Charles University

Faculty of Science

Study programme: Molecular Biology and Biochemistry of organisms

Branch of study: Bachelor

Melikov Aleksandr

Functional screening of *de novo* proteins

Charakterizace funkcí *de novo* proteinů

Type of thesis:

Bachelor’s thesis

Supervisor:

Mgr. Klára Hlouchová, Ph.D.

Prague, 2022

Poděkování:

Chtěl bych moc poděkovat své školitelce Kláře Hlouchové a svému konzultantovi Valerio Guido Giacobelli za jejich trpělivost a neocenitelnou pomoc v napsání dané bakalářské práce. Také bych chtěl poděkovat své rodině a přítelkyni za duševní podporu v této době.

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 04.05.2022

Abstract

Synthetic biology relies upon working with two main types of biological macromolecules - nucleic acids and proteins. Natural proteins represent only a small percentage of the whole amino-acid sequence space. Most of it may conceal an enormous potential (unexplored by nature as well as scientific endeavor), which has started to be carefully explored only in the recent decades. Characterization of non-native proteins includes several key aspects: structure and its stability, function, patterns of interaction with other molecules (of different chemical nature) and *in vivo* tolerance. This work focuses on the functional testing of *de novo* polypeptide molecules, either appearing as novelties of genome non-coding regions or as products of artificial bioengineering design.

Key words: *de novo* proteins, function screening, protein libraries, protein design, sequence space

Abstrakt

Syntetická biologie obnáší práci zejména se dvěma hlavními typy biologických makromolekul - nukleovými kyselinami a proteiny. Přírodní proteiny představují zanedbatelnou část celého sekvenčního prostoru. Většina tohoto prostoru je dosud neprozkoumaná (jak biologickými systémy tak i vědeckým úsilím) a může skýtat nepoznaný strukturní a funkční potenciál, kterým se syntetická biologie a proteinové inženýrství zabývá zhruba jen poslední dvě desetiletí. Charakterizace nepřirozených proteinů zahrnuje několik klíčových aspektů: strukturu a její stabilitu, funkci, možnosti interakce s jinými proteiny, nukleovými kyselinami či kofaktory a v neposlední řadě *in vivo* tolerance. Tato práce se zaměřuje prioritně na funkční testování *de novo* proteinů, vzešlých buď z dříve nekodujících genomových oblastí, nebo jako produkty proteinového inženýrství.

Klíčová slova: *de novo* proteiny, funkční screening, proteinové knihovny, design proteinů, sekvenční prostor

Table of contents

1. Properties of protein sequence space

From the beginning of protein biochemistry, it was clear that combinatorial possibilities of the amino acid alphabet are tremendous. Since the genetic code was deciphered and the first protein was sequenced, new opportunities for bioengineering appeared (*Sanger and Tuppy 1951; reviewed by Nirenberg 2004*). There is one frequently referenced example of combinatorial diversity even for relatively short proteins – for a 100 amino acid polypeptide and 20 canonical amino acids, 20^{100} possible sequences could be constructed. Such numbers are beyond the possibilities of what nature can experiment with during evolution and even more inconceivable for larger proteins. Still, huge amounts of genomic and proteomic data were processed and analyzed during the last 50 years in order to systematize and classify all sequences found in nature (estimated to be in the order of 10^{15} in total). Many patterns of protein evolution have been discovered, providing the first insights about ancestral sequences and functions (*Goodman 1981; reviewed by Pál et al 2006; Kolodny et al. 2021*).

Why and how nature selected the specific proteins to sustain life is not clear. In general it is considered that non-efficient, deleterious sequences are rapidly erased from genomes. The longer the biological life exists, the more variants are eliminated from the world gene pool. But how much of the whole amino acid sequence space has already undergone natural selection or just disappeared due to genetic drift? Which fraction of possibilities was expelled? On the basis of existing data, Dryden and colleagues calculated the size of sequence space as a function of the amino acid alphabet size used in the protein (**Fig. 1**)**.** It illustrates that a limited repertoire of amino acids severely reduces the number of possible sequence variants as does the sequence length (*Dryden et al. 2008*). Most amino acids in protein structures are not obligatory essential, they could be substituted by analogues with similar side chain properties and the general structure could still be maintained *(Lau and Dill 1990; reviewed by Cordes et al. 1996*). Some proteins have been shown to obtain new functions by changing only a few most conserved and therefore important amino acids, retaining the original structural fold (*Nagano et al. 2002; reviewed by Anantharaman et al. 2003*). These observations have been considered by researchers who were exploring functional features of partially designed protein libraries based on specific folds, e.g., 4-alpha-helix bundle protein

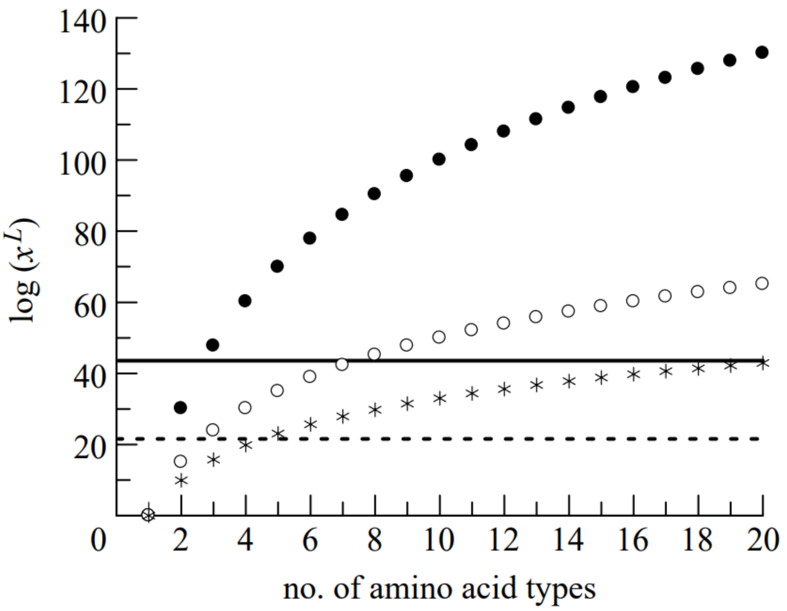


Fig.1. A graph representing the number of possible sequences log(x^L) as the function of the number of amino acid types for proteins of different length: 33 amino acids (asterisks), 50 (empty circles) and 100 (filled circles). The solid and dashed horizontal lines define the estimated amount of explored sequence variants since the origin of life on Earth. Taken from *Dryden et al. 2006.*

libraries, where no specific function was designed (*Fisher et al. 2011*). Although some studies constrain the size of protein sequence space, such an approach still remains largely unexplored and will probably help us comprehend the vast possibilities of the unexplored protein space.

1.1. Non-coding genome

Exploring sequence space isn't focused only on the research of translated open reading frames (ORFs) - non-coding genome regions also contain a lot of essential biological actors. Cis-regulatory elements, such as promoters, terminators, enhancers and inhibitors, different kinds of non-translated RNA and other components of non-coding regions. Although such elements do not provide any peptide product, they play an important role in cell existence (*reviewed by Ludwig 2002; reviewed by Shabalina and Spiridonov 2004*). Nevertheless, it turns out that a substantial fraction of non-coding DNA gets transcribed, producing plenty of different RNAs, which furthermore could be translated (*Ruiz-Orera et al. 2018*). Small ORFs may play an important role in the cell's metabolism in stressful conditions, as it was shown in e.g. *Escherichia coli* (*Hemm et al. 2010*) and many of them are located across intergenic regions, probably getting expressed due to environmental stress. Many other small translated ORFs were identified in eukaryotes as well (*Bazzini et al. 2014*). These *de novo* born peptides could serve as possible templates for novel gene selection (*reviewed by Ruiz-Orera et al. 2020*). Furthermore, the studies investigating potential of non-coding sequences indicate the probability of new ORFs evolving from them (*Wu and Zhang 2013; reviewed by Tautz 2014*). Such native *de novo* proteins are of great interest particularly for their hypothetical new function and evolutionary role.

1.2. *De novo* gene birth

During billions of years, evolution created and tuned huge amounts of genomic sequences to carry out specific functions in given organisms. Processes responsible for that are ongoing, with new functions evolving constantly by several different mechanisms, ranging from evolution based on preexisting promiscuous activities to gene recombination (*reviewed by Copley 2020*). Getting a stable and functional protein structure from intergenic non-coding sequences may seem to be implausible compared with the better characterized and more prevalent phenomena of gene duplication or horizontal gene transfer (*reviewed by Long et al. 2003*). Statistically, *de novo* gene birth is not very common: e.g., for *Drosophila* the frequency has been estimated about 1 event per million years (*Heames et al. 2020*), for three-spined stickleback - around 80 events per million years (*Schmitz et al. 2020*), for rice the reported number is roughly 50 per million years (*Zhang et al. 2019*). Despite that, the novel sequences may have a substantial effect on evolutionary and physiological characteristics of an organism. At the same time, totally new biological activities potentially coded by such sequences can affect the whole cell metabolism. But how can the rate of function gain be assessed and which features of newborn coding DNA may be considered as functional?

To help resolve this question, the Pittsburgh model was constructed by Keeling et al. It defines the main interpretations of what is regarded as the function in terms of *de novo* protein birth (*Keeling et al. 2019*). Based on a comprehensive meta-analysis, 5 "meanings of function" were distinguished (**Fig. 2**)**:** evolutionary role, physiological function, patterns of interaction, intrinsic biophysical capacity (including structural flexibility) and expression profile. Besides, these aspects of function can be considered hierarchically: in order to obtain a physiological role, the sequence must positively affect the natural selection and provide a beneficial phenotype and so on. According to the authors, a gradual acquisition of the discussed functional properties is essential for a novel ORF to be transformed into a true *de novo* gene. In addition, Wu and Zhang proposed a general model for *de novo* gene function origin**,** which reflects an order of events starting from transcription and leading to product adaptation (*Wu and Zhang 2013,* **Fig. 3a**). This "adaptation following neutrality" model gives an insight into how rapid non-coding sequences could be evolved under the positive selection. Newly transcribed and translated gene products may have a narrow interacting pattern range at the beginning, providing a starting platform for subsequent modifications. Structure flexibility and starting set of interactions are the two main requirements for rapid evolvability of function for such *de novo* proteins. Under certain conditions, novel genes may acquire a significant role, which would be quickly fixed due to adaptive evolution and increasing the fitness of a given organism. This scenario was demonstrated in *Drosophila* (*Chen et al. 2010*), although it isn't the sole one. New function acquisition can happen in a different way *via* several stages of intermediate proto-genes, so the continuum between non-coding sequences and novel genes is observable (*Carvunis et al. 2012,* **Fig. 3b**).

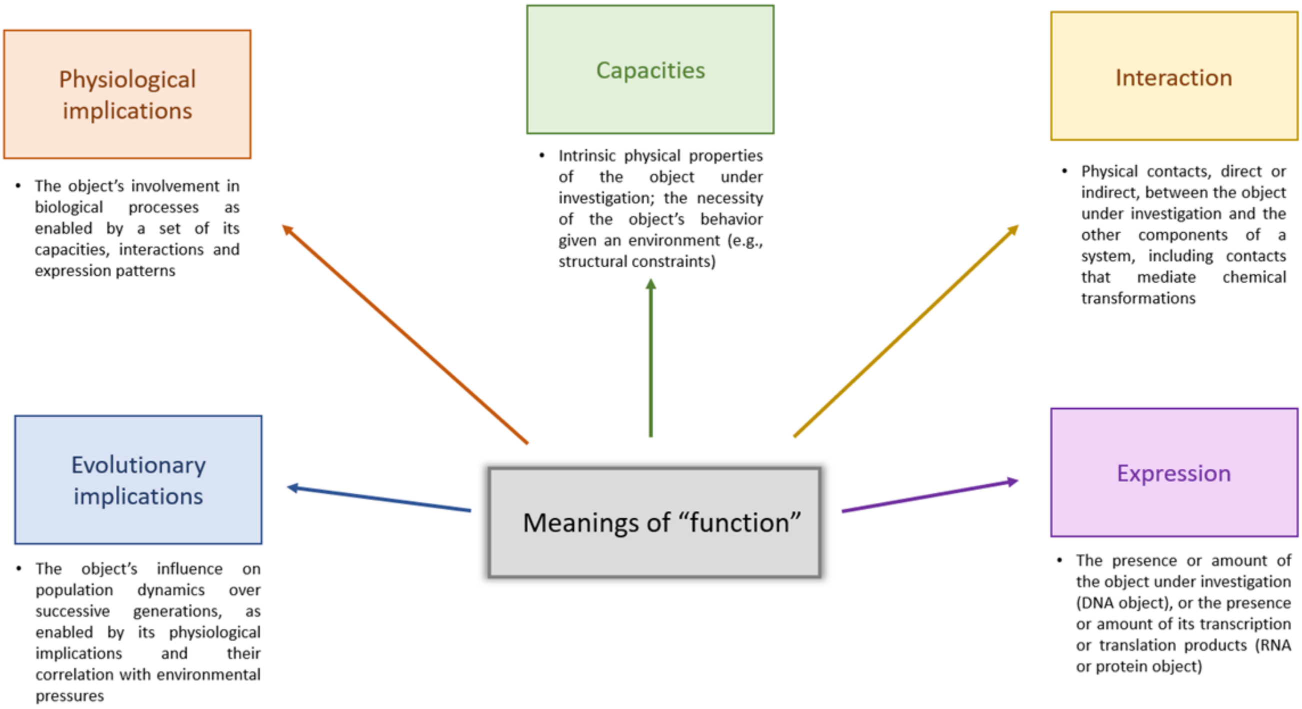

Physiological implications
• The object's involvement in biological processes as enabled by a set of its capacities, interactions and expression patterns
Capacities
• Intrinsic physical properties of the object under investigation; the necessity of the object's behavior given an environment (e.g., structural constraints)
Interaction
• Physical contacts, direct or indirect, between the object under investigation and the other components of a system, including contacts that mediate chemical transformations
Evolutionary implications
• The object's influence on population dynamics over successive generations, as enabled by its physiological implications and their correlation with environmental pressures
Meanings of "function"
Expression
• The presence or amount of the object under investigation (DNA object), or the presence or amount of its transcription or translation products (RNA or protein object)

Fig.2. The Pittsburgh model of function. 5 key aspects that can guide in seeking and characterizing a novel gene. Adapted from *Keeling et al. 2019***.**

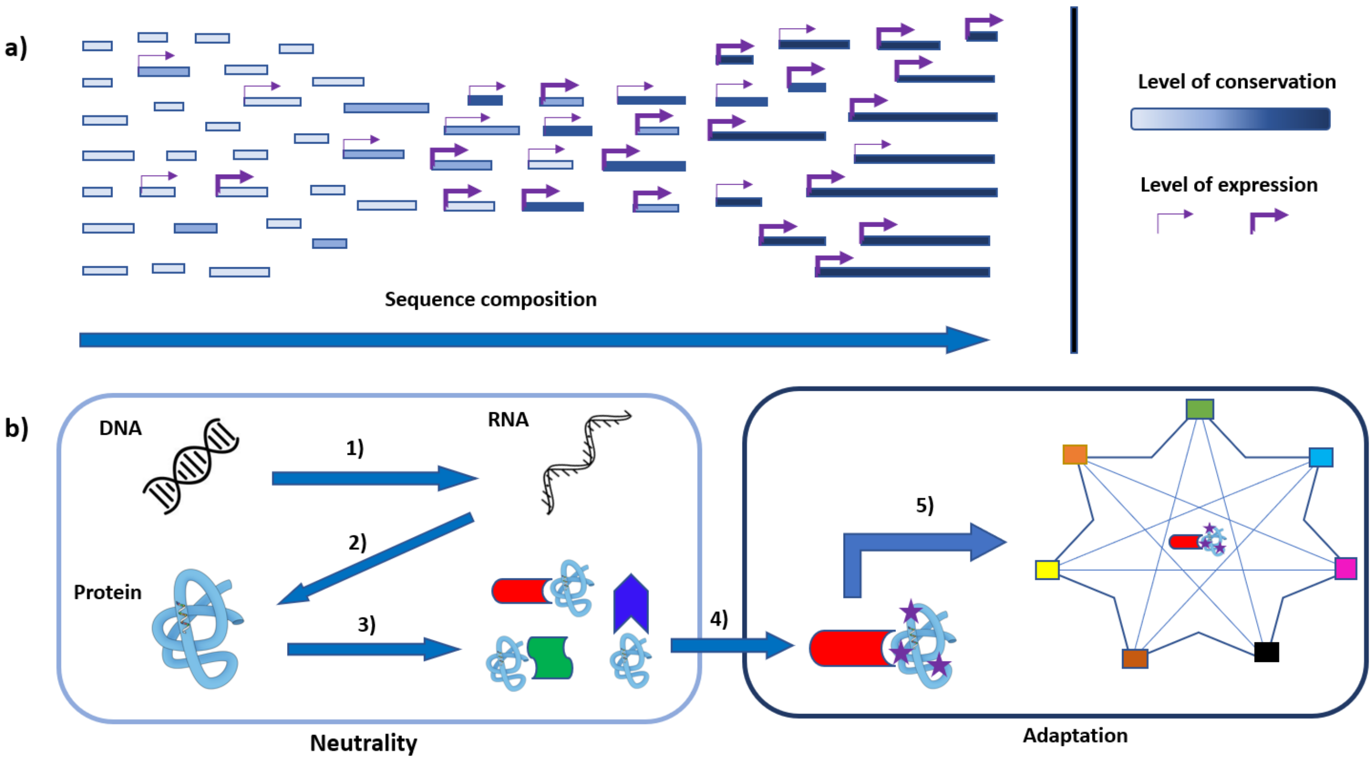


Fig.3. Models of *de novo* gene emergence. **a)** The Continuum model. ORFs from non-coding regions are expressed, but only a certain fraction of them is retained. Persisted proto-genes will gradually evolve and diversify into specific independent protein-coding units under the positive selection and adaptive evolution**.** Adapted from *Carvunis et al. 2012*). **b)** Adaptation following neutrality model. (1) Non-coding regions of DNA get transcribed and (2) translated, giving a protein product. Due to conformational variability, novel polypeptides possess a range of interacting patterns with other native proteins or other substrates (3). Under positive selection and adaptive evolution, the specificity of interactions increases, which can lead a significant function to emerge (4,5). Adapted from *Wu and Zhang 2013*).

Studying the processes of *de novo* gene emergence on the *Oryza* plant model, Zhang proposed the following mechanism (**Fig. 4**): for a sequence to become a gene, a series of events must occur, including transcription gain due to a frameshift (indel, i.e., insertion-deletion event) and removing of premature stop codons. Afterwards, transcribed sequence is able to be translated providing a polypeptide product, which consequently undergoes the natural selection process. These 2 events, the transcription and translation gain, do not necessarily occur simultaneously or within a short period of time: an intergenic sequence may firstly evolve into a functional non-coding RNA, which consequently becomes a translated ORF. Remarkably, both RNA and protein products of such sequence can concurrently carry out distinctive functions (*Dinger et* *al. 2008*). These bifunctional RNAs can be considered as intermediates between non-coding RNA and completely evolved novel gene.

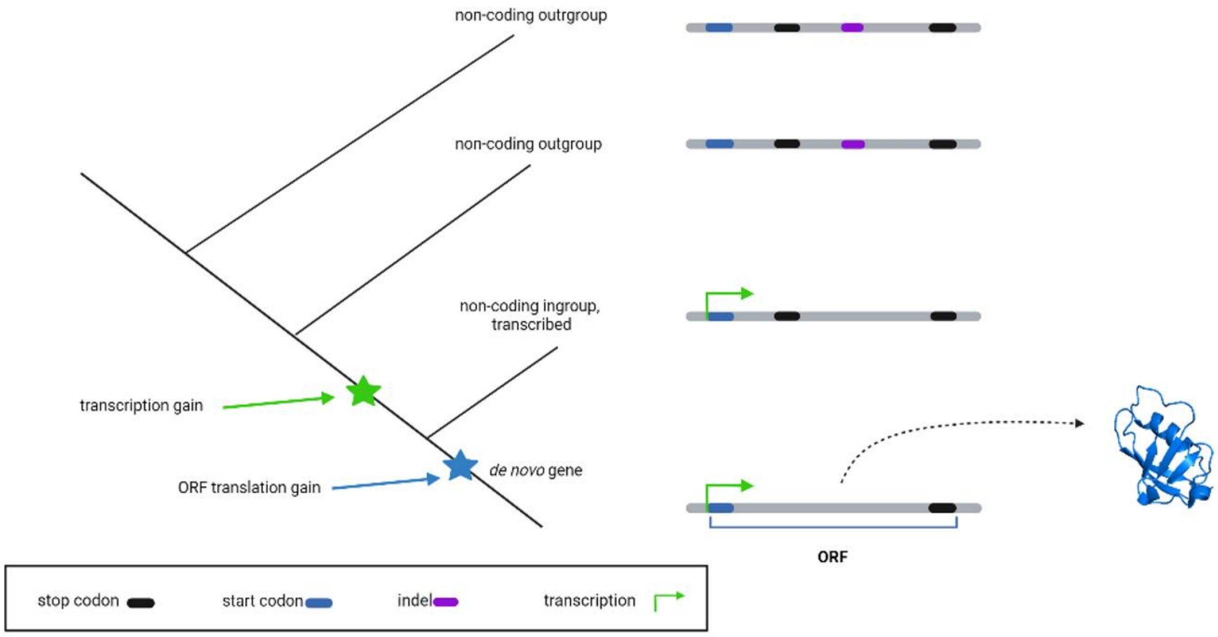


Fig.4. A possible model of de novo gene emergence. Due to the insertion-deletion event, which causes a frameshift, the non-coding sequence gets transcribed and consequently translated as an ORF. Possible premature stop codons are to be erased, so the transcript of functionable length could appear. *Adapted from Zhang et al. 2019.*

Another mechanism of *de novo* gene birth lies in alternative translation of preexisting ORFs, i.e., *via* overprinting: the point mutation occurring inside a coding sequence may lead to emergence of the novel start codon, so two overlapping genes ORFs would arise. This scenario was demonstrated in *E. coli* (*Delaye et al. 2008*) and mouse (*Neme and Tautz 2013*). However, while the mechanisms of *de novo* gene birth have been uncovered significantly over the last decade, the structural and functional properties of de novo proteins are still heavily understudied and this limited knowledge is discussed in the following chapters of this thesis.

1.3. Synthetic proteins

Recent advances of synthetic and computational biology brought tremendous progress in design and selection of *de novo* artificial proteins: 4-helix bundle structures, beta sheet, beta barrel and more complicated structures were designed or selected from randomized scaffolds or fully random sequences (*Hecht et al. 1990; Xu et al., 2001; Park et al. 2006; Fisher et al. 2011; Thomson et al. 2014; Anishchenko et al. 2021*). There are three main approaches for *de novo* protein design (*reviewed by Huang, Boyken, and Baker 2016*):

a) **Structure prediction.** It is applied when the amino acid sequence is known, but the backbone arrangement is not given. For a created sequence, researcher can apply either homology modeling, which is dependent on the presence of homologous protein templates in databases, or *ab initio* prediction, when the prediction is carried out without any native templates (*Chothia and Lesk 1986; Bowie, Lüthy and Eisenberg 1991; Duan and Kollman 1998; Bonneau and Baker 2001*).

b) **Fixed-backbone design.** This case is the opposite of the previous one: for the known desired structure, a possible protein sequence is sought; for instance, it was used for the generation of structurally constrained peptides or evolution of novel function based on a specific structural scaffold (*Chao et al. 2013; Bhardwaj et al. 2016*).

c) ***De novo* design.** When neither structure, nor sequence is specified, *ab initio* design can be applied. Starting with a backbone generation and adjustment of amino acid side chains, a sequence candidate is selected. Using *de novo* design, alpha-helical, alpha-beta and repeat protein structures were created (*Thomson et al. 2014; Lin et al. 2015; Doyle et al. 2015*).

In this work, synthetic proteins are considered together with *de novo* born sequences. Moreover, as will be shown further, both synthetic and *de novo* proteins are often analyzed within or selected from sequence libraries**.**

Non-native protein sequences may represent an unexplored reservoir of biological functionality. Nowadays, a number of approaches (summarized in this thesis) have been developed to study the potential of *de novo* and random synthetic polypeptides. In the following chapters, different types of artificial and *de novo* sequences and respective methodologies of functional analysis are compared and discussed.

2. Characterization of naturally evolved *de novo* proteins

Many naturally evolved de novo proteins and designed proteins have been studied one by one, taking advantage of the late advances of molecular and cell biology. However, recent developments in the field of synthetic biology have also made it possible to characterize such proteins in a high-throughput format and to screen their properties in whole libraries of sequences. Specifically, this has been made possible mainly by the tremendous progress in the methods of parallel DNA synthesis, assembly and sequencing (*W. P. C. Stemmer et al. 1995; Zhou 2004*; *Engler et al. 2009; Currin et al. 2014; Heather and Chain 2016.*)

2.1. *De novo* protein library construction

Non-coding and artificial sequence space has been mimicked with a pool of DNA molecules (i.e., *de novo* library), which can be subjected to mutagenesis and selection occurring in nature. If the library is to best represent non-coding genome regions, parameters such as GC content, codon bias, amino acid frequencies and distribution, must be set in the libraries in accordance with the chosen model organism (*Galtier et al. 2018*). The various ways of library design are also distinguished by the amount of constraints applied:

1) The first case is represented by a totally random pool of DNA constructs, when almost no constraints were applied (*Cho et al. 2000; Chiarabelli, Vrijbloed, De Lucrezia, et al. 2006*). The nucleotide distributions in codons are set up to almost equal values, with the exception of 3rd nucleotide frequency for adenine, which was minimized to zero to reduce the stop codons frequencies (TAA and TGA). Another case is the introduction of required purification tags or restriction sites – these modifications are considered to be purely technical and do not affect probable functional potential of expressed random ORFs.
2) Next, some of the above-mentioned parameters can be considered in the design; for instance, genomic nucleotide and codon frequencies can be modeled (*Heames et al. 2022*) or a particular set of amino acids is favored, e.g., primordial amino acids pool (*Knopp et al. 2019*).

The physical implementation of the synthesized library strongly depends on the applied screening methods. In case of *in vivo* selections (bacterial or eukaryotic), the library should be delivered to cells in a plasmid or phage vector (*Gunge 1983*). Precise techniques of gene plasmid cloning were developed recently: for instance, Golden Gate assembly uses IIS type restriction enzymes cleaving the target sequence outside of recognition sites, therefore producing unique overhangs (*reviewed by Szybalski et al. 1991; Engler et al. 2009*). Knowing a distance between recognition and cleavage sites and designing proper flanking sequences serve as a very elegant solution even for big size gene assemblies, when up to 9 fragments can be simultaneously cloned into a vector in a desired order. An alternative solution can be represented by technology of phage display, when the target *de novo* sequence is linked to a bacteriophage capsid sequence (*G. P. Smith 1985; Malys et al. 2002*). The further paragraphs describing the particular experiments will show that sequence libraries and their selection are mostly indispensable for the screening of novel proteins.

2.2. Biophysical properties of random and *de novo* proteins

The propensity of any protein for function is tightly related to its ability to express and its biophysical properties. Aggregation propensities, solubilities and secondary structure occurrence of *de novo* and random libraries were recently examined. Heames et al. studied a set of putative *de novo* genes from *Homo sapiens* and *Drosophila melanogaster* and designed a comparable set of random genes based on the same length, GC content and amino acid composition. Both types of sequences were reported to contain similar amounts of secondary structure elements. However, higher solubility for the *de novo* library was observed suggesting that sequences with increased solubility get selected during *de novo* gene birth (*Heames et al. 2022*). Lower solubility is correlated with lower secondary structure content and therefore a protein's solubility seems to be more determining than its structural content (*Tretyachenko et al. 2017*). This is consistent with an observation that random proteins with higher proportion of intrinsically disordered regions (IDRs) and therefore with lower structural content are better tolerated *in vivo* (*Tretyachenko et al. 2017*). Generally, only a limited fraction of random sequences has been reported to be soluble upon overexpression (*Tanaka et al. 2010*). However, it has been pointed out that only very low expression is observed upon gain of *de novo* ORF translation and that no general trends in the structure of *de novo* proteins have been observed, besides perhaps the increased IDR content (*reviewed* by *Bornberg-Bauer et al. 2021*). More studies of *de novo* protein biophysical properties will therefore be needed to draw general conclusions about their properties. Discrepancies among different studies have been caused not only by the methodologies of screening, but also by the selection of specific sequences and potentially also at the stage of library design (*reviewed by Bornberg-Bauer et al. 2021*).

3. Approaches for functional design and selection

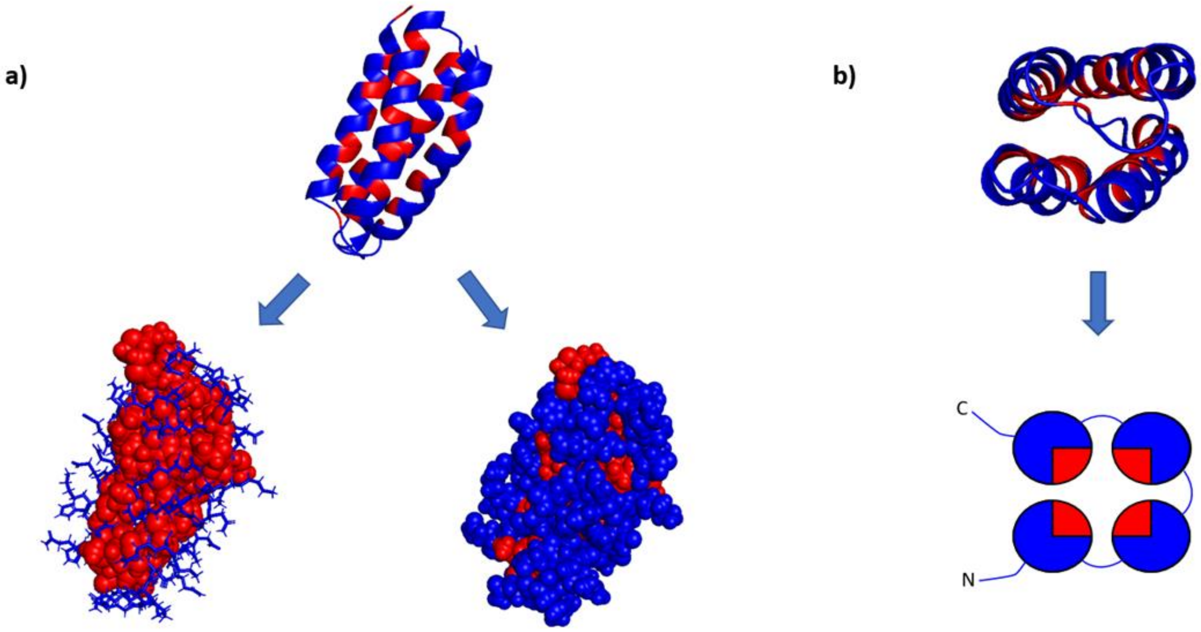
Unlike in the studies of naturally evolved or random *de novo* proteins (where the aim is to search if the protein may have some function), the main purpose of *de novo* protein design is obtaining a desired function. Protein design and selection for function have been performed using the following distinct approaches.

3.1. Combinatorial libraries

The first approach relies on selecting the most appropriate combination of variant protein libraries with a selection procedure. For example, in the case of searching for an efficient ATP-binder in the random sequence library, several rounds of ATP-immobilized chromatography combined with an mRNA-displayed library were used (*Keefe and Szostak 2001*) (see paragraph **4.2**). Such approaches start with a specifically designed library from which the gene coding for the desired function is amplified and selected.

Another possible approach is a continuous evolution of the starting library, most typically using error prone PCR. This method often builds on a starting protein target but avoids a specific design step. When used with an appropriate function selection method, this approach can be used for selecting both a binding as well as enzymatic activity. For example, the *de novo* ferric enterobactin esterase was derived from the member of 4-helix bundle combinatorial library by mutagenic error-prone PCR and *in vivo* selection (*Donnelly et al. 2018*). The primary protein template was able to rescue an *E. coli* knock-out mutant lacking the native esterase coded by *fes* gene (Δ*fes*) and threonine esterase (Δ*ilvA*); the Syn-IF gene cloned in the expression vector was used as template for the next step of mutagenesis in order to introduce the enterobactin esterase enzymatic activity (*B. A. Smith, Mularz, and Hecht 2015*). The new mutated version, Syn-F4, not only hydrolyzed the target molecule, but also demonstrated enantioselectivity degrading only L-enterobactin; it suggests that a native enzymatic function can be successfully reproduced from combinatorial libraries. In this experiment the error-prone PCR served as a source of sequence diversification; the method is further discussed within section **3.4**.

The next example of combinatorial library design is patterning the order of amino acid types in a sequence in order to produce the desired structural scaffold, as it can be demonstrated on 4-helix bundle libraries (*Regan and DeGrado 1988; Hecht et al. 1990; Kamtekar et al. 1993*). These small polypeptides with a stable secondary structure can be simply designed by alternating hydrophobic and hydrophilic amino acids, for example the typical pattern may be **PNPPNNPPNPPNNP**, where **P** is for polar, **N** is for non-polar; the structure of representative 4-helix bundle construct is depicted on **Fig. 5.** Binary patterned libraries have shown a high secondary structure formation propensity in NMR-studies (*Go et al. 2008*) and different functions rescuing in *E. coli* knock-out strains (*Fisher et al. 2011*). Besides purely alpha-helical structures that are easier to produce, novel beta-sheet proteins were also created (*West et al. 1999*). In that case the pattern was accordingly modified to obtain high beta-sheet formation, e.g. **PNPNPNPN,** which corresponds to the structural periodicity of a typical beta-sheet (**Fig.6.)**



c) (N)MYGKLNDLLEDLQEVLKHVNQHWQ*GGQK*NMNKVDHHLQNVIEDIHDFMQ*G*
*GGSGG*KLQEMMKEFQQVLDEIKQQLQ*GGDN*SLHNVHENIKEIFHHLEELVHR(C)

Fig.5. S-836, the 4-helix bundle synthesized and analyzed by *Go et al. 2008.* Nonpolar residues are colored blue, polar residues are colored red. **a)** Lateral projection of NMR-solved structure; nonpolar residues protruding from each helix form a stable core, whereas polar side chains cover the central core and interact with aqueous environment. **b)** Apical projection of S-836 – experimentally determined structure is totally in agreement with the theoretical model of 4-helix bundle fold. **c)** The S-836 amino acid sequence; the highlighted polar regions disrupting the given pattern represent turns, which connect the individual helices. The structure was adapted from https://www.rcsb.org/structure/2JUA

Beta-sheet strands tend to aggregate, for this reason Hecht and Wang had to change the pattern in order to disfavor the intermolecular hydrophobic interactions. The novel pattern was **PNPKPNP,** where **K** is for lysine (*Wang and Hecht 2001*).

The introduction of amino acid patterning is the initial step towards protein design by promoting stable secondary structures that will increase the probability of new function identification, although no specific function is designed. Patterned libraries can't be considered as totally random, since a strong sequence bias is introduced. Thus, the amino acid patterning represents an intermediate case between random and designed protein libraries.

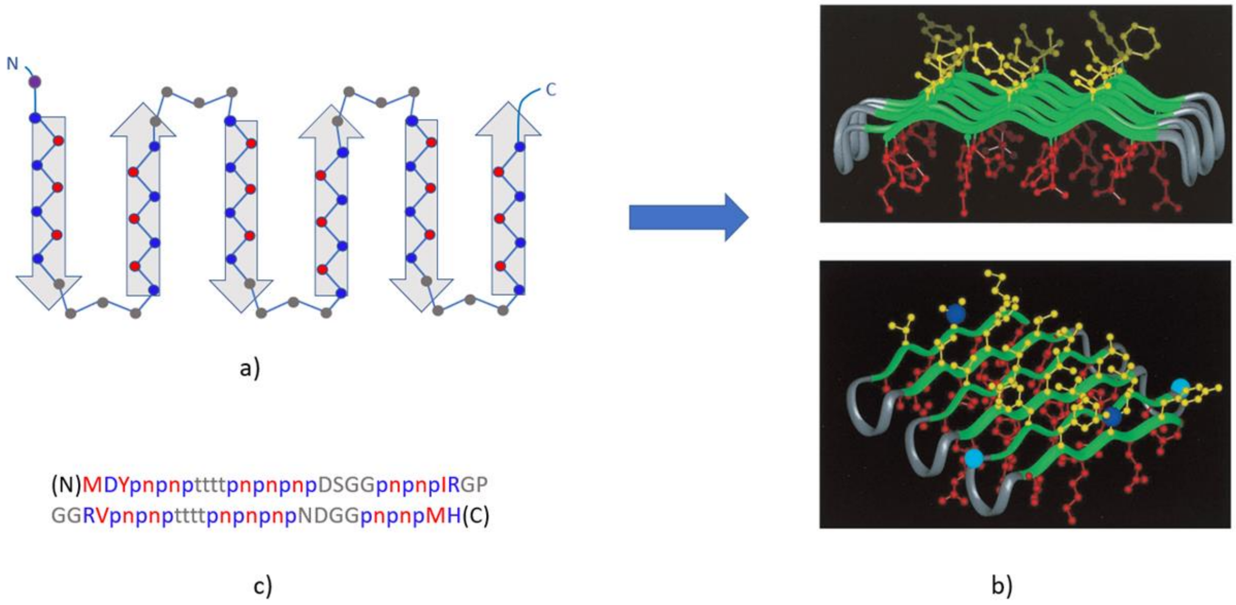

N
C
a)
(N)MDYpnpnpttttpnpnpnpDSGGpnpnpIRGP
GGRVpnpnpttttpnpnpnpNDGGpnpnpMH(C)
c)
b)

Fig.6. The *de novo* designed beta-protein structural pattern. Polar residues are colored blue, nonpolar residues are colored red, and turn amino acids are colored gray. The total sequence length is 63 amino acids, each beta-sheet fragment consists of 4 polar and 3 nonpolar amino acids, which alternate one after another. **a)** The scheme of the given beta protein. **b)** Computational model created in INSIGHT/DISCOVER package of programs, Molecular Simulations, Waltham, MA (taken from *Xu et al. 2001*); due to patterning the polar and nonpolar residues are protruding from opposite sides of the backbone plane, which allows the polymeric monolayer formation (not shown here). **c)** The sequence pattern used for the design; n stands for nonpolar amino acids, p for polar, t for turn. Adapted from *Xu et al. 2001.*

3.2. Design from preexisting native scaffold

Other approaches start with a pre-existing protein structure and target specific sites (such as the active or binding sites), whereas the general fold and remaining sequence are kept unchanged. For instance, changing 1 catalytic amino acid (tyrosine to alanine) was sufficient to convert PLP-dependent alanine racemase from *Geobacillus stearothermophilus* into an aldolase (*Seebeck and Hilvert 2003*). In other cases, several mutations were applied; the methodology of simultaneous incorporation and adjustment of functional elements (SIAFE) together with directed evolution techniques was used on the glyoxalase II αβ/βα metallohydrolase scaffold in order to generate beta-lactamase activity (*Park et al. 2006*). SIAFE includes 3 types of mutations naturally occurring in genomes: deletions, insertions and substitutions of particular protein elements, which, along with point mutations emerging during the directed evolution, realize the natural protein evolvability (*Aharoni et al. 2005*).

Another example of enzymatic redesign is represented by novel nucleic acid (NA) polymerases capable of synthesis of sequences assembled from non-conventional nucleotides, i.e., xeno-nucleic acid polymerases (*Pinheiro et al. 2012*). An original methodology called CSF (compartmentalized self-tagging) was developed to solve the problems connected with the strict polymerase substrate specificity. As a template, a variant of DNA polymerase from *Thermococcus gorgonarius* (TgoT) was used and subjected to several mutagenesis cycles that produced the mutant library used for the further rounds of CSF selection and functional screening. A single round of CSF includes the steps depicted on **Fig. 7**. The efficacy of vector isolation on streptavidin beads is critically dependent on primer extension – if the DNA synthesis occurs, the stronger binding pattern between attached primer and related vector is observed, thus only sufficiently extended primers enable the isolation of vector coding a functional enzyme.

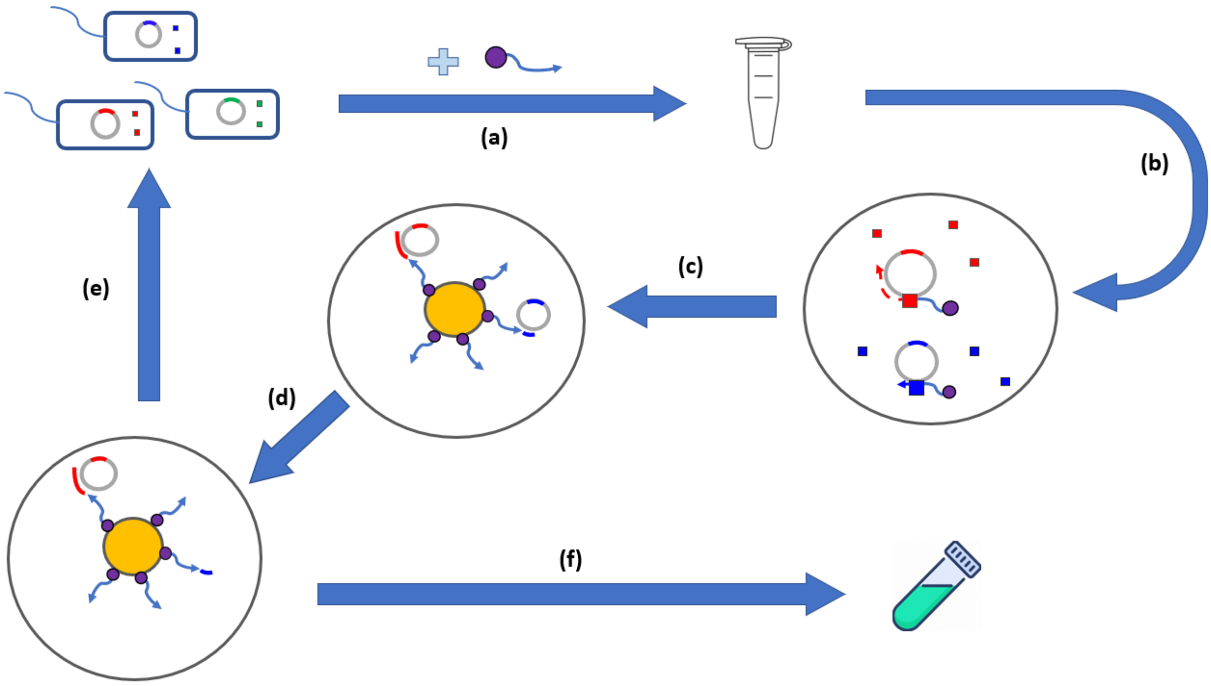


Fig.7. Compartmentalized self-tagging technique. The library of designed polymerases is cloned into a vector and transformed into *E. coli* strain. a) The *E. coli* library is incubated with modified nucleotides and respective biotinylated primers in water-in-oil emulsion. b) The obtained suspension undergoes PCR-like amplification. c) The extended biotinylated primers are incubated with streptavidin beads and the corresponding vectors attach to primers on the beads. d) The washing step allows to get rid of unspecific binding. e) The isolated sequences are subjected to further rounds of mutagenesis and CSF selection or f) screened *in vitro* for the polymerase activity. Adapted from *Pinheiro et al. 2012.*

3.3. Rational design

The rational computational protein design aims for specific protein engineering. It is often used in combination with directed evolution, that helps tuning the assembled structure and to select the most successful construct. One of the disadvantages of the rational approach is often the use of long demands in terms of cost calculation power. Recently, the important rules of design strategies were proposed by Koga et al.; they define the relations between local secondary structures and the respective tertiary motifs *via* controlling the lengths and the amino acid contents motifs, that allowed to precisely design and synthesize different Rossman-like, ferredoxin-like and other folds with high accuracy (*Koga et al. 2012*). The control of the structure quality in these experiments is carried out with NMR or X-ray crystallography. All the calculations were performed using the Rosetta program developed by David Baker's laboratory (*Rohl et al. 2004*). These and other related data (*Thomson et al. 2014; Lin et al. 2015;*

Huang et al. 2016) suggest that engineering of *de novo* stable protein scaffolds is feasible, which is the first step towards designing proteins with a specific function.

However, proteins designed in this fashion are not necessarily long and complex structures. As it was already demonstrated, small oligopeptides also possess functional activity (*reviewed by Storz et al. 2014*). The notable advantage of smaller proteins lies in the relative simplicity of their composition. The rules for high-throughput peptides design were proposed by Rocklin et al., as well as synthesis and screening techniques (*Rocklin et al. 2017*)*;* 15000 *de novo* miniproteins were created and tested for protease susceptibility assay in order to characterize their stability and folding. According to these principles, Chevalier et al. engineered and tested more than 20000 peptides (with length ranging from 37 to 42 amino acids) for the botulotoxin B and influenza hemagglutinin binding (*Chevalier et al. 2017*)*;* the computational part was done using the Rosetta prediction tool. This software enables modeling high-affinity peptide interactions *via* docking. Progress in this field also led to the possibility of creating novel patterns of interaction between various cationic amino acids and organic cofactors. A relatively small three-helical metalloenzyme with carbonic anhydrase activity was designed by Zastrow et al. using a TRI peptide backbone (*Zastrow et al. 2012*). The obtained three-stranded coiled coil was coordinated by Hg^{2+}, which played a key role in structure stabilization, and by Zn^{2+} necessary for the catalytic activity. Moreover, the same group used a non canonical sulfur-containing amino acid, penicillamine, instead of cysteine in order to obtain a more accurate X-ray crystallography model. These studies provide examples (among many other studies mentioned here for space reasons) of the computational design potential, which now makes it possible to approach proteins of different sizes, folds and cofactors.

However, searching for a new enzymatic activity does not inevitably involve only redesign of protein scaffolds. In some studies, significant results were reached by targeting both inorganic and organic cofactors. For example, substituting iron cation with other metal analogues (e.g., iridium) in porphyrin IX rings and their insertion into the native apo-PIX-protein scaffolds resulted in emergence of *de novo* heme metalloenzymes, which were able to catalyze new types of chemical transformation (*Key et al. 2016; Dydio et al. 2016*). It suggests that cofactor chemistry can serve as a robust source of catalytic variability. Similarly, the exposure of nicotinamide to light led to the photoexcitation event on the hydrogen atoms. In the excited state, nicotinamide enabled the catalysis of the new reactions *via* radical mechanism. It was demonstrated on the example of nicotinamide-dependent ketoreductases (KRED): these enzymes commonly catalyze the reduction of carbonyl to hydroxyl group; after the excitation, KREDs were able to induce radical formation and to dehalogenate halogen lactones. Moreover, a previously racemic mixture of halogen lactones was converted into chiral lactones of one type (*Emmanuel et al. 2016*).

Combination of *de novo* scaffold with a native cofactor may provide an activity resembling the activity of natural proteins: it was observed on the example of designed dimeric 4-helical protein with 4 heme groups installed. The measured electrochemical properties and the spectra resembled the native members of redox heme protein family, such as cytochrome *bc$_1$* or cytochrome *c* oxidase (*Roberston et al. 1994*). In the experiment performed by Roberston et al. was also utilized 4-helix bundle fold, as well as in case of *de novo* combinatorial libraries with binary patterning discussed in the previous paragraph (**Fig.5**). However, the selection of the proper protein-cofactor interaction is done through a computational design, when not only binary patterning was introduced, but also the necessary histidines were placed into binding sites in order to coordinate the prosthetic groups. It's noteworthy that the library of unevolved *de novo* 4-helix bundle proteins paired to naturally hemes was reported to have peroxidase activity, as well as esterase and lipase activity (*Moffet et al. 2000; Patel et al. 2009*).

3.4. The application of directed evolution

The computational approach combined with the directed evolution and the methods of structural biology can serve as an efficient way to achieve natural-like protein activity (*Karanicolas et al. 2011; Blomberg et al. 2013; Dydio et al. 2016*). The methodology for *de novo* protein production can be proposed:

1) A protein structure is computationally designed;
2) The protein corresponding the designed model is expressed, purified and analyzed by the methods of structural biology (e.g., X-ray crystallography, NMR, circular dichroism);
3) The necessary structural tuning is carried out by directed evolution;
4) The resulting protein variant is functionally screened.
5) Obtained data is used for further rounds of computational design, directed evolution and screening.

The described iterative technique was applied in the series of experiments done by Privett et al., when the first-generation designed version of Kemp eliminase (KE) showed no catalytic activity (*Privett et al. 2012*). X-ray crystallography and molecular dynamics (MD) simulations helped to investigate the reasons: the active site was too flexible; as a result, structural fluctuations led to exposure to aqueous environment, so water molecules were disturbing the interactions between the polar catalytic amino acid residues. The solvent barrier led to the next round of KE computational design, where the previous mistakes and inaccuracies were taken into account. Second KE design was subsequently analyzed *in silico* (by MD simulations) and *in vitro* (measuring the kinetics of catalyzed reaction). The obtained data allowed the design of the third KE construct. The created KE was used in another study done by Blomberg et al., where the same *de novo* enzyme was additionally improved by directed evolution (*Blomberg et al. 2013*). The first mutations were introduced into KE scaffold according to the data obtained from ligand docking – the computational approach, which simulates a ligand placement into the active site of designed structure (*Lassila 2006*). The following rounds of mutagenesis and *in vitro* selection allowed to isolate the best catalyst from the mutant library. This research demonstrates a practical application of several mutagenesis techniques in *de novo* protein creation:

a) **Error-prone PCR**. This modification of standard PCR utilizes the relatively low fidelity of *Taq* polymerase, which can be decreased even more by introducing of specific reaction conditions, for instance: increasing the concentration of $MgCl_2$, adding $MnCl_2$, unbalancing the concentrations of dNTPs, increasing the concentration of polymerase and increasing the elongation time (*Cadwell and Joyce 1992*). Error-prone PCR serves as a source of non-specific random mutations.

b) **DNA shuffling**. This technique allows to rearrange a pre-existing sequence; it consists of several steps:
   - fragmentation by DNAse I;
   - following PCR amplification of obtained fragments;
   - reassembly of fragments performed by restriction endonucleases and DNA ligase.

   DNA shuffling represents an additional tool for random mutagenesis, as well as previous technique (*W. P. Stemmer 1994*).

c) **PCR site-directed mutagenesis (SDM)**. It exploits overlapping primers, which already contain desirable mutations in their sequences (*Ho et al. 1989; Aiyar, Xiang, and Leis 1996*)*;* however, there are plenty of variations of this method (*reviewed by Shen 2002*).

   In contrast with two previous methods, PCR SDM is controllable, because the desired mutations are introduced into primers by design, not randomly.

Based on the above, the general experimental workflow for *de novo* protein design and analysis was proposed and depicted in **Fig.8.** The various approaches of directed evolution were reviewed by Yuan et al. 2005 (*Yuan et al. 2005*).

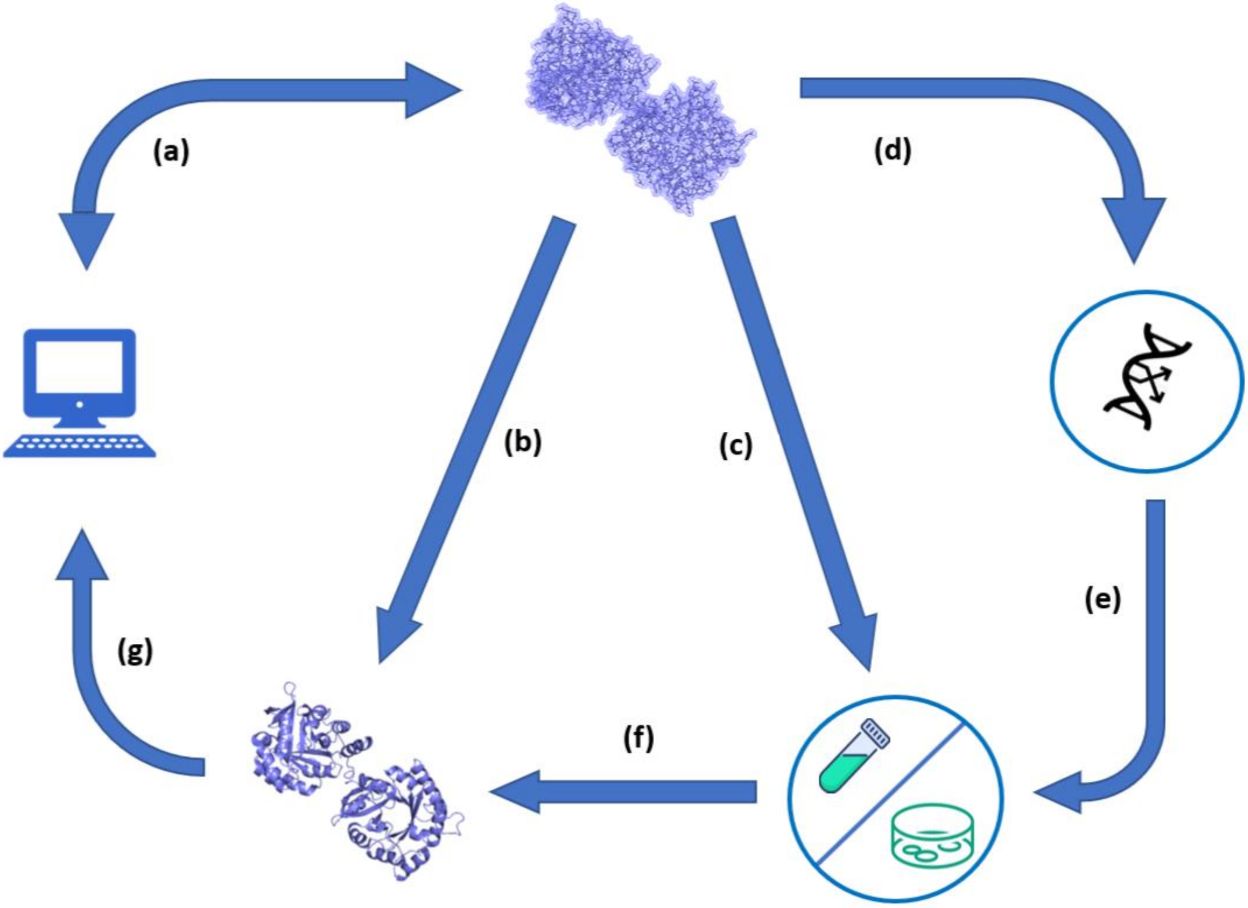


Fig.8. The proposed experimental workflow for iterative artificial protein design, tuning and testing. **(a)** first-generation computational protein model is created. **b)** The generated protein is expressed, purified and analyzed by structural methods, i.e., X-ray crystallography and/or NMR. **c)** the purified protein is experimentally characterized either *in vivo* or *in vitro*. **d)** The structure could go to directed evolution, which results in a mutant library to emerge; **e)** the desirable variants are similarly screened, selected and **f)** structurally analyzed. **g)** The overall obtained data may be further utilized for a new round of computational assessment and design. The number of iterations is variable, as well as involvement of individual steps: the simplest experimental arrangement can be restricted to the steps **(a), (b)** and **(g).** Based on *Privett et al. 2012 and Blomberg et al. 2013*.

4. *De novo* proteins functional characterization and selection

4.1. *In vivo* screening methods

One of the possible ways to study functional properties of *de novo* genes is cell system exploitation. A whole *de novo* protein library can be expressed in cells. In this regard, the advantages of using bacterial expression systems are indisputable (*reviewed by Terpe 2006*). Since bacteria are haploid, phenotypic manifestation of a new gene could be easily observed. Furthermore, numerous screening techniques for bacteria were developed, allowing testing of a wide range of possible functions. Unfortunately, the low level of post-translational modifications (PTMs) in prokaryotes reduces the functional potential of newly expressed proteins (*reviewed by Macek et al. 2019*).

4.1.1. Selections for improved fitness

As it was discussed above, the non-coding sequence space represents a potential source of variable functional activities, which can stimulate or inhibit bacterial cell growth (*Neme et al. 2017*). Neme et al. constructed a random sequence library (of equal nucleotide frequencies throughout the sequences) and used it as a proxy of non-coding genomic sequences. All constructs had the same length of 50 amino acids. The library was cloned into a pFLAG-CTC expression vector, which contains isopropylthiogalactoside (IPTG)-inducible promoter, and transformed into *E. coli* strain. The effects on cell growth with and without IPTG-induction were monitored in order to verify if the physiological effect was connected to peptides expressed from the library. Bacterial colonies with the empty vector were used as negative control. Clones with significant growth rate change were then identified via plasmid extraction and PCR with gene-specific primers, which allowed them to compare frequencies of different library variants and correlate them with growth effects. The screening itself was done initially *in vivo* in liquid LB media with different rounds of reinoculation combined with an *in vitro* identification step. The expression level was also checked by Western blot analysis. This study provided a starting platform for all similar screening techniques, the most important and discussed steps in this kind of experiment being: 1) an appropriate negative control; 2) use of a tunable expression vector; 3) extraction and characterization of the selected variants. The experimental conditions in the experiment done by Neme et al. were set up as optimal, whereas in other studies a stressful environment was simulated (*Stepanov and Fox 2007*). The reason for inducing cellular stress during such experiments is that the selection pressure for novel genes extremely increases: under the accelerated selection in stressful conditions, genes that confer a resistant phenotype will be favored and quickly fixed in the population. In the work of Stepanov and Fox, IPTG-induction was also used, although the size of ORFs was reduced from 50 to 20 amino acids. The stressful environment was mimicked by subinhibitory concentrations of $NiCl_2$, $AgNO_3$, or K_2TeO_3. These conditions were chosen due to a simple controllability of chemical concentration, while ultraviolet radiation or oxygen stress are more lethal and much more difficult to control. Stress-induced mutagenesis brings an additional source of genetic variability for *de novo* libraries selection (*reviewed by Foster 2007*).

Similarly, the 4-helix bundle designed protein was reported to increase the cell's tolerance to higher concentrations of copper cations (*Hoegler and Hecht 2016*). The MIC of $CuCl_2$ was determined for *E. coli* negative control cells, which were transformed with an empty vector carrying beta-galactosidase. The same strain was consequently transformed with the cloned library, in this way the negative control could be compared to the library transformants. After several rounds of directed evolution, the selected variants were able to grow at 7.0 mM concentration of $CuCl_2$ and $CuSO_4$, whereas WT maximally tolerated only 4.4 mM. The protein library did not show any further resistance to high concentrations of other divalent metal cations, such as Ni^{2+}, Co^{2+} and Zn^{2+}.

4.1.2 Selections for specific functions

In vivo screening/selection techniques can also be used to search for a specific function. This pipeline has been pioneered by Knopp et al. when searching if random sequences can confer antibiotic resistance (*Knopp et al. 2019*). The sizes of ORFs ranged from 10 to 50 amino acids. The library was cloned into low-copy-number expression vector pRD2 and transformed into *E. coli*, which was grown in the liquid culture and consequently spread on a solid agar medium with 12 different antibiotics types. It's noteworthy that selecting individual clones on solid media, e.g., on different agars, is more convenient and doable than doing so in liquid cultures – required colonies can be easily scraped and isolated from solid surface. As in the previous experiments, the expression was controlled by an IPTG-inducible promoter. For each antibiotic screening the negative control was made by transforming an empty vector. From plates containing kanamycin, 3 peptide products were isolated (named Arp1-3), and their minimal inhibitory concentrations (MIC) were measured also with all other 11 antibiotics. For Arp1, a 48-fold increase of MIC on amikacin plate was observed. The Arp1-3 mechanism of action was thoroughly examined: the chosen Arp1 was fused to His-tag on the C-terminus, transformed into the cells and incubated with anti-His-tag antibodies labeled with gold particles. Transmission electron microscopy revealed the localization of non-aggregated protein molecules at the plasma membrane. However, the protein aggregates were also observed across the cell cytoplasm, which could be explained by the hydrophobic sequence of Arp proteins. Hydrophobic sequences tend to form inclusion bodies when expressed in cells (*Upadhyay et al. 2012*). To further analyze the Arp proteins mechanism of action, the changes in membrane electrochemical gradient were assessed by using fluorescent dye bis-(1,3-dibutylbarbituric acid)trimethine oxonol [DiBAC$_4$(3)], a voltage-sensitive fluorophore: its uptake by cell strongly depends on the membrane electrochemical potential – when the membrane depolarizes, the increased DiBAC$_4$(3) uptake can be observed (*Epps, Wolfe, and Groppi 1994*). As a negative control, a non-treated *E. coli* strain was used. As a positive control, the protonophore carbonyl cyanide m-chlorophenyl hydrazone (CCCP) was applied, since it disrupts the membrane potential. The DiBAC$_4$(3) assay revealed the ability of Arp1 to reduce the electrochemical potential. The results suggest that the selected library members act as small hydrophobic membrane peptides (the sizes vary from 22 to 25 amino acids) causing membrane depolarization and therefore lower antibiotic uptake by cells. A similar strategy was used in more recent research, where a similar random library was screened for colistin resistance (*Knopp et al. 2021*). 6 ORFs were identified to provide the given phenotype *via* specific interaction with sensor kinase PmrB and modification of surface lipid A domain of lipopolysaccharide (LPS). In both cases the antibiotic resistance is due to altered membrane features, even though *via* different mechanisms. Nevertheless, novel function of random short proteins was manifested by new interacting patterns with the native cell structures, such as plasma membrane or protein components of different pathways (*reviewed by Storz et al. 2014*).

Another possible way to search for a specific function *in vivo* lies in "function rescue" experiments. This approach was pioneered by the Hecht group (*Fisher et al. 2011*), who performed this kind of screening on *E. coli* single-gene knockout mutants (*Baba et al. 2006*). The synthesized 4-helix bundle library of 1.5 x 10^6 variants was cloned into the pCA24NMAF2 expression vector (each library sequence was flanked by an IPTG-inducible promoter). For the experiment 27 different knock-out auxotrophic strains were chosen according to the inability to grow on the minimal agar medium (M9-glucose medium). The main goal was to discover a new protein, which would rescue a particular knockout mutant. Different controls were prepared in order to obtain more robust results: the positive control was represented by the wild-type gene cloned into the same vector and transformed into a chosen strain while the negative one contained only the empty vector. In result, only 4 out of 27 knock-out strains were rescued. All of those knockout genes code enzymes participating in different biosynthetic pathways: phosphoserine phosphatase (*serB,* the final enzyme of serine biosynthetic pathway); citrate synthase (*gltA,* glutamate biosynthesis); threonine deaminase (*ilvA,* isoleucine biosynthesis from threonine); enterobactin esterase (*fer*, hydrolase cleaving iron-enterobactin complex). The next isolation and analysis steps were supposed to reveal the mechanism of how these new proteins were able to rescue those functions. The authors took into account different scenarios: either the library protein adopts the activity of deleted enzymes or the selected variants interact with alternative pathways providing a bypass for the original one (*Moffet et al. 2000; Patel et al. 2009*). Furthermore, they could function as transcription regulators enhancing or inhibiting some gene expression; stimulating general stress response; interacting with endogenous proteins and allosterically modifying their activity etc. In order to analyze biochemical activities of the studied library variants *in vitro*, the selected *E. coli* clones were isolated and sequenced. After the identification step, the plasmids were retransformed into *E. coli* strain for protein expression and purification. The purified *de novo* proteins were then tested for enzymatic activity: each library protein was assayed for the biochemical function of corresponding native protein, which was deleted in the rescued strain. In this case no biochemical activity was detected during the *in vitro* enzymatic assays. The alternative mechanisms mentioned above could be investigated in the following way:

1) Novel biochemical bypass emergence could be tested *via* deletion of other enzymatic components of the respective biochemical pathways. For instance, this is how a novel bypass pathway of pyridoxal-5'-phosphate (PLP) synthesis was discovered in the work of Kim et al. The experiment was performed on knockout *E. coli* mutants, which contained the deletion of 4-phosphoerythronate dehydrogenase (PdxB): it is the key enzyme of main PLP biosynthetic pathway in *E. coli* (*Kim et al. 2010*). The cells lacking more than one enzyme could be rescued, if the *de novo* sequence enables a bypass of the whole pathway. This option was tested and excluded by Fisher et al. in their study.
2) Assessment of interacting patterns is more complicated and is mostly performed *in vitro* (see chapter **4.2**), nevertheless there are ways to analyze interactions with endogenous proteins or nucleic acids *in vivo*. Therefore, *de novo* library variants may be able to interact with genomic *cis*-regulatory elements and alter gene expression profiles. Remarkably, Patrick et al. discovered the multicopy suppression genes in the *E. coli* genome: when overexpressed, these multicopy suppressors are able to rescue functions of non-related genes, which was demonstrated in *E. coli* deletion mutants (*Patrick et al. 2007*). In addition, the induced overexpression of the particular genes was reported to

be responsible for toxin and antibiotic resistance (*Soo, Hanson-Manful, and Patrick 2011*). In the work of Fisher et al., multicopy suppressors were identified for 3 out of 4 rescued E. coli auxotrophic mutants (*Fisher et al. 2011*). To test the ability of artificial proteins to enhance the expression of related multicopy suppressors, the double-deletion mutants are created lacking both the rescued gene and its respective suppressor; thus, if the library protein is acting *via* the discussed mechanism, such double mutants won't be rescued. That option was tested and rejected by Fisher et al. for most of the target genes. Only two double-knockouts weren't created:

- Δ*fes*Δ*thiL,* since the Δ*thiL* mutant is non-viable (*thiL* encodes thiamine-monophosphate kinase, responsible for the last step thiamine-pyrophosphate biosynthesis)
- Δ*serB*Δ*hisB* (*hisB* encodes histidinol phosphate phosphatase, the participant of histidine biosynthetic pathways), because authors didn't consider the situation, where that double auxotroph would be rescued by single 4-helix bundle novel protein.

Nonetheless, the effect on transcription regulation was analyzed in a follow-up study: the isolated library sequence (named SynSerB3) was able to rescue Δ*serB* mutant, even though not *via* enzymatic catalysis (*Digianantonio and Hecht 2016*). After transformation of the vector carrying SynSerB3 into non-deletion *E. coli* strain and sequencing of the cell's mRNA pool, a 10-fold overexpression from histidine biosynthetic operon was revealed. This discovery was additionally supported by the results of quantitative PCR performed on cDNA, which was produced from mRNA pools from non-deletion transformants and Δ*serB* transformants. These data suggest that the library variant affected the expression of his-operon.

3) The sequences selected from the library may support the cell growth via stimulating the stress response, since misfolded protein structures represent one of the major sources for cellular stress induction (*reviewed by Kültz 2005*). To check this option, structural analysis could be applied. The *de novo* proteins, especially those selected from libraries with no defined amino acid pattern, may contain less secondary structure elements. On the other hand, the 4-helix bundle library is considered to be another case: as was shown in previous research of Go et al., the designed alpha-helical content was conserved in practice (*Go et al. 2008*)

Finally, a more recent technique to select for function from large protein libraries is represented by microfluidics droplets - microscopic compartments of picoliter volumes, which can emerge in solutions consisting of two immiscible phases, such as oil and water mixtures. The generation, physical properties and applications of microfluidic droplets were reviewed by Teh et al. (*Teh et al. 2008*). In particular, microfluidic droplets were demonstrated to effectively sequester biomolecules and cells in microscopic compartments, increasing the probability of possible interactions and the following processes. The isolation and the consequent lysis of individual cells in these microscopic droplets allow to analyze and measure the activity of a particular protein pool originating from one cell. Moreover, it allows to link genotype to phenotype: the respective DNA sequence is released during the cell lysis and retained in a droplet, where the active library protein is functioning. This droplet microfluidics application was demonstrated in the work of Kintses et al., who analyzed the mutant library of arylsulfatase from *Pseudomonas aeruginosa* (*Kintses et al. 2012*). The library was created by error-prone PCR (see paragraph **3.4**) in order to select the variants with the increased catalytic activity to cleave phosphonate. The specific fluorescent assay was used: a phosphonate-like substrate, bis(methylphosphonyl)-fluorescein, was applied. The product of the substrate hydrolysis is fluorescein and therefore easy to detect and sort. The mutant arylsulfatase library was transformed into the *E. coli* strain, after the protein expression the individual cells were compartmentalized into microfluidics droplets and lysed. The released library proteins were assayed and the enzyme kinetics were measured. This methodology combines both *in vivo* and *in vitro* steps and represents an effective high-throughput approach of *de novo* and mutant enzyme libraries screening (*Pinheiro et al. 2012; Fischlechner et al. 2014; Obexer et al. 2017*)

According to the above, there are multiple options on how to screen/select for the function of *de novo* sequences *in vivo* (**Fig. 9**). The cell systems were demonstrated to be quite useful in determining physiological effects of artificial sequences. The important feature of *in vivo* expression systems lies in the ability to independently produce and assemble native or *de novo* protein structures, which may be further isolated, purified and additionally analyzed *in vitro,* as it was done in every experiment described in this section.

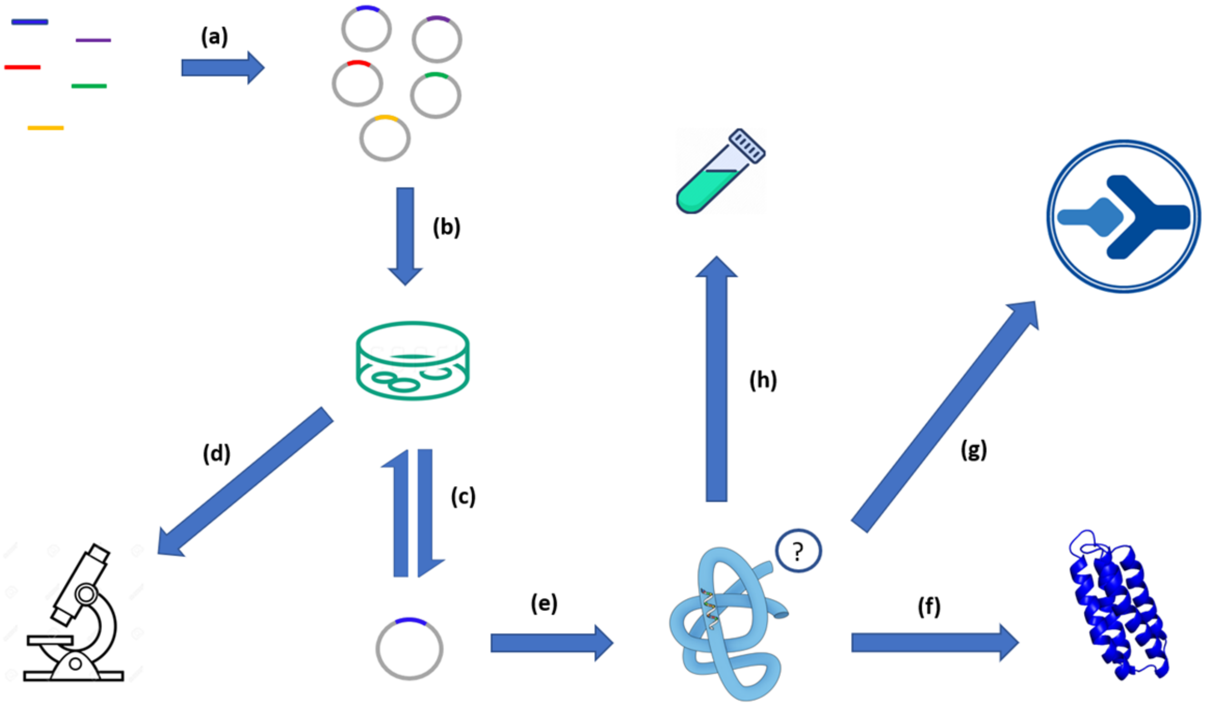


Fig.9. *In vivo* screening experiment workflow. **a)** *De novo* sequence pool is cloned into an expression vector. **b)** The cloned library is transformed into a cell strain for direct selection. **c)** The DNA sequences responsible for the respective phenotype are isolated and characterized. **d)** The microscopical observation, optionally combined with various *in vivo* immunohistochemical assays, may be carried out in order to detect the protein localization in a cell. **e)** The coded *de novo* protein is expressed and purified for a further analysis: **f)** solving the structure by NMR, X-ray crystallography or other methods; **g)** *in vitro* binding assays; **h)** *in vitro* biochemical assays.

4.2. *In vitro* screening and selection

In the pioneering work of Keefe and Szostak, the mRNA display technique was used to search for ATP binding in a library of completely random sequences. The methodology relies on construction of the mRNA-peptide fused library: each *de novo* DNA sequence is converted into the respective mRNA, which is ligated to a puromycin linker and translated; as the result, the appearing peptide is bound *via* its carboxy-terminus to its respective mRNA. Then, this construct undergoes the reverse transcription leading to the formation of cDNA-peptide fusion library, which can be further screened for the binding abilities (**Fig. 10).** In the discussed experiment, the final library is screened for the ATP binding by column affinity chromatography with the stationary phase represented by immobilized ATP-agarose granules.

The whole procedure is performed iteratively (the number of iterations is determined by the researcher at will) in order to select the most successful binders (*Roberts and Szostak 1997; Keefe and Szostak 2001*).

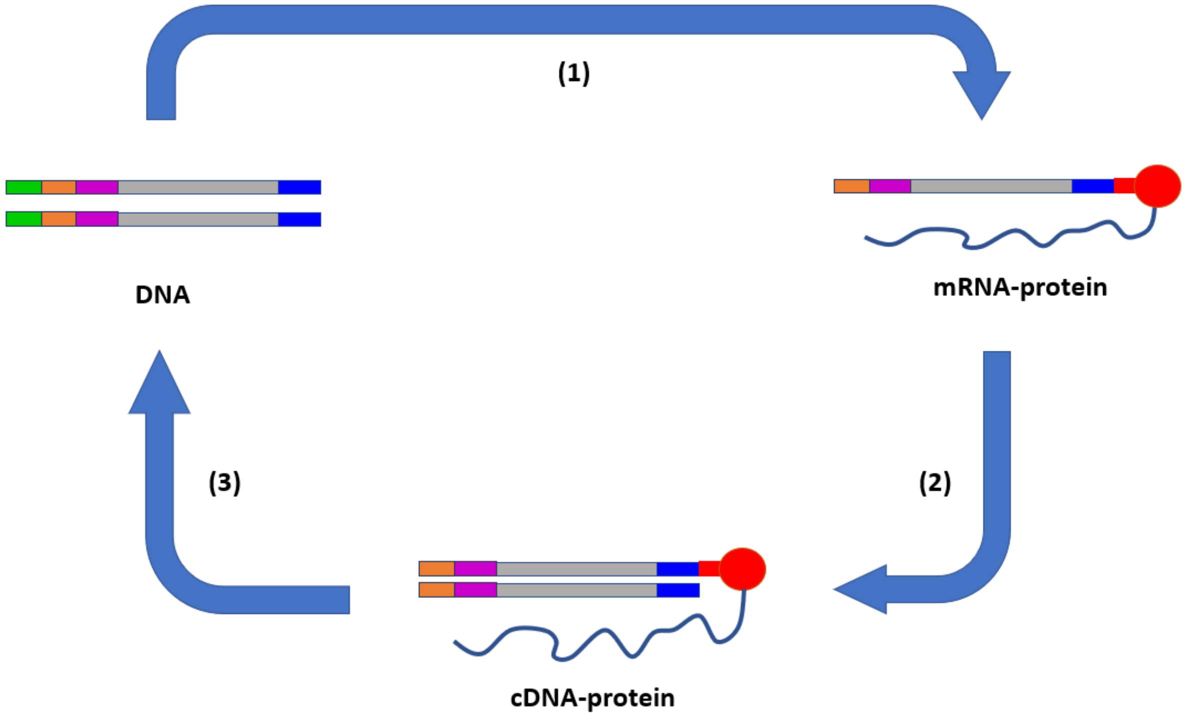


Fig.10. Preparation and selection of the *de novo* mRNA-displayed proteins. **(1)** The DNA sequence (gray) coding a library variant is flanked by two different purification tags (magenta and blue); a strong RNA polymerase promoter (green) and a translation enhancer (orange) are added upstream the translated ORF region. The DNA library constructed in this fashion undergoes transcription, ligation to puromycin linker and translation in order to produce mRNA-protein fused library. **2)** mRNA-protein library is then purified by oligo(dT) and other affinity chromatography techniques (according to the added purification tags). After purification, the cDNA-protein fused library is created by reverse transcription. **3)** The obtained constructs are selected by affinity chromatography with stationary phase represented by target molecules (in the work of Keefe and Szostak, the ATP-agarose beads were chosen). The strongest protein binders are eluted and the linked cDNA is amplified by PCR. The resulting DNA pool is enriched with the sequences coding the most successful ATP-binders. The whole cycle is repeated in order to select the most efficient protein binders. Based on *Cho et al. 2000* and *Keefe and Szostak 2001.*

An alternative approach is represented by phage display, a widely used technique for protein interaction assessment, including antibody screening (*Barbas et al. 1991*). It is based on the genetically engineered bacteriophages, when the phage DNA coding the coat protein is fused to the DNA sequence of tested protein. A produced phage library, in which an individual virion carries the tested protein on its surface and also contains the respective DNA inside, is screened *in vitro* on the plate with immobilized binding targets, such as other proteins or nucleic acids (**Fig.11**). This method combines both *in vivo* phage production and *in vitro* screening step. The phage type is chosen between two main classes - filamentous and T7 phages; each of these groups has its own pros and cons. The filamentous phages are applied more frequently (for instance, M13 phage), because they possess several coat proteins (e.g., pIII or pVIII), which can be easily fused to a foreign protein without disrupting a coat structure, and because of the sufficient capsid volume, which is able to fit longer DNA sequences (*G. P. Smith 1985; Il'ichev AA 1989*). Nonetheless, the non-lytic phage production in this case requires the export of all phage components separately through the bacterial inner membrane into periplasmic space, where the phage particle is assembled, so it isn’t suitable for every tested protein; T7 phage display overcomes this problem, although the size of constructed DNA is constrained by the smaller volume of T7 virion (*Danner and Belasco 2001*). The phage DNA with an inserted foreign sequence can contain all necessary information for virion production, as it was initially developed (*McCafferty et al. 1990*). Further modifications of phage DNA led to the construction of phagemid vectors – genetic constructs, which combine features of a bacterial plasmid and a phage DNA. Because of several deletions in phage genome, the phagemid vector is unable to establish normal production of progeny virions, thus the helper phage is required for the DNA packaging and virus particle assembly inside a bacterial cell; on the other hand, longer fragments of foreign DNA may fit into a phagemid due to deletions of original phage genes (*reviewed by Qi et al. 2012*). The phagemid-based phage display was used in the studies by Chiarabelli et al., who developed and tested the method for the screening of folded random peptide sequences (*Chiarabelli, Vrijbloed, Thomas, et al. 2006; Chiarabelli, Vrijbloed, De Lucrezia, et al. 2006*). The phage display has found an application in the antibody design and screening: for instance, the library of heavy and light variable chains (V_H, V_K and V_λ) was created and displayed on the surface of filamentous phage in order to simulate the natural antibody selection, which occurs normally in B lymphocytes (*Marks et al. 1991; Winter et al. 1994*).

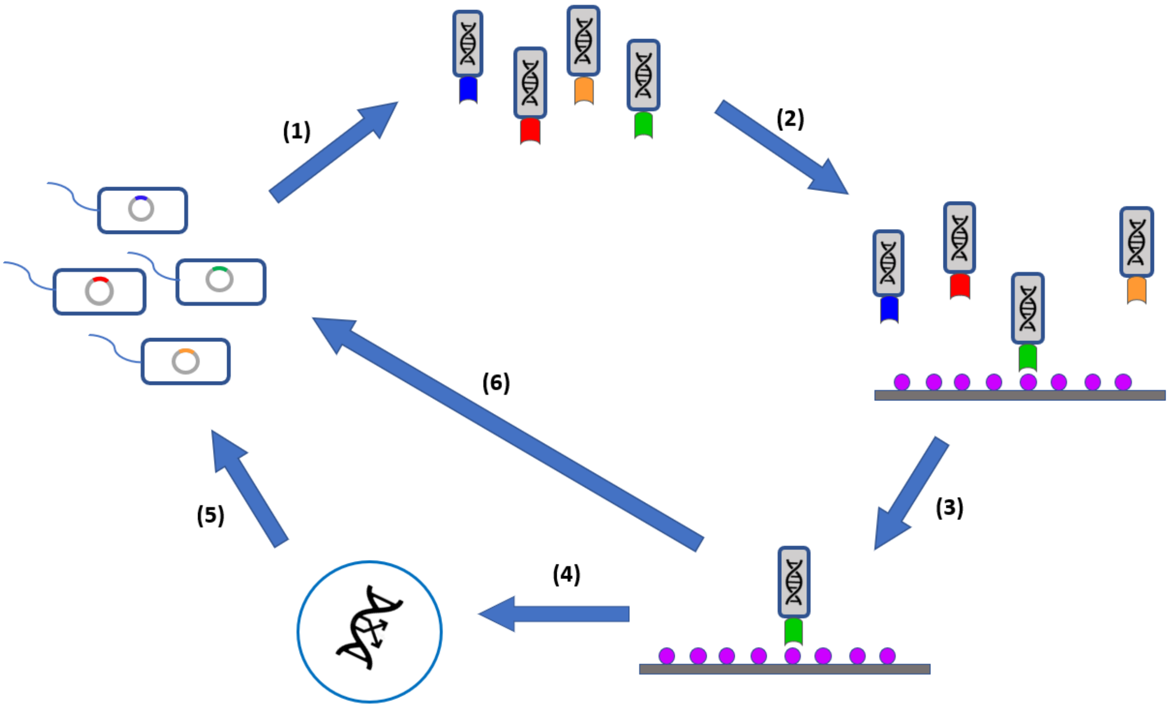


Fig.11. The phage display workflow. **1)** After the phage library vectors are delivered to bacterial cells, the phage population is produced and purified. Each virion contains a library protein fused to the capsid on the outer surface. **2)** The phage pool is screened on a surface with immobilized target molecules. **3)** Non-binders are washed away; the bound phage is eluted. **4)** Subsequently, the isolated phage DNA can be subjected to mutagenesis in order to create a new mutant library. **5), 6)** The isolated phage DNA is used for reinfection. The step of mutagenesis can be skipped. Based on the review of *Smith and Petrenko 1997.*

According to the above, both phage and mRNA display represent a very useful tool for binding testing and they are applicable for a wide range of interactions (protein-protein, protein-small molecule, protein-nucleic acid etc.).

5. Conclusions

This thesis is focused on the phenomenon of *de novo* proteins, i.e., novel proteins that arise either from genome non-coding regions or as products of bioengineering design. As such, it is related to the vast sequence space and its propensity to form structure and function.

De novo gene evolution represents one of the many ways of how novel genes can arise in nature. Compared with other mechanisms of protein evolution (such as gene duplication and recombination), such events are probably extremely rare. While myriads of random transcripts appear continuously, the vast majority of them vanish fast and only a small fraction is retained and does not get purged upon translation. The sequence properties of such *de novo* proteins are almost indistinguishable from random sequences and such sequences have been used as proxies to study what properties make them selected during the *de novo* gene birth. Moreover, how potentially functional proteins evolve in nature is of great interest in the field of protein design. The methodology to study the phenomenon of function emergence overlap significantly for these two directions of study and both are therefore summarized here.

The experiments exploring functional properties of non-native proteins typically take advantage of large sequence libraries, where the probability of encountering a functional variant is maximized. This field of research therefore pioneered numerous methods of how combinatorial peptide libraries can be designed and constructed.

The naturally evolved *de novo* proteins have been studied mainly by fully random libraries in which parameters such as GC content (of the DNA template) and amino acid composition have been controlled. In contrast, the "designed" *de novo* proteins have been selected from libraries that were typically constructed from specific templates or scaffolds by mutagenesis or rational design.

To select functional protein variants from such libraries (often exceeding millions of variants), many high-throughput screening and selection pipelines have been evolved over the last two decades. While listing all such methods was beyond the scope of this thesis, the ones that have been most relevant for the research of *de novo* proteins were thoroughly summarized and illustrated on selected exemplary studies. These include solely *in vitro* or *in vivo* techniques, however, the combination of both can be applied as well. For proteins that exhibit totally novel non-natural activity, the standard screening methods have been adapted in very unique ways and such examples were also described in the thesis.

The continuous progress in the protein sequence space investigation and advances of computational design will definitely keep inspiring more creativity and will lead to development of new specific techniques to search for protein function, natural or novel.

6. References

1. Aharoni, Amir, Leonid Gaidukov, Olga Khersonsky, Stephen McQ Gould, Cintia Roodveldt, and Dan S Tawfik. 2005. "The 'evolvability' of Promiscuous Protein Functions." *Nature Genetics* 37 (1): 73–76. https://doi.org/10.1038/ng1482.
2. Aiyar, Ashok, Yan Xiang, and Jonathan Leis. 1996. "Site-Directed Mutagenesis Using Overlap Extension PCR." In *In Vitro Mutagenesis Protocols*, by Michael K. Trower, 57:177–92. New Jersey: Humana Press. https://doi.org/10.1385/0-89603-332-5:177.
3. Anantharaman, Vivek, L Aravind, and Eugene V Koonin. 2003. "Emergence of Diverse Biochemical Activities in Evolutionarily Conserved Structural Scaffolds of Proteins." *Current Opinion in Chemical Biology* 7 (1): 12–20. https://doi.org/10.1016/S1367-5931(02)00018-2.
4. Anishchenko, Ivan, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, et al. 2021. "De Novo Protein Design by Deep Network Hallucination." *Nature* 600 (7889): 547–52. https://doi.org/10.1038/s41586-021-04184-w.
5. Baba, Tomoya, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. 2006. "Construction of *Escherichia Coli* K-12 In-frame, Single-gene Knockout Mutants: The Keio Collection." *Molecular Systems Biology* 2 (1). https://doi.org/10.1038/msb4100050.
6. Barbas, C F, A S Kang, R A Lerner, and S J Benkovic. 1991. "Assembly of Combinatorial Antibody Libraries on Phage Surfaces: The Gene III Site." *Proceedings of the National Academy of Sciences* 88 (18): 7978–82. https://doi.org/10.1073/pnas.88.18.7978.
7. Bazzini, A. A., T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, et al. 2014. "Identification of Small ORFs in Vertebrates Using Ribosome Footprinting and Evolutionary Conservation." *The EMBO Journal* 33 (9): 981–93. https://doi.org/10.1002/embj.201488411.
8. Bhardwaj, Gaurav, Vikram Khipple Mulligan, Christopher D. Bahl, Jason M. Gilmore, Peta J. Harvey, Olivier Cheneval, Garry W. Buchko, et al. 2016. "Accurate de Novo Design of Hyperstable Constrained Peptides." *Nature* 538 (7625): 329–35. https://doi.org/10.1038/nature19791.
9. Blomberg, Rebecca, Hajo Kries, Daniel M. Pinkas, Peer R. E. Mittl, Markus G. Grütter, Heidi K. Privett, Stephen L. Mayo, and Donald Hilvert. 2013. "Precision Is Essential for Efficient Catalysis in an Evolved Kemp Eliminase." *Nature* 503 (7476): 418–21. https://doi.org/10.1038/nature12623.
10. Bonneau, Richard. 2001. "AB INITIO PROTEIN STRUCTURE PREDICTION: Progress and Prospects," 18.
11. Bornberg-Bauer, Erich, Klara Hlouchova, and Andreas Lange. 2021. "Structure and Function of Naturally Evolved de Novo Proteins." *Current Opinion in Structural Biology* 68 (June): 175–83. https://doi.org/10.1016/j.sbi.2020.11.010.
12. Bowie, James U, ROLAND LtCY, and David Eisenberg. 1991. "A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Stucture" 253: 7.
13. Cadwell, R C, and G F Joyce. 1992. "Randomization of Genes by PCR Mutagenesis." *Genome Research* 2 (1): 28–33. https://doi.org/10.1101/gr.2.1.28.
14. Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, et al. 2012. "Proto-Genes and de Novo Gene Birth." *Nature* 487 (7407): 370–74. https://doi.org/10.1038/nature11184.
15. Chao, Fa-An, Aleardo Morelli, John C Haugner Iii, Lewis Churchfield, Leonardo N Hagmann, Lei Shi, Larry R Masterson, Ritimukta Sarangi, Gianluigi Veglia, and Burckhard Seelig. 2013. "Structure and Dynamics of a Primordial Catalytic Fold Generated by in Vitro Evolution." *Nature Chemical Biology* 9 (2): 81–83. https://doi.org/10.1038/nchembio.1138.
16. Chen, Sidi, Yong E. Zhang, and Manyuan Long. 2010. "New Genes in *Drosophila* Quickly Become Essential." *Science* 330 (6011): 1682–85. https://doi.org/10.1126/science.1196380.
17. Chevalier, Aaron, Daniel-Adriano Silva, Gabriel J. Rocklin, Derrick R. Hicks, Renan Vergara, Patience Murapa, Steffen M. Bernard, et al. 2017. "Massively Parallel de Novo Protein Design for Targeted Therapeutics." *Nature* 550 (7674): 74–79. https://doi.org/10.1038/nature23912.
18. Chiarabelli, Cristiano, Jan W. Vrijbloed, Davide De Lucrezia, Richard M. Thomas, Pasquale Stano, Fabio Polticelli, Tiziana Ottone, Ester Papa, and Pier Luigi Luisi. 2006. "Investigation Ofde Novo Totally Random Biosequences, Part II: On the Folding Frequency in a Totally Random Library Ofde Novo Proteins Obtained by Phage Display." *Chemistry & Biodiversity* 3 (8): 840–59. https://doi.org/10.1002/cbdv.200690088.
19. Chiarabelli, Cristiano, Jan W. Vrijbloed, Richard M. Thomas, and Pier Luigi Luisi. 2006. "Investigation Ofde Novo Totally Random Biosequences, Part I: A General Method Forin Vitro Selection of Folded Domains from a Random Polypeptide Library Displayed on Phage." *Chemistry & Biodiversity* 3 (8): 827–39. https://doi.org/10.1002/cbdv.200690087.

20. Cho, Glen, Anthony D Keefe, Rihe Liu, David S Wilson, and Jack W Szostak. 2000. “Constructing High Complexity Synthetic Libraries of Long ORFs Using In Vitro Selection.” *Journal of Molecular Biology* 297 (2): 309–19. https://doi.org/10.1006/jmbi.2000.3571.
21. Chothia, C., and A.M. Lesk. 1986. “The Relation between the Divergence of Sequence and Structure in Proteins.” *The EMBO Journal* 5 (4): 823–26. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.
22. Copley, Shelley D. 2020. “The Physical Basis and Practical Consequences of Biological Promiscuity.” *Physical Biology* 17 (5): 051001. https://doi.org/10.1088/1478-3975/ab8697.
23. Cordes, Matthew HJ, Alan R Davidson, and Robert T Sauer. 1996. “Sequence Space, Folding and Protein Design.” *Current Opinion in Structural Biology* 6 (1): 3–10. https://doi.org/10.1016/S0959-440X(96)80088-1.
24. Currin, Andrew, Neil Swainston, Philip J. Day, and Douglas B. Kell. 2014. “SpeedyGenes: An Improved Gene Synthesis Method for the Efficient Production of Error-Corrected, Synthetic Protein Libraries for Directed Evolution.” *Protein Engineering, Design and Selection* 27 (9): 273–80. https://doi.org/10.1093/protein/gzu029.
25. Danner, Stefan, and Joel G. Belasco. 2001. “T7 Phage Display: A Novel Genetic Selection System for Cloning RNA-Binding Proteins from CDNA Libraries.” *Proceedings of the National Academy of Sciences* 98 (23): 12954–59. https://doi.org/10.1073/pnas.211439598.
26. Delaye, Luis, Alexander DeLuna, Antonio Lazcano, and Arturo Becerra. 2008. “The Origin of a Novel Gene through Overprinting in Escherichia Coli.” *BMC Evolutionary Biology* 8 (1): 31. https://doi.org/10.1186/1471-2148-8-31.
27. Digianantonio, Katherine M., and Michael H. Hecht. 2016. “A Protein Constructed de Novo Enables Cell Growth by Altering Gene Regulation.” *Proceedings of the National Academy of Sciences* 113 (9): 2400–2405. https://doi.org/10.1073/pnas.1600566113.
28. Dinger, Marcel E., Ken C. Pang, Tim R. Mercer, and John S. Mattick. 2008. “Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities.” Edited by Johanna McEntyre. *PLoS Computational Biology* 4 (11): e1000176. https://doi.org/10.1371/journal.pcbi.1000176.
29. Donnelly, Ann E, Grant S Murphy, Katherine M Digianantonio, and Michael H Hecht. 2018. “A de Novo Enzyme Catalyzes a Life-Sustaining Reaction in Escherichia Coli.” *Nature Chemical Biology* 14 (3): 253–55. https://doi.org/10.1038/nchembio.2550.
30. Doyle, Lindsey, Jazmine Hallinan, Jill Bolduc, Fabio Parmeggiani, David Baker, Barry L. Stoddard, and Philip Bradley. 2015. “Rational Design of α-Helical Tandem Repeat Proteins with Closed Architectures.” *Nature* 528 (7583): 585–88. https://doi.org/10.1038/nature16191.
31. Dryden, David T.F, Andrew R Thomson, and John H White. 2008. “How Much of Protein Sequence Space Has Been Explored by Life on Earth?” *Journal of The Royal Society Interface* 5 (25): 953–56. https://doi.org/10.1098/rsif.2008.0085.
32. Duan Yong and Kollman Peter A. 1998. “Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution.” *Science* 282 (5389): 740–44. https://doi.org/10.1126/science.282.5389.740.
33. Dydio, P., H. M. Key, A. Nazarenko, J. Y.-E. Rha, V. Seyedkazemi, D. S. Clark, and J. F. Hartwig. 2016. “An Artificial Metalloenzyme with the Kinetics of Native Enzymes.” *Science* 354 (6308): 102–6. https://doi.org/10.1126/science.aah4427.
34. Emmanuel, Megan A., Norman R. Greenberg, Daniel G. Oblinsky, and Todd K. Hyster. 2016. “Accessing Non-Natural Reactivity by Irradiating Nicotinamide-Dependent Enzymes with Light.” *Nature* 540 (7633): 414–17. https://doi.org/10.1038/nature20569.
35. Engler, Carola, Ramona Gruetzner, Romy Kandzia, and Sylvestre Marillonnet. 2009. “Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes.” Edited by Jean Peccoud. *PLoS ONE* 4 (5): e5553. https://doi.org/10.1371/journal.pone.0005553.
36. Engvall, Eva, and Peter Perlmann. 1972. “Enzyme-Linked Immunosorbent Assay, Elisa.” *The Journal of Immunology* 109 (1): 129.
37. Epps, Dennis E., Mark L. Wolfe, and Vince Groppi. 1994. “Characterization of the Steady-State and Dynamic Fluorescence Properties of the Potential-Sensitive Dye Bis-(1,3-Dibutylbarbituric Acid)Trimethine Oxonol (Dibac4(3)) in Model Systems and Cells.” *Chemistry and Physics of Lipids* 69 (2): 137–50. https://doi.org/10.1016/0009-3084(94)90035-3.
38. Fischlechner, Martin, Yolanda Schaerli, Mark F. Mohamed, Santosh Patil, Chris Abell, and Florian Hollfelder. 2014. “Evolution of Enzyme Catalysts Caged in Biomimetic Gel-Shell Beads.” *Nature Chemistry* 6 (9): 791–96. https://doi.org/10.1038/nchem.1996.
39. Fisher, Michael A., Kara L. McKinley, Luke H. Bradley, Sara R. Viola, and Michael H. Hecht. 2011. “De Novo Designed Proteins from a Library of Artificial Sequences Function in Escherichia Coli and Enable Cell Growth.” Edited by Mark Isalan. *PLoS ONE* 6 (1): e15364. https://doi.org/10.1371/journal.pone.0015364.
40. Foster, Patricia L. 2007. “Stress-Induced Mutagenesis in Bacteria.” *Critical Reviews in Biochemistry and Molecular Biology* 42 (5): 373–97. https://doi.org/10.1080/10409230701648494.

41. Galtier, Nicolas, Camille Roux, Marjolaine Rousselle, Jonathan Romiguier, Emeric Figuet, and Sylvain Gl. 2018. “Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion,” 12.
42. Go, Abigail, Seho Kim, Jean Baum, and Michael H. Hecht. 2008. “Structure and Dynamics of de Novo Proteins from a Designed Superfamily of 4-Helix Bundles.” *Protein Science* 17 (5): 821–32. https://doi.org/10.1110/ps.073377908.
43. Goodman, Morris. 1981. “Decoding the Pattern of Protein Evolution.” *Progress in Biophysics and Molecular Biology* 38: 105–64. https://doi.org/10.1016/0079-6107(81)90012-2.
44. Gunge, Norio. 1983. “YEAST DNA PLASMIDS.” *Annual Review of Microbiology* 37 (1): 253–76. https://doi.org/10.1146/annurev.mi.37.100183.001345.
45. Heames, Brennen, Filip Buchel, Margaux Aubel, Vyacheslav Tretyachenko, Andreas Lange, Erich Bornberg-Bauer, and Klara Hlouchova. 2022. “Experimental Characterisation of *de Novo* Proteins and Their Unevolved Random-Sequence Counterparts.” Preprint. Evolutionary Biology. https://doi.org/10.1101/2022.01.14.476368.
46. Heames, Brennen, Jonathan Schmitz, and Erich Bornberg-Bauer. 2020. “A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila.” *Journal of Molecular Evolution* 88 (4): 382–98. https://doi.org/10.1007/s00239-020-09939-z.
47. Heather, James M., and Benjamin Chain. 2016. “The Sequence of Sequencers: The History of Sequencing DNA.” *Genomics* 107 (1): 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.
48. Hecht, Michael H, Jane S Richardson, David C Richardson, and RIcHARD C Ogden. 1990. “Protein of Native-Like Sequence” 249: 8.
49. Hemm, Matthew R., Brian J. Paul, Juan Miranda-Ríos, Aixia Zhang, Nima Soltanzad, and Gisela Storz. 2010. “Small Stress Response Proteins in *Escherichia Coli* : Proteins Missed by Classical Proteomic Studies.” *Journal of Bacteriology* 192 (1): 46–58. https://doi.org/10.1128/JB.00872-09.
50. Ho, Steffan N., Henry D. Hunt, Robert M. Horton, Jeffrey K. Pullen, and Larry R. Pease. 1989. “Site-Directed Mutagenesis by Overlap Extension Using the Polymerase Chain Reaction.” *Gene* 77 (1): 51–59. https://doi.org/10.1016/0378-1119(89)90358-2.
51. Hoegler, Kenric J., and Michael H. Hecht. 2016. “A *de Novo* Protein Confers Copper Resistance in *E Scherichia Coli*.” *Protein Science* 25 (7): 1249–59. https://doi.org/10.1002/pro.2871.
52. Huang, Po-Ssu, Scott E. Boyken, and David Baker. 2016. “The Coming of Age of de Novo Protein Design.” *Nature* 537 (7620): 320–27. https://doi.org/10.1038/nature19946.
53. Il’ichev, AA, OO Minenkova, SI Tat’kov, NN Karpyshev, AM Eroshkin, VI Ofitserov, ZA Akimenko, VA Petrenko, and LS Sandakhchiev. 1990. “[The use of filamentous phage M13 in protein engineering].” *Mol Biol (Mosk)* 24 (2): 530–35.
54. Kamtekar, Satwik, Jarad M Schiffer, Huayu Xiong, Jennifer M Babik, and Michael H Hechtt. 1993. “Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids” 262: 6.
55. Karanicolas, John, Jacob E. Corn, Irwin Chen, Lukasz A. Joachimiak, Orly Dym, Sun H. Peck, Shira Albeck, et al. 2011. “A De Novo Protein Binding Pair By Computational Design and Directed Evolution.” *Molecular Cell* 42 (2): 250–60. https://doi.org/10.1016/j.molcel.2011.03.010.
56. Keefe, Anthony D., and Jack W. Szostak. 2001. “Functional Proteins from a Random-Sequence Library.” *Nature* 410 (6829): 715–18. https://doi.org/10.1038/35070613.
57. Keeling, Diane Marie, Patricia Garza, Charisse Michelle Nartey, and Anne-Ruxandra Carvunis. 2019. “The Meanings of ‘function’ in Biology and the Problematic Case of de Novo Gene Emergence.” *ELife* 8 (November): e47014. https://doi.org/10.7554/eLife.47014.
58. Key, Hanna M., Paweł Dydio, Douglas S. Clark, and John F. Hartwig. 2016. “Abiological Catalysis by Artificial Haem Proteins Containing Noble Metals in Place of Iron.” *Nature* 534 (7608): 534–37. https://doi.org/10.1038/nature17968.
59. Kim, Juhan, Jamie P Kershner, Yehor Novikov, Richard K Shoemaker, and Shelley D Copley. 2010. “Three Serendipitous Pathways in *E. Coli* Can Bypass a Block in Pyridoxal-5′-phosphate Synthesis.” *Molecular Systems Biology* 6 (1): 436. https://doi.org/10.1038/msb.2010.88.
60. Kintses, Balint, Christopher Hein, Mark F. Mohamed, Martin Fischlechner, Fabienne Courtois, Céline Lainé, and Florian Hollfelder. 2012. “Picoliter Cell Lysate Assays in Microfluidic Droplet Compartments for Directed Enzyme Evolution.” *Chemistry & Biology* 19 (8): 1001–9. https://doi.org/10.1016/j.chembiol.2012.06.009.
61. Knopp, Michael, Arianne M. Babina, Jónína S. Gudmundsdóttir, Martin V. Douglass, M. Stephen Trent, and Dan I. Andersson. 2021. “A Novel Type of Colistin Resistance Genes Selected from Random Sequence Space.” Edited by Carmen Buchrieser. *PLOS Genetics* 17 (1): e1009227. https://doi.org/10.1371/journal.pgen.1009227.
62. Knopp, Michael, Jonina S. Gudmundsdottir, Tobias Nilsson, Finja König, Omar Warsi, Fredrika Rajer, Pia Ädelroth, and Dan I. Andersson. 2019. “*De Novo* Emergence of Peptides That Confer Antibiotic Resistance.” Edited by Gerard D. Wright. *MBio* 10 (3): e00837-19. https://doi.org/10.1128/mBio.00837-19.

63. Koga, Nobuyasu, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, and David Baker. 2012. "Principles for Designing Ideal Protein Structures." *Nature* 491 (7423): 222–27. https://doi.org/10.1038/nature11600.
64. Kolodny, Rachel, Sergey Nepomnyachiy, Dan S Tawfik, and Nir Ben-Tal. 2021. "Bridging Themes: Short Protein Segments Found in Different Architectures." Edited by Julian Echave. *Molecular Biology and Evolution* 38 (6): 2191–2208. https://doi.org/10.1093/molbev/msab017.
65. Kültz, Dietmar. 2005. "MOLECULAR AND EVOLUTIONARY BASIS OF THE CELLULAR STRESS RESPONSE." *Annual Review of Physiology* 67 (1): 225–57. https://doi.org/10.1146/annurev.physiol.67.040403.103635.
66. Lassila, Jonathan Kyle, Heidi K. Privett, Benjamin D. Allen, and Stephen L. Mayo. 2006. "Combinatorial Methods for Small-Molecule Placement in Computational Enzyme Design." *Proceedings of the National Academy of Sciences* 103 (45): 16710–15. https://doi.org/10.1073/pnas.0607691103.
67. Lau, K F, and K A Dill. 1990. "Theory for Protein Mutability and Biogenesis." *Proceedings of the National Academy of Sciences* 87 (2): 638–42. https://doi.org/10.1073/pnas.87.2.638.
68. Lin, Yu-Ru, Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Amanda F. Clouser, Gaetano T. Montelione, and David Baker. 2015. "Control over Overall Shape and Size in de Novo Designed Proteins." *Proceedings of the National Academy of Sciences* 112 (40). https://doi.org/10.1073/pnas.1509508112.
69. Long, Manyuan, Esther Betrán, Kevin Thornton, and Wen Wang. 2003. "The Origin of New Genes: Glimpses from the Young and Old." *Nature Reviews Genetics* 4 (11): 865–75. https://doi.org/10.1038/nrg1204.
70. Ludwig, M. 2002. "Functional Evolution of Noncoding DNA." *Current Opinion in Genetics & Development* 12 (6): 634–39. https://doi.org/10.1016/S0959-437X(02)00355-6.
71. Macek, Boris, Karl Forchhammer, Julie Hardouin, Eilika Weber-Ban, Christophe Grangeasse, and Ivan Mijakovic. 2019. "Protein Post-Translational Modifications in Bacteria." *Nature Reviews Microbiology* 17 (11): 651–64. https://doi.org/10.1038/s41579-019-0243-0.
72. Malys, Naglis, Dau-Yin Chang, Richard G. Baumann, Dongmei Xie, and Lindsay W. Black. 2002. "A Bipartite Bacteriophage T4 SOC and HOC Randomized Peptide Display Library: Detection and Analysis of Phage T4 Terminase (Gp17) and Late σ Factor (Gp55) Interaction." *Journal of Molecular Biology* 319 (2): 289–304. https://doi.org/10.1016/S0022-2836(02)00298-X.
73. Marks, James D., Hennie R. Hoogenboom, Timothy P. Bonnert, John McCafferty, Andrew D. Griffiths, and Greg Winter. 1991. "By-Passing Immunization." *Journal of Molecular Biology* 222 (3): 581–97. https://doi.org/10.1016/0022-2836(91)90498-U.
74. McCafferty, John, Andrew D. Griffiths, Greg Winter, and David J. Chiswell. 1990. "Phage Antibodies: Filamentous Phage Displaying Antibody Variable Domains." *Nature* 348 (6301): 552–54. https://doi.org/10.1038/348552a0.
75. Moffet, David A., Laura K. Certain, Allison J. Smith, Adam J. Kessel, Katharine A. Beckwith, and Michael H. Hecht. 2000. "Peroxidase Activity in Heme Proteins Derived from a Designed Combinatorial Library." *Journal of the American Chemical Society* 122 (31): 7612–13. https://doi.org/10.1021/ja001198q.
76. Nagano, Nozomi, Christine A Orengo, and Janet M Thornton. 2002. "One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on Their Sequences, Structures and Functions." *Journal of Molecular Biology* 321 (5): 741–65. https://doi.org/10.1016/S0022-2836(02)00649-6.
77. Neme, Rafik, Cristina Amador, Burcin Yildirim, Ellen McConnell, and Diethard Tautz. 2017. "Random Sequences Are an Abundant Source of Bioactive RNAs or Peptides." *Nature Ecology & Evolution* 1 (6): 0127. https://doi.org/10.1038/s41559-017-0127.
78. Neme, Rafik, and Diethard Tautz. 2013. "Phylogenetic Patterns of Emergence of New Genes Support a Model of Frequent de Novo Evolution." *BMC Genomics* 14 (1): 117. https://doi.org/10.1186/1471-2164-14-117.
79. Nirenberg, M. 2004. "Historical Review: Deciphering the Genetic Code – a Personal Account." *Trends in Biochemical Sciences* 29 (1): 46–54. https://doi.org/10.1016/j.tibs.2003.11.009.
80. Obexer, Richard, Alexei Godina, Xavier Garrabou, Peer R. E. Mittl, David Baker, Andrew D. Griffiths, and Donald Hilvert. 2017. "Emergence of a Catalytic Tetrad during Evolution of a Highly Active Artificial Aldolase." *Nature Chemistry* 9 (1): 50–56. https://doi.org/10.1038/nchem.2596.
81. Pál, Csaba, Balázs Papp, and Martin J. Lercher. 2006. "An Integrated View of Protein Evolution." *Nature Reviews Genetics* 7 (5): 337–48. https://doi.org/10.1038/nrg1838.
82. Park, Hee-Sung, Sung-Hun Nam, Jin Kak Lee, Chang No Yoon, Bengt Mannervik, Stephen J. Benkovic, and Hak-Sung Kim. 2006. "Design and Evolution of New Catalytic Activity with an Existing Protein Scaffold." *Science* 311 (5760): 535–38. https://doi.org/10.1126/science.1118953.
83. Patel, Shona C., Luke H. Bradley, Sayuri P. Jinadasa, and Michael H. Hecht. 2009. "Cofactor Binding and Enzymatic Activity in an Unevolved Superfamily of *de Novo* Designed 4-Helix Bundle Proteins:

Binding and Activity of *De Novo* Designed Proteins." *Protein Science* 18 (7): 1388–1400. https://doi.org/10.1002/pro.147.

84. Patrick, W. M., E. M. Quandt, D. B. Swartzlander, and I. Matsumura. 2007. "Multicopy Suppression Underpins Metabolic Evolvability." *Molecular Biology and Evolution* 24 (12): 2716–22. https://doi.org/10.1093/molbev/msm204.
85. Pinheiro, Vitor B., Alexander I. Taylor, Christopher Cozens, Mikhail Abramov, Marleen Renders, Su Zhang, John C. Chaput, et al. 2012. "Synthetic Genetic Polymers Capable of Heredity and Evolution." *Science* 336 (6079): 341–44. https://doi.org/10.1126/science.1217622.
86. Privett, Heidi K., Gert Kiss, Toni M. Lee, Rebecca Blomberg, Roberto A. Chica, Leonard M. Thomas, Donald Hilvert, Kendall N. Houk, and Stephen L. Mayo. 2012. "Iterative Approach to Computational Enzyme Design." *Proceedings of the National Academy of Sciences* 109 (10): 3790–95. https://doi.org/10.1073/pnas.1118082108.
87. Qi, Huan, Haiqin Lu, Hua-Ji Qiu, Valery Petrenko, and Aihua Liu. 2012. "Phagemid Vectors for Phage Display: Properties, Characteristics and Construction." *Journal of Molecular Biology* 417 (3): 129–43. https://doi.org/10.1016/j.jmb.2012.01.038.
88. Regan, Lynne, and William F. DeGrado. 1988. "Characterization of a Helical Protein Designed from First Principles." *Science* 241 (4868): 976–78. https://doi.org/10.1126/science.3043666.
89. Roberts, Richard W., and Jack W. Szostak. 1997. "RNA-Peptide Fusions for the *in Vitro* Selection of Peptides and Proteins." *Proceedings of the National Academy of Sciences* 94 (23): 12297–302. https://doi.org/10.1073/pnas.94.23.12297.
90. Robertson, Dan E., Ramy S. Farid, Christopher C. Moser, Jeffrey L. Urbauer, Stephen E. Mulholland, Ravindernath Pidikiti, James D. Lear, A. Joshua Wand, William F. DeGrado, and P. Leslie Dutton. 1994. "Design and Synthesis of Multi-Haem Proteins." *Nature* 368 (6470): 425–32. https://doi.org/10.1038/368425a0.
91. Rocklin, Gabriel J., Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, et al. 2017. "Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing." *Science* 357 (6347): 168–75. https://doi.org/10.1126/science.aan0693.
92. Rohl, Carol A., Charlie E.M. Strauss, Kira M.S. Misura, and David Baker. 2004. "Protein Structure Prediction Using Rosetta." In *Methods in Enzymology*, 383:66–93. Elsevier. https://doi.org/10.1016/S0076-6879(04)83004-0.
93. Ruiz-Orera, Jorge, Pol Verdaguer-Grau, José Luis Villanueva-Cañas, Xavier Messeguer, and M. Mar Albà. 2018. "Translation of Neutrally Evolving Peptides Provides a Basis for de Novo Gene Evolution." *Nature Ecology & Evolution* 2 (5): 890–96. https://doi.org/10.1038/s41559-018-0506-6.
94. Ruiz-Orera, Jorge, José Luis Villanueva-Cañas, and M. Mar Albà. 2020. "Evolution of New Proteins from Translated SORFs in Long Non-Coding RNAs." *Experimental Cell Research* 391 (1): 111940. https://doi.org/10.1016/j.yexcr.2020.111940.
95. Sanger, F., and H. Tuppy. 1951. "The Amino-Acid Sequence in the Phenylalanyl Chain of Insulin. 1. The Identification of Lower Peptides from Partial Hydrolysates." *Biochemical Journal* 49 (4): 463–81. https://doi.org/10.1042/bj0490463.
96. Schmitz, Jonathan F., Frédéric J. J. Chain, and Erich Bornberg-Bauer. 2020. "Evolution of Novel Genes in Three-Spined Stickleback Populations." *Heredity* 125 (1–2): 50–59. https://doi.org/10.1038/s41437-020-0319-7.
97. Seebeck, Florian P., and Donald Hilvert. 2003. "Conversion of a PLP-Dependent Racemase into an Aldolase by a Single Active Site Mutation." *Journal of the American Chemical Society* 125 (34): 10158–59. https://doi.org/10.1021/ja036707d.
98. Shabalina, Svetlana A, and Nikolay A Spiridonov. 2004. "The Mammalian Transcriptome and the Function of Non-Coding DNA Sequences." *Genome Biology*, 8.
99. Shen, Binzhang. 2002. "PCR Approaches to DNA Mutagenesis and Recombination: An Overview." In *PCR Cloning Protocols*, by Bing-Yuan Chen and Harry W. Janes, 192:167–74. New Jersey: Humana Press. https://doi.org/10.1385/1-59259-177-9:167.
100. Smith, Betsy A., Ann E. Mularz, and Michael H. Hecht. 2015. "Divergent Evolution of a Bifunctional *de Novo* Protein: Divergent Evolution of Bifunctional De Novo Protein." *Protein Science* 24 (2): 246–52. https://doi.org/10.1002/pro.2611.
101. Smith, George P. 1985. "Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface." *Science* 228 (4705): 1315–17. https://doi.org/10.1126/science.4001944.
102. Smith, George P., and Valery A. Petrenko. 1997. "Phage Display." *Chemical Reviews* 97 (2): 391–410. https://doi.org/10.1021/cr960065d.
103. Soo, Valerie W. C., Paulina Hanson-Manful, and Wayne M. Patrick. 2011. "Artificial Gene Amplification Reveals an Abundance of Promiscuous Resistance Determinants in *Escherichia Coli*." *Proceedings of the National Academy of Sciences* 108 (4): 1484–89. https://doi.org/10.1073/pnas.1012108108.
104. Stemmer, W P. 1994. "DNA Shuffling by Random Fragmentation and Reassembly: In Vitro Recombination for Molecular Evolution." *Proceedings of the National Academy of Sciences* 91 (22): 10747–51. https://doi.org/10.1073/pnas.91.22.10747.

105.Stemmer, Willem P.C., Andreas Crameri, Kim D. Ha, Thomas M. Brennan, and Herbert L. Heyneker. 1995. “Single-Step Assembly of a Gene and Entire Plasmid from Large Numbers of Oligodeoxyribonucleotides.” *Gene* 164 (1): 49–53. https://doi.org/10.1016/0378-1119(95)00511-4.

106.Stepanov, V. G., and G. E. Fox. 2007. “Stress-Driven In Vivo Selection of a Functional Mini-Gene from a Randomized DNA Library Expressing Combinatorial Peptides in Escherichia Coli.” *Molecular Biology and Evolution* 24 (7): 1480–91. https://doi.org/10.1093/molbev/msm067.

107.Storz, Gisela, Yuri I. Wolf, and Kumaran S. Ramamurthi. 2014. “Small Proteins Can No Longer Be Ignored.” *Annual Review of Biochemistry* 83 (1): 753–77. https://doi.org/10.1146/annurev-biochem-070611-102400.

108.Szybalskia, Waclaw. n.d. “Class-LlS Restriction Enzymes- a Review,” 1.

109.Tanaka, Junko, Nobuhide Doi, Hideaki Takashima, and Hiroshi Yanagawa. 2010. “Comparative Characterization of Random-Sequence Proteins Consisting of 5, 12, and 20 Kinds of Amino Acids: Random-Sequence Proteins with Limited Alphabets.” *Protein Science* 19 (4): 786–95. https://doi.org/10.1002/pro.358.

110.Tautz, Diethard. 2014. “The Discovery of De Novo Gene Evolution.” *Perspectives in Biology and Medicine* 57 (1): 149–61. https://doi.org/10.1353/pbm.2014.0006.

111.Teh, Shia-Yen, Robert Lin, Lung-Hsin Hung, and Abraham P. Lee. 2008. “Droplet Microfluidics.” *Lab on a Chip* 8 (2): 198. https://doi.org/10.1039/b715524g.

112.Terpe, Kay. 2006. “Overview of Bacterial Expression Systems for Heterologous Protein Production: From Molecular and Biochemical Fundamentals to Commercial Systems.” *Applied Microbiology and Biotechnology* 72 (2): 211–22. https://doi.org/10.1007/s00253-006-0465-8.

113.Thomson, Andrew R., Christopher W. Wood, Antony J. Burton, Gail J. Bartlett, Richard B. Sessions, R. Leo Brady, and Derek N. Woolfson. 2014. “Computational Design of Water-Soluble α-Helical Barrels.” *Science* 346 (6208): 485–88. https://doi.org/10.1126/science.1257452.

114.Tretyachenko, Vyacheslav, Jiří Vymětal, Lucie Bednárová, Vladimír Kopecký, Kateřina Hofbauerová, Helena Jindrová, Martin Hubálek, et al. 2017. “Random Protein Sequences Can Form Defined Secondary Structures and Are Well-Tolerated in Vivo.” *Scientific Reports* 7 (1): 15449. https://doi.org/10.1038/s41598-017-15635-8.

115.Upadhyay, Arun K., Aruna Murmu, Anupam Singh, and Amulya K. Panda. 2012. “Kinetics of Inclusion Body Formation and Its Correlation with the Characteristics of Protein Aggregates in Escherichia Coli.” Edited by Christophe Herman. *PLoS ONE* 7 (3): e33951. https://doi.org/10.1371/journal.pone.0033951.

116.Wang, Weixun, and Michael H Hecht. n.d. “Rationally Designed Mutations Convert de Novo Amyloid-like Fibrils into Monomeric,” 6.

117.West, Michael W., Weixun Wang, Jennifer Patterson, Joseph D. Mancias, James R. Beasley, and Michael H. Hecht. 1999. “*De Novo* Amyloid Proteins from Designed Combinatorial Libraries.” *Proceedings of the National Academy of Sciences* 96 (20): 11211–16. https://doi.org/10.1073/pnas.96.20.11211.

118.Wilhelm, Brian T., and Josette-Renée Landry. 2009. “RNA-Seq—Quantitative Measurement of Expression through Massively Parallel RNA-Sequencing.” *Methods* 48 (3): 249–57. https://doi.org/10.1016/j.ymeth.2009.03.016.

119.Winter, Greg, Andrew D Griffiths, Robert E Hawkins, and Hennie R Hoogenboom. n.d. “Making Antibodies by Phage Display Technology,” 23.

120.Wu, Dong-Dong, and Ya-Ping Zhang. 2013. “Evolution and Function of de Novo Originated Genes.” *Molecular Phylogenetics and Evolution* 67 (2): 541–45. https://doi.org/10.1016/j.ympev.2013.02.013.

121.Xu, Guofeng, Weixun Wang, John T Groves, and Michael H Hecht. n.d. “Self-Assembled Monolayers from a Designed Combinatorial Library of de Novo,” 6.

122.Yuan, Ling, Itzhak Kurek, James English, and Robert Keenan. 2005. “Laboratory-Directed Protein Evolution.” *Microbiology and Molecular Biology Reviews* 69 (3): 373–92. https://doi.org/10.1128/MMBR.69.3.373-392.2005.

123.Zastrow, Melissa L., Anna F. A. Peacock, Jeanne A. Stuckey, and Vincent L. Pecoraro. 2012. “Hydrolytic Catalysis and Structural Stabilization in a Designed Metalloprotein.” *Nature Chemistry* 4 (2): 118–23. https://doi.org/10.1038/nchem.1201.

124.Zhang, Li, Yan Ren, Tao Yang, Guangwei Li, Jianhai Chen, Andrea R. Gschwend, Yeisoo Yu, et al. 2019. “Rapid Evolution of Protein Diversity by de Novo Origination in Oryza.” *Nature Ecology & Evolution* 3 (4): 679–90. https://doi.org/10.1038/s41559-019-0822-5.

125.Zhou, X. 2004. “Microfluidic PicoArray Synthesis of Oligodeoxynucleotides and Simultaneous Assembling of Multiple DNA Sequences.” *Nucleic Acids Research* 32 (18): 5409–17. https://doi.org/10.1093/nar/gkh879.