Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

De novo proteins from random sequences through *in vitro* evolution

Cher Ling Tong[1,2], Kun-Hwa Lee[1,2] and Burckhard Seelig[1,2]

Natural proteins are the result of billions of years of evolution. The earliest predecessors of today's proteins are believed to have emerged from random polypeptides. While we have no means to determine how this process exactly happened, there is great interest in understanding how it reasonably could have happened. We are reviewing how researchers have utilized *in vitro* selection and molecular evolution methods to investigate plausible scenarios for the emergence of early functional proteins. The studies range from analyzing general properties and structural features of unevolved random polypeptides to isolating *de novo* proteins with specific functions from synthetic randomized sequence libraries or generating novel proteins by combining evolution with rational design. While the results are exciting, more work is needed to fully unravel the mechanisms that seeded protein-dominated biology.

Addresses
[1] Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN, USA
[2] BioTechnology Institute, University of Minnesota, St. Paul, MN, USA

Corresponding author: Seelig, Burckhard (seelig@umn.edu)

Current Opinion in Structural Biology 2021, **68**:129–134

This review comes from a themed issue on **Sequences and topology**

Edited by **Nir Ben-Tal** and **Andrei Lupas**

https://doi.org/10.1016/j.sbi.2020.12.014

Introduction

Biology as we know it today has evolved over eons. The vast diversity of proteins in nature is simply astounding, comprising complex structures that enable an impressive variety of functions. By applying the principles of Darwinian evolution, scientists have learned how to trace back sophisticated modern proteins to their likely simpler ancestors [1–4]. However, the origin of those simple ancient proteins is less well understood.

Could the earliest functional proteins have emerged by chance from random polypeptides? Unfortunately, finding a *specific* sequence by chance for even just a small 100 residue long protein is practically impossible. The number of amino acid combinations possible for such a protein is literally astronomical: 20^{100} is greater than the estimated number of atoms in the universe. Nature could not have exhausted the entire sequence space of all possible combinations and yet is populated with millions of well-structured, functional proteins. Therefore, the important question is not how to find a *specific* protein sequence known to biology but rather how to find *any* sequence with properties useful to biology. In other words — how are functional proteins distributed across the vast sequence space comprising all possible proteins? (Figure 1)

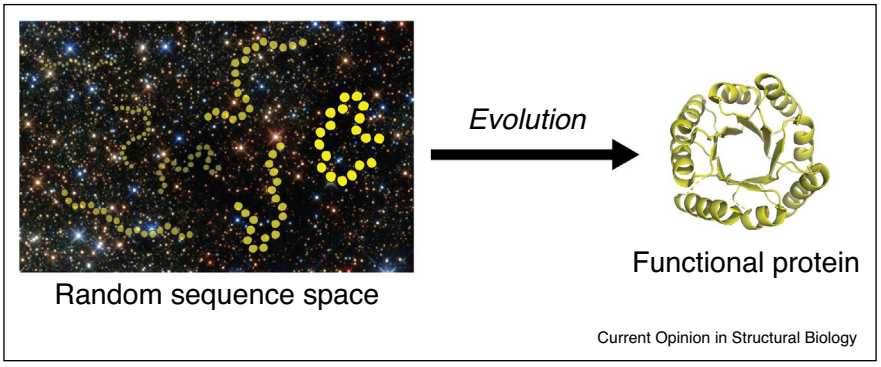
We define such a protein that has just been brought into existence from protein sequence space without an evolutionary history as a *de novo* protein [5]. The search of sequence space for proteins with useful properties can be performed through screening or selection methods. The methods range from manually screening tens of protein variants to automated screening of millions. Even higher throughput can be achieved with *in vivo* or *in vitro* selections such as cell-based selections, phage display, ribosome display, or mRNA display that can identify proteins from mixtures of up to 10^{13} different variants [6•,7]. The selection techniques are commonly expanded to directed evolution strategies by including a mutagenesis step that can further improve the proteins' initial properties.

In this review, we focus on *de novo* proteins identified through screening, selection, and directed evolution. The search of random sequence space for proteins with distinct properties is motivated by the desire to investigate the origin of protein-based life and can provide unique proteins for applications in medicine or biotechnology.

General properties of random proteins

Natural evolution has yielded proteins that share several fundamental characteristics: (i) They fold into mostly well-defined 3D-structures built from α-helices, β-strands, or a mixture of those. (ii) They often bind to other proteins or non-protein molecules through surface contacts. (iii) Many proteins act upon their binding partners through specific interactions to induce conformational changes or even catalyze a chemical reaction. To investigate the idea that the earliest ancestors of modern proteins emerged by chance from random polypeptide sequences, it is crucial to first understand the general biophysical and biochemical properties of such unevolved random sequences. Do unevolved *de novo* proteins have similar properties as naturally evolved proteins? How

Figure 1



Emergence of *de novo* proteins from protein sequence space that comprises all possible sequences. (Background image: Hubble space telescope's view of the milky way (NASA Image and Video Library; URL: https://images.nasa.gov/details-GSFC_20171208_Archive_e000717). Protein structure: Protein Data Bank code 5BVL).

likely are random sequences to fold into a 3D-structure, which is considered a prerequisite for a functional protein?

One of the earliest experimental demonstrations showing that folded proteins are common among random amino acid sequences was published by the Sauer group [8]. A library encoding 80-to100-residue random sequences consisting of only the amino acids Q, L, and R was transformed into *Escherichia coli*. 5% of the random sequences could be expressed in *E. coli*. These expressed proteins possessed α-helical content, and their structures showed cooperative thermal denaturation transitions and resistance to protease degradation [9]. A different library of 100-residue random sequences made up mainly of the amino acids G, A, V, D, and E was found to be highly soluble, yet the eight arbitrarily chosen clones did not possess any secondary structure [10]. When all 20 amino acids were used to prepare a random protein library of 141 amino acidsin length, 20% of the successfully expressed proteins were also found to be soluble [11]. These findings suggest that a large fraction of unevolved random sequences could be soluble like naturally evolved proteins. Even a protein library as short as only 50 residues made from all 20 amino acids was reported to contain about 20% folded protein, as shown by resistance to protease digestion and circular dichroism [12]. Various biophysical techniques were used to screen a library of 71-residue random proteins with an amino acid composition similar to natural proteins [13]. The unevolved random proteins possessed secondary structure content and unfolding behavior similar to natural proteins, suggesting that distinct structural properties of proteins were not necessarily evolved by natural selection but are intrinsic to polypeptides.

The modern 20 amino acid alphabet originated from a much smaller early genetic code of only a few prebiotic amino acids. A wealth of information established a likely chronological order of amino acids entering the genetic code [14]. An extensive study by Newton et al. explored the solubility and secondary structure content of random proteins made from the likely earliest 5, 9, 16, and the modern 20 amino acids [15••]. This work compared unevolved proteins at various stages of the genetic code evolution under identical conditions. Unlike previous attempts [10,16], a different library synthesis method allowed for the generation of 80-residue random proteins of the desired ancient amino acid compositions. The study found that between one-third and two-thirds of the library members could be expressed in *E. coli*. Surprisingly, the majority (66–86%) of those expressed random proteins from the three reduced alphabet libraries was highly soluble. The fraction of highly soluble clones increased from the 5 to 9 to 16 amino acids alphabets. In contrast, the 20 amino acid library yielded only one highly soluble protein out of 16 proteins that could be expressed (6%). Most of the soluble variants from all four librarieswere shown to possess secondary structure content by circular dichroism in the presence of structure-promoting trifluoroethanol. For one clone from the 16 amino acid alphabet, even tertiary structure was detected. [15••].

An interesting computational approach was taken by the Hlouchová group to compare random polypeptides with biological proteins [17•]. They applied bioinformatic tools to predict secondary structure content, degree of disorder, and aggregation state of a library of 100-residue random proteins and then experimentally characterized several of these proteins. The study showed that the secondary structure content and general physicochemical properties were surprisingly similarly for proteinswith random sequences and those from nature. However, the unevolved random polypeptides with the least secondary structure content, yet highest disorder, were found to be most soluble as they were less likely to aggregate. This is not the case for natural proteins.

The combined studies above demonstrate that random sequence space contains a considerable fraction of proteins that are soluble and have the ability to fold to some degree.

Functional proteins identified from random sequence space

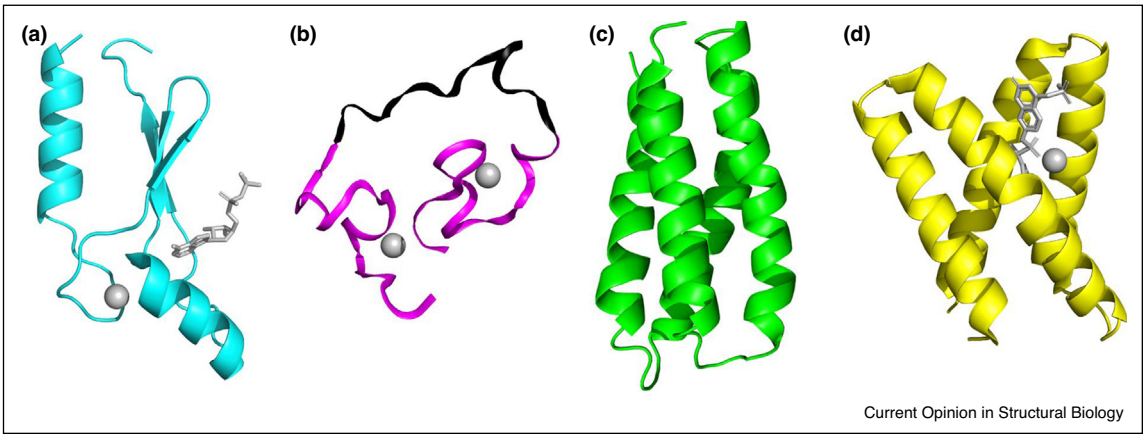
While the ability of a protein to fold is usually a precondition for function, only the function itself is what renders a *de novo* protein potentially useful to biology and, therefore, selectable through natural evolution. Several studies demonstrated that functional polypeptides can be isolated from random sequence libraries. As proteins that are structured *and* functional are likely more rare than just structured proteins, higher throughput selection and evolution methods were necessary to find them.

An *in vivo* selection strategy was used to interrogate several random peptide libraries of 10^8 variants that were 10–50 residues in length for their ability to increase antibiotic resistance in bacteria [18••]. Three active peptides of only 22–25 residues were isolated and found to be composed of mostly hydrophobic amino acids. These active peptides behaved similarly to natural transmembrane proteins. They have a short α-helical structure and insert into the cell membrane, thereby decreasing the membrane potential and thus the aminoglycoside uptake, leading to an up to 48-fold increase in antibiotic resistance.

In a landmark study, Keefe and Szostak isolated *de novo* ATP-binding proteins from random sequences [19]. Using rounds of *in vitro* selection and evolution by mRNA display [6•,20] and starting from a library of 6×10^{12} polypeptides of 80 amino acid in length, they identified four unrelated proteins that bound ATP with high affinity and specificity. The 3D structure of the tightest ATP binder solved by X-ray crystallography [21] and NMR [22] consists of two α-helices and three antiparallel β-strands that are further stabilized by coordination to zinc (Figure 2a). The ATP-binding pocket shows similarities to natural adenine binding proteins [23•], with the adenine bound by aromatic stacking interactions and hydrogen bonds. The ribose moiety and phosphate groups interact with polar side chains through additional hydrogen bonds. While the overall structure of the artificial protein constitutes a novel fold that does not match known proteins, the part of the protein structure that coordinates the zinc ion does have a natural analog in the treble clef finger [24]. This finding is remarkable because it is one of the first experimental proofs for the existence of a true structural analog. In a different *in vitro* selection project using the closely related cDNA display method, an ATP-binding protein was isolated from a random protein library of 108 residues consisting of only the 15 likely early amino acids [25]. Preliminary characterization indicated that the protein possesses α-helix and random coil content and potentially ATP hydrolysis activity.

The *in vitro* selection technology mRNA display was modified to enable the isolation of *de novo* enzymes. From a library of 4×10^{12} randomized polypeptides, proteins were found to catalyze an RNA ligation reaction for which there are no enzymes known in nature [26]. The enzymes exhibited rate enhancements of more than two million-fold. Unlike the examples above, the starting library was based on a small stable protein domain of 77 amino acids,

Figure 2



3-D structures of *de novo* proteins from random sequences or partly designed randomized sequence libraries. **(a)** Artificial ATP-binding protein [21] (Protein Data Bank code 1UW1). Bound ADP is shown in gray and zinc as a gray sphere. **(b)** Artificial RNA ligase enzyme [27] (Protein Data Bank code 2LZE). The two highly structured regions (purple) frame the more dynamic loop (black). Flexible termini were omitted for clarity, and zinc ions are shown as gray spheres. **(c)** Four-helix bundle protein from a binary-pattern random library [30] (Protein Data Bank code 2JUA). (D) Engineered metalloesterase complexed with a phosphonate transition state analog (gray) [31••] (Protein Data Bank code 5OD1). Zinc is shown as a gray sphere.

of which about a quarter of residues was completely randomized. However, during the selection and evolution process, the original structure was lost, and the protein instead adopted an entirely different fold [27]. NMR spectroscopy of a thermostable enzyme variant [28] revealed two highly structured regions coordinated by two zinc ions and embedded in more dynamic regions (Figure 2b). Surprisingly, secondary structure elements like α-helices and β-strands are essentially absent from this laboratory-evolved enzyme. While the protein is stably folded and displays cooperative unfolding at 72°C, the structure has increased flexibility compared to natural proteins. This artificial enzyme also has practical applications for the selective labeling of certain classes of RNA [29].

Despite these exciting discoveries of functional proteins with novel structures, the number of examples for truly *de novo* proteins that emerged from random sequence space solely through laboratory selection and evolution is still extremely small. This dearth is likely due to the perceived difficulty of such a project. To prove that the origin of proteins from random is a reliable scenario, more examples of *de novo* proteins with a wider variety of functions need to be identified, which is still no easy task. In contrast, there are more examples of *de novo* proteins generated through a combination of some initial rational design with subsequent directed evolution as described in the following section.

Combining rational design with directed evolution for novel proteins

While the main focus of this review is the emergence of function from random sequence space, the clever use of rational design aspects provides additional insights into the emergence of functional proteins. The rational design of new protein structures has made substantial progress [32]. However, the design of specific functions like catalysis is still challenging [33]. Fortunately, some studies have shown that the subsequent *in vitro* evolution of designed proteins can identify variants with a substantial improvement of their respective functions to levels that have not been attainable by rational design alone. We will discuss a few select examples but are not able to review this area comprehensively here.

The Hecht group designed a semi-random library of polypeptides with a binary pattern of polar and non-polar side chains that resulted in proteins prone to fold into four-helix bundle structures, yet without any designed functions [34] (Figure 2c). About 10^6 variants of this 102-residue-long library were transformed into several strains of *E. coli* that were conditional auxotrophs due to a single-gene knockout. Four strains were rescued by the overexpression of specific library variants, demonstrating that *de novo* four-helix bundle proteins are capable of affording cell growth. The same group subsequently showed that these *de novo* proteins enabled survival of the knockout strains through different mechanisms. Two of the artificial proteins affected the upregulation of endogenous enzymes with an enzymatic activity similar to the deleted enzyme [35,36]. In contrast, a third selected artificial protein instead was shown to act as a bona fide enzyme that replaced the function of the deleted naturally evolved enzyme [37••]. The rescue activity of the initially isolated protein was improved by several rounds of directed evolution [38]. The enzyme catalyzed the hydrolysis of the siderophore ferric enterobactin and thereby enabled cells to assimilate iron. Remarkably, the *de novo* enzyme used a completely different sequence, structure, and mechanism compared to the natural *E. coli* enzyme, which it replaced. Other proteins originating from the semi-random four-helix bundle library design performed several additional functions [39], including an ATPase activity [40•]. About 1100 library members were first screened for fatty acid ester hydrolysis. One of the five most active variants was then found to also hydrolyze the terminal phosphodiester bond in ATP, as well as GTP, CTP, and UTP. As the rate acceleration of about 100-fold above the uncatalyzed reaction is rather low, it will be interesting to see how much this *de novo* activity can be improved by future directed evolution.

The tremendous power of directed laboratory evolution to turn barely functional primitive proteins into highly functional enzymes has already been demonstrated for *de novo* enzymes originally generated not from random sequence but by rational design. This success suggests that low-level functionalities of random sequence origin might also be evolvable to proficient levels. Therefore, we will describe two examples in more detail.

The first example from the Hilvert group took a semi-rational approach to create a highly active and enantiospecific *de novo* metalloenzyme. They started from a computationally designed homodimer of a 46-residue peptide that coordinated zinc at the dimer interface [41]. The zinc was found to serendipitously catalyze the hydrolysis of an ester bond weakly [42]. The two peptides of the dimer were fused, and the protein was subjected to several rounds of mutagenesis and screening, eventually yielding a variant with >10 000-fold higher activity compared to the originally designed progenitor [31••] (Figure 2d). This level of activity is similar to typical naturally evolved enzymes, which is still rarely achieved for *de novo* enzymes. Compared to the original design, the protein structure had changed to some degree during the course of laboratory evolution, which included the replacement of one of the zinc-coordinating residues. This project also demonstrates how a simple peptide can be the starting point for an evolutionary path to a highly active globular metalloenzyme through metal-mediated assembly, domain fusion, and diversification.

The second example for the efficient directed laboratory evolution involves a *de novo* aldolase enzyme [43•]. However, in contrast to the cases above, only the activity of this enzyme was designed *de novo*, while its structure was that of a naturally evolved protein. The original aldolase was computationally designed [44] and subsequently improved through directed evolution. Interestingly, a variant with a >4400-fold improved activity had abandoned the rationally designed catalytic apparatus and instead used a new catalytic residue and an extensively remodeled active site [45]. Further evolution using the high-throughput method of droplet-based microfluidic screening increased the aldolase activity by another 30-fold. The resulting enzyme had a rate enhancement of $>10^9$-fold, which is comparable to the efficiency of an average natural enzyme [43•]. The accumulated mutations led to a sophisticated active site with a catalytic tetrad — a catalytic feature also found in natural enzymes. This general strategy of installing computationally modeled catalytic residues into a suitable natural protein cavity has been successful also for several other reactions and has been reviewed elsewhere [46].

Conclusion

Natural evolution has only sampled a minute fraction of the vast sequence space of all possible proteins. Therefore, we asked at the beginning of this article whether simple, functional proteins could be found from this vast space by random chance. The work reviewed here demonstrates that random sequences are indeed a viable source for *de novo* proteins to emerge. Several studies have convincingly shown that a substantial fraction of unevolved random proteins already possesses natural protein-like properties such as solubility, foldability, and thermostability. In contrast, only very few studies have been able to take the next step and prove that sequence space is sufficiently populated with not just structured but also functional proteins that can be identified by high-throughput selection methods. One of the challenges of this approach is the limitation in throughput of the existing selection technologies, currently capped at libraries with about 10^{13} proteins. For comparison, it was shown that proteins that tightly bind ATP occur about 1 in 10^{11} random sequences [19]. Therefore, improvements to the throughput of selection methods would increase chances for identifying additional *de novo* proteins from random for a wider range of functions. To date, the structure of this ATP binder remains the only example of a *de novo* globular protein that was isolated from naïve random sequence space. Without additional examples, it is premature to draw general conclusions about the distribution of functions in random protein sequence space. Furthermore, there is still no structure of a genuine enzyme isolated from random sequences.

In summary, whether your main interest lies in unraveling the origin of proteins or in engineering efficient novel enzymes, the articles reviewed here demonstrate that directed molecular evolution is tremendously helpful to achieve your goals. To paraphrase our initial query: Can we re-enact plausible scenarios for the emergence of early functional proteins from random sequence space through laboratory evolution? Yes, we can.

Conflict of interest statement

Nothing declared.

Acknowledgments

We thank Stephen P. Miller and Romas J. Kazlauskas for critical reading of the manuscript. We gratefully acknowledge financial support from the US National Aeronautics and Space Administration (NASA)80NSSC18K1277, the US National Institutes of Health (NIH) (GM108703), Human Frontier Science Program (RGP0041), and the Simons Collaboration on the Origins of Life (340762).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Randall RN, Radford CE, Roof KA, Natarajan DK, Gaucher EA: **An experimental phylogeny to benchmark ancestral sequence reconstruction**. *Nat Commun* 2016, **7**:12847.

2. Merkl R, Sterner R: **Ancestral protein reconstruction: techniques and applications**. *Biol Chem* 2016, **397**:1-21.

3. Alva V, Soding J, Lupas AN: **A vocabulary of ancient peptides at the origin of folded proteins**. *eLife* 2015, **4**.

4. Longo LM, Despotovic D, Weil-Ktorza O, Walker MJ, Jablonska J, Fridmann-Sirkis Y, Varani G, Metanis N, Tawfik DS: **Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion**. *Proc Natl Acad Sci U S A* 2020, **117**:15731-15739.

5. Golynskiy MV, Seelig B: ***De novo* enzymes: from computational design to mRNA display**. *Trends Biotechnol* 2010, **28**:340-345.

6. • Newton MS, Cabezas-Perusse Y, Tong CL, Seelig B: ***In vitro* selection of peptides and proteins-advantages of mRNA display**. *ACS Synth Biol* 2020, **9**:181-190

The mRNA display technology is compared to other selection methods such as phage display, cell-surface display, or ribosome display. The review explains how mRNA display outperforms other methods, especially on throughput and the large degree of control on experimental conditions.

7. Lane MD, Seelig B: **Advances in the directed evolution of proteins**. *Curr Opin Chem Biol* 2014, **22**:129-136.

8. Davidson AR, Sauer RT: **Folded proteins occur frequently in libraries of random amino acid sequences**. *Proc Natl Acad Sci U S A* 1994, **91**:2146-2150.

9. Davidson AR, Lumb KJ, Sauer RT: **Cooperatively folded proteins in random sequence libraries**. *Nat Struct Biol* 1995, **2**:856-864.

10. Doi N, Kakukawa K, Oishi Y, Yanagawa H: **High solubility of random-sequence proteins consisting of five kinds of primitive amino acids**. *Protein Eng Des Sel* 2005, **18**:279-284.

11. Prijambada ID, Yomo T, Tanaka F, Kawama T, Yamamoto K, Hasegawa A, Shima Y, Negoro S, Urabe I: **Solubility of artificial proteins with random sequences**. *FEBS Lett* 1996, **382**:21-25.

12. Chiarabelli C, Vrijbloed JW, Thomas RM, Luisi PL: **Investigation of *de novo* totally random biosequences, Part I: A general method for *in vitro* selection of folded domains from a random polypeptide library displayed on phage**. *Chem Biodivers* 2006, **3**:827-839.

13. Labean TH, Butt TR, Kauffman SA, Schultes EA: **Protein folding absent selection**. *Genes (Basel)* 2011, **2**:608-626.

14. Trifonov EN: **The triplet code from first principles**. *J Biomol Struct Dyn* 2004, **22**:1-11.

15. •• Newton MS, Morrone DJ, Lee KH, Seelig B: **Genetic code evolution investigated through the synthesis and characterisation of proteins from reduced-alphabet libraries**. *Chembiochem* 2019, **20**:846-856

This study compares ancient genetic codes by comparing the properties of unevolved random proteins made from four different alphabets of the earliest 5, 9, 16, and all 20 amino acids. An unexpectedly high fraction of arbitrarily chosen random proteins are soluble and show some degree of structure.

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

17. • Tretyachenko V, Vymetal J, Bednarova L, Kopecky V Jr, Hofbauerova K, Jindrova H, Hubalek M, Soucek R, Konvalinka J, Vondrasek J *et al.*: **Random protein sequences can form defined secondary structures and are well-tolerated *in vivo***. *Sci Rep* 2017, **7**:15449

Bioinformatic tools were applied to compare natural proteins and unevolved random protein sequences for their propensity to form secondary structure, form aggregates, and be expressible in cells.

18. •• Knopp M, Gudmundsdottir JS, Nilsson T, Konig F, Warsi O, Rajer F, Adelroth P, Andersson DI: ***De novo* emergence of peptides that confer antibiotic resistance**. *mBio* 2019, **10**

In vivo selection of a random peptide library yielded peptides that insert into the cell membrane and reduce the membrane potential resulting in decreased drug uptake. This process led to a 48-fold increase in antibiotic resistance.

19. Keefe AD, Szostak JW: **Functional proteins from a random-sequence library**. *Nature* 2001, **410**:715-718.

20. Roberts RW, Szostak JW: **RNA-peptide fusions for the *in vitro* selection of peptides and proteins**. *Proc Natl Acad Sci U S A* 1997, **94**:12297-12302.

21. Lo Surdo P, Walsh MA, Sollazzo M: **A novel ADP- and zinc-binding fold from function-directed *in vitro* evolution**. *Nat Struct Mol Biol* 2004, **11**:382-383.

22. Mansy SS, Zhang J, Kummerle R, Nilsson M, Chou JJ, Szostak JW, Chaput JC: **Structure and evolutionary analysis of a non-biological ATP-binding protein**. *J Mol Biol* 2007, **371**:501-513.

23. • Narunsky A, Kessel A, Solan R, Alva V, Kolodny R, Ben-Tal N: **On the evolution of protein-adenine binding**. *Proc Natl Acad Sci U S A* 2020, **117**:4701-4709

An inventive computational pipeline was created to mine the Protein Data Bank for adenine-binding proteins and superimpose their binding pockets. This method enabled new insights into the evolution of protein-ligand interactions, suggesting that adenine-binding emerged on several independent occasions.

24. Krishna SS, Grishin NV: **Structurally analogous proteins do exist!** *Structure* 2004, **12**:1125-1127.

25. Kang SK, Chen BX, Tian T, Jia XS, Chu XY, Liu R, Dong PF, Yang QY, Zhang HY: **ATP selection in a random peptide library consisting of prebiotic amino acids**. *Biochem Biophys Res Commun* 2015, **466**:400-405.

26. Seelig B, Szostak JW: **Selection and evolution of enzymes from a partially randomized non-catalytic scaffold**. *Nature* 2007, **448**:828-831.

27. Chao FA, Morelli A, Haugner JC 3rd, Churchfield L, Hagmann LN, Shi L, Masterson LR, Sarangi R, Veglia G, Seelig B: **Structure and dynamics of a primordial catalytic fold generated by *in vitro* evolution**. *Nat Chem Biol* 2013, **9**:81-83.

28. Morelli A, Haugner J, Seelig B: **Thermostable artificial enzyme isolated by *in vitro* selection**. *PLoS One* 2014, **9** e112028.

29. Haugner JC 3rd, Seelig B: **Universal labeling of 5'-triphosphate RNAs by artificial RNA ligase enzyme with broad substrate specificity**. *Chem Commun (Camb)* 2013, **49**:7322-7324.

30. Go A, Kim S, Baum J, Hecht MH: **Structure and dynamics of *de novo* proteins from a designed superfamily of 4-helix bundles**. *Protein Sci* 2008, **17**:821-832.

31. •• Studer S, Hansen DA, Pianowski ZL, Mittl PRE, Debon A, Guffy SL, Der BS, Kuhlman B, Hilvert D: **Evolution of a highly active and enantiospecific metalloenzyme from short peptides**. *Science* 2018, **362**:1285-1288

A previously designed zinc-binding peptide of 46 residues was evolved into a highly efficient globular enzyme through domain fusion, random mutagenesis, and rounds for screening.

32. Korendovych IV, DeGrado WF: ***De novo* protein design, a retrospective**. *Q Rev Biophys* 2020, **53**:33.

33. Baker D: **An exciting but challenging road ahead for computational enzyme design**. *Protein Sci* 2010, **19**:1817-1819.

34. Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH: ***De novo* designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth**. *PLoS One* 2011, **6**:e15364.

35. Digianantonio KM, Hecht MH: **A protein constructed *de novo* enables cell growth by altering gene regulation**. *Proc Natl Acad Sci U S A* 2016, **113**:2400-2405.

36. Digianantonio KM, Korolev M, Hecht MH: **A non-natural protein rescues cells deleted for a key enzyme in central metabolism**. *ACS Synth Biol* 2017, **6**:694-700.

37. •• Donnelly AE, Murphy GS, Digianantonio KM, Hecht MH: **A *de novo* enzyme catalyzes a life-sustaining reaction in *Escherichia coli***. *Nat Chem Biol* 2018, **14**:253-255

In vivo selection from a library of binary patterned proteins with a high probability of folding into four-helix bundle structures yielded a new enzyme that rescues a bacterial strain deficient in a single natural enzyme. The *de novo* enzymatic activity was confirmed *in vitro* through assays with the purified artificial protein.

38. Smith BA, Mularz AE, Hecht MH: **Divergent evolution of a bifunctional *de novo* protein**. *Protein Sci* 2015, **24**:246-252.

39. Hecht MH, Zarzhitsky S, Karas C, Chari S: **Are natural proteins special? Can we do that?**. *Curr Opin Struct Biol* 2018, **48**:124-132.

40. • Wang MS, Hecht MH: **A completely *de novo* ATPase from combinatorial protein design**. *J Am Chem Soc* 2020, **142**:15230-15234

The four-helix bundle library mentioned above yielded proteins with ATPase activity. Both the protein structure and its catalytic mechanism are completely different from naturally evolved ATPases.

41. Der BS, Machius M, Miley MJ, Mills JL, Szyperski T, Kuhlman B: **Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer**. *J Am Chem Soc* 2012, **134**:375-385.

42. Der BS, Edwards DR, Kuhlman B: **Catalysis by a *de novo* zinc-mediated protein interface: implications for natural enzyme evolution and rational enzyme engineering**. *Biochemistry* 2012, **51**:3933-3940.

43. • Obexer R, Godina A, Garrabou X, Mittl PR, Baker D, Griffiths AD, Hilvert D: **Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase**. *Nat Chem* 2017, **9**:50-56

Directed evolution of an artificial aldolase enzyme that had originally been rationally designed led to the creation of catalytic tetrad, which is a complex catalytic center also found in natural enzymes.

44. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF 3rd *et al.*: ***De novo* computational design of retro-aldol enzymes**. *Science* 2008, **319**:1387-1391.

45. Giger L, Caner S, Obexer R, Kast P, Baker D, Ban N, Hilvert D: **Evolution of a designed retro-aldolase leads to complete active site remodeling**. *Nat Chem Biol* 2013, **9**:494-498.

46. Kiss G, Celebi-Olcum N, Moretti R, Baker D, Houk KN: **Computational enzyme design**. *Angew Chem Int Ed Engl* 2013, **52**:5700-5725.