leucinem (DPCK-LH) nebo se všemi aromatickými kyselinami včetně histidinu substituovanými racionálně navrženými prebioticky dostupnými hydrofobními aminokyselinami (DPCK-M) (**Obr. 8A**).

A

Pozice	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

Pozice	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

B

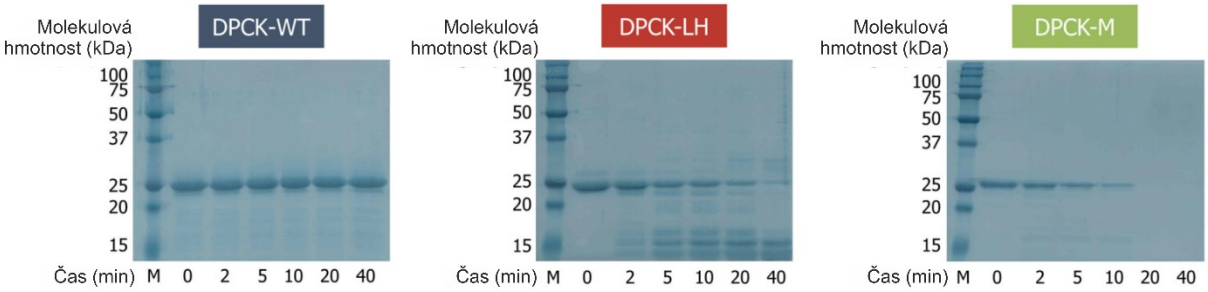


Obrázek 8. Souhrn substituovaných aromatických reziduí ve variantách DPCK-LH a DPCK-M v porovnání s přírodní verzí DPCK-WT (A) a (B) struktura DPCK enzymu s vyznačenými aromatickými aminokyselinami (šedě) a navázáným ATP (žlutě)

Všechny proteiny byly připraveny v buňkách *E. coli* BL21-DE3, purifikovány pomocí třístupňového purifikačního protokolu a otestovány na přítomnost katalytické aktivity. Přírodní forma DPCK-WT měla podobnou aktivitu jako dříve charakterizována DPCK z *E. histolytica*. Zatímco DPCK-WT a -LH varianty vykázaly ATP hydrolytickou i dCoA-dependentní fosfotransferázovou aktivitu, varianta DPCK-M disponovala pouze ATPázovou aktivitou. Při kvantifikaci produktu katalýzy varianta DPCK-LH vyprodukovala 100× méně CoA než přírodní varianta DPCK-WT.

Strukturní charakterizace všech variant DPCK pomocí spektroskopie cirkulárního dichroismu a 1D/2D HN NMR spektroskopie odhalila vyšší obsah neuspořádaných struktur v obou mutovaných

variantách ve srovnání s DPCK-WT. Nicméně, na rozdíl od varianty DPCK-M disperze NMR signálů DPCK-LH varianty naznačuje částečnou formaci terciárního uspořádání. Komplementárně, dynamika proteolýzy všech variant DPCK proteázou LysC ukázala na rozdílné terciární uspořádání proteinů. Zatímco varianta DPCK-WT proteolýze odolávala, varianta DPCK-M je kompletně degradována v čase štěpného experimentu a varianta DPCK-LH je štěpená za vzniku velkých 15 kDa fragmentů (**Obr. 9**).



Obrázek 9. 14% SDS-polyakrylamidové gely znázorňující průběh selektivní proteolýzy DPCK proteinů pomocí proteázy LysC. Gel byl vizualizován imidazol-zinkovou metodou barvení

Na závěr této studie jsme se soutředili na konformační dynamiku varianty DPCK-LH související s vazbou ATP pomoci spektroskopie nukleární magnetické resonance, dynamického rozptylu světla a titrací 8-aminonaftalen-1-sulfonové kyseliny (ANS). Pomocí těchto technik jsme potvrdili relaxovanou strukturu (angl. molten globule) DPCK-LH a výraznou změnu v konformaci při vazbě ATP. Podle měření dynamického rozptylu světla této varianty se hydrodynamický poloměr zmenšil o 20 % a dosáhl tak hodnoty odpovídající struktuře DPCK-WT (**Tabulka 1**). Tato pozorování podporují hypotézu dle níž kofaktory mohly hrát nezaměnitelnou roli při stabilizaci raných proteinových struktur.

Tabulka 1. Souhrn měření dynamického rozptylu světla variant DPCK-WT a DPCK-LH v přítomnosti a absenci ATP

	Střední hydrodynamický poloměr (nm)	Index polydisperzity (%)
DPCK-WT	2.52 ± 0.15	6.8
DPCK-WT + 200 μM ATP	2.44 ± 0.17	8.3
DPCK-LH	3.30± 0.15	8.0
DPCK-LH + 200 μM ATP	2.68± 0.18	18.2

Souhrn

Cílem této práce bylo (i) prostudovat vlastnosti náhodného sekvenčního prostoru a vztah náhodných proteinových sekvencí k přírodním proteinům a (ii) studovat vliv aminokyselinového repertoáru na proteinovou strukturu a funkci.

Byly získány níže uvedené výsledky jež jsou součástmi tří přiložených publikací a popsaných předběžných výsledků které budou rozvinuty v další studii:

- Výpočetní analýza náhodné proteinové knihovny ukázala její podobnost v obsahu sekundárních struktur a zároveň odhalila rozdíly v agregačních tendencích ve srovnání s přírodními proteiny.

- Experimentální charakterizace 45 náhodných proteinových sekvencí potvrdila *in silico* predikce obsahu sekundárních struktur a agregace a odhalila, že neuspořádáné náhodné sekvence jsou nejvíce tolerovanými proteiny v intracelulárním prostředí.

- Byl vyvinut nástroj CoLiDe pro návrh kombinatoriálních proteinových knihoven a byl poskytnut širokému vědeckému publiku.

- Nástroj CoLiDe byl otestován experimentálně, testování odhalilo nástrahy jež vedou k odchylkám daným experimentální přípravou knihovny.

- Byly připraveny kombinatoriální proteinové knihovny s různými aminokyselinovými repertoáry. Výsledky jejich biochemické charakterizace naznačují odlišné strukturní tendence v náhodném sekvenčním prostoru.

- Charakterizace variant enzymu defosfo koenzym A kinázy bez aromatických aminokyselin ukázala na důležitou roli těchto reziduí při stabilizaci nativní proteinové struktury.

- Redukce hydrodynamického poloměru varianty DPCK bez aromatických aminokyselin při vazbě kofaktoru poukazuje na významnost kofaktorů při stabilizaci raných proteinových struktur.

Bibliografie

1. 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).

2. 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).

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

4. 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).

5. 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).

6. 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).

7. 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).

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

9. Caetano-Anollés, G. & Caetano-Anollés, D. An evolutionarily structured universe of protein architecture. *Genome Res.* **13**, 1563–1571 (2003).

10. 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).

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

12. 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).

13. 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).

14. 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.

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

16. 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).

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

18. Longo, L. M., Despotovi, D., Weil-ktorza, O., Walker, M. 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. **117**, (2020).

19. 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).

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

21. 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).

22. 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).

23. 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 Suppl 6**, S22 (2009).

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

25. 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).

26. 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).

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

28. 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).

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

30. 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).

31. 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).

32. 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).

33. 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).

34. 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).

35. 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).

Seznam publikací

Publikace přímo související s touto prací:

4) **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)
5) **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)
6) 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)

Další publikace autora:

4) 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)
5) 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)
6) 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)

Curriculum Vitae

Education

2015 – now	PhD in **Biochemistry** at Faculty of Science of Charles University **PhD thesis**: The effect of amino acid repertoire on protein structure and function **Supervisor**: Klára Hlouchová PhD
2014 – 2018	MSc in **Chemical Informatics and Bioinformatics** at the University of Chemical Technology **Master's thesis**: Aplication of random sequences in protein engineering **Supervisor**: Daniel Svozil, PhD
2013 – 2015	MSc in **Biochemistry** at Faculty of Science of Charles University **Principal courses**: Biochemistry, Molecular Biology, Analytical Biochemistry, Enzymology, Biophysical Chemistry, Structural Biology, Proteomics, Genetic Engineering **Master's thesis**: Never Born Proteins: Occurrence and characterization of secondary structure motifs **Supervisor**: Klára Hlouchová PhD
2010 – 2013	BSc in **Biochemistry** at Faculty of Science of Charles University **Bachelor's thesis**: Expression of recombinant form of nepenthesin I from *Nepenthes gracilis*. **Supervisor**: Petr Man PhD

Courses, fellowships and collaborations

February, 2019	Creating is understanding: Synthetic biology masters complexity, Heidelberg, Germany **EMBO Workshop, travel fellowship**
February, 2019	Methods for studying phase separation in biology, Dresden, Germany **EMBO practical course, travel fellowship**
June, 2018	Astrobiology Graduate Conference travel fellowship. Atlanta, USA
March, 2018	Charles University graduate student research fellowship (GAUK) **funded project (1686218)** - "Evolution of protein structure by interaction with RNA"
January, 2018	6th ELSI International Symposium **travel fellowship**. Tokyo, Japan
September, 2017	Spetsai Summer School 2017: Proteins and organized complexity **EMBO/FEBS event.** Spetses, Greece
Jan., 2016 – March, 2016	Virtual screening for allosteric ligands of HIV-1 protease, research fellowship EEA and Norway grants, **funded project**: NF-CZ07-INS-5-172-2015 research stay at laboratory of Medical Pharmacology and Toxicology, Arctic University of Tromso, Norway

Talks and poster presentations

February, 2019	Creating is understanding: Synthetic biology masters complexity, Heidelberg, Germany flash talk & poster presentation
July, 2018	Gordon Research Conference: Intrinsically Disordered Proteins 2018, Les Diablerets, Switzerland poster presentation – Into the wild: Expression and characterization of random protein libraries
June, 2018	Astrobiology Graduate Conference, Atlanta, GA, USA poster presentation – Search into unevolved protein space
April, 2018	Prague Protein Spring Conference, Prague, Czech Republic poster presentation – Into the wild: Expression and characterization of random protein libraries
March, 2018	XV Discussions in Structural Molecular Biology, Nove Hrady, Czech Rep. poster presentation – Into the wild: Expression and characterization of random protein libraries, **best poster award**
January, 2018	6th ELSI International Symposium, Tokyo, Japan poster presentation - Search into unevolved protein space
April, 2017	Astrobiology Science Conference (AbSciCon) 2017, Phoenix, Arizona, USA poster presentation – Search into unevolved protein space
January, 2017	Diamond Light Source, Oxfordshire, Great Britain invited talk - How does the amino acid repertoire affect the protein structure universe?
November, 2016	Students Scientific Conference (SVK) at University of Chemical Technology, Prague, Czech Republic conference talk – Dark protein space: from theory to high throughput experiment, **best talk award**
September,2016	CSBMB Conference, Prague, Czech Republic poster presentation - The Effect of Genetic Code Evolution on Protein Structure Space
September, 2016	Origins and evolution of life on Earth and the Universe, Liblice, Czech Rep. poster presentation - The Effect of Genetic Code Evolution on Protein Structure Space
June, 2016	ENBIK2016 Conference, Loučeň, Czech Republic

	conference talk – Occurrence of secondary structure in protein sequence space
May, 2016	Prague Protein Spring Conference, Prague, Czech Republic poster presentation – Occurrence of secondary structure in the vast protein sequence space
December, 2015	Students Scientific Conference (SVK) at University of Chemical Technology, Prague, Czech Republic presentation of master's thesis – Occurrence of structure in random protein sequences, **2nd best talk award**

Research papers

(1) Makarov, M., Meng, J., **Tretyachenko, V**., Srb, P., Brezinova, A., Giacobelli, V. G., ... & Hlouchova, K. (2020). Enzyme catalysis prior to aromatic residues: reverse engineering of a dephosphoCoA kinase. *bioRxiv*.

(2) **Tretyachenko, V**., Voráček, V., Souček, R., Fujishima, K., & Hlouchová, K. (2020). CoLiDe: Combinatorial Library Design tool for probing protein sequence space. *Bioinformatics*.

(3) **Tretyachenko, V.**,Vymětal, J., …& Hlouchová, K., Random protein sequences can form defined secondary structures and are well-tolerated in vivo., *Sci. Rep*.. 7:15449.

(4) Makukhin, N., **Tretyachenko, V.**, Moskowitz, J., Misek, J. A Ratiometric Fluorescent Probe for Imaging of the Activity of Methionine Sulfoxide Reductase A in Cells. *Angew Chem Int Ed Engl.,* September 2016

(5) Fejfarová, K., Kádek, A., Mrázek, H., Hausner, J., **Tretyachenko, V**., Koval, T., ... & Dohnálek, J. (2016). Crystallization of nepenthesin I using a low-pH crystallization screen. *Acta Crystallographica Section F: Structural Biology Communications*, 72(1), 24-28.

(6) Kadek, A., **Tretyachenko**, V., Mrazek, H., Ivanova, L., Halada, P., Rey, M., Schriemer, D.C., Man, P. Expression and characterization of plant aspartic protease nepenthesin-1 from Nepenthes gracilis. *Protein Expr Purif*,Vol. 95, March 2014, p. 121–128