Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Towards descriptor of elementary functions for protein design

Igor N Berezovsky[1,2]

We review studies of the protein evolution that help to formulate rules for protein design. Acknowledging the fundamental importance of Dayhoff's provision on the emergence of functional proteins from short peptides, we discuss multiple evidences of the omnipresent partitioning of protein globules into structural/functional units, using which greatly facilitates the engineering and design efforts. Closed loops and elementary functional loops, which are descendants of ancient ring-like peptides that formed fist protein domains in agreement with Dayhoff's hypothesis, can be considered as basic units of protein structure and function. We argue that future developments in protein design approaches should consider descriptors of the elementary functions, which will help to complement designed scaffolds with functional signatures and flexibility necessary for their functions.

Addresses
[1] Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix 138671, Singapore
[2] Department of Biological Sciences (DBS), National University of Singapore (NUS), 8 Medical Drive, 117579, Singapore

Corresponding author: Berezovsky, Igor N (igorb@bii.a-star.edu.sg)

Current Opinion in Structural Biology 2019, **58**:159–165

This review comes from a themed issue on **Biophysical and computational methods**

Edited by **Laura Itzhaki** and **Alan Lowe**

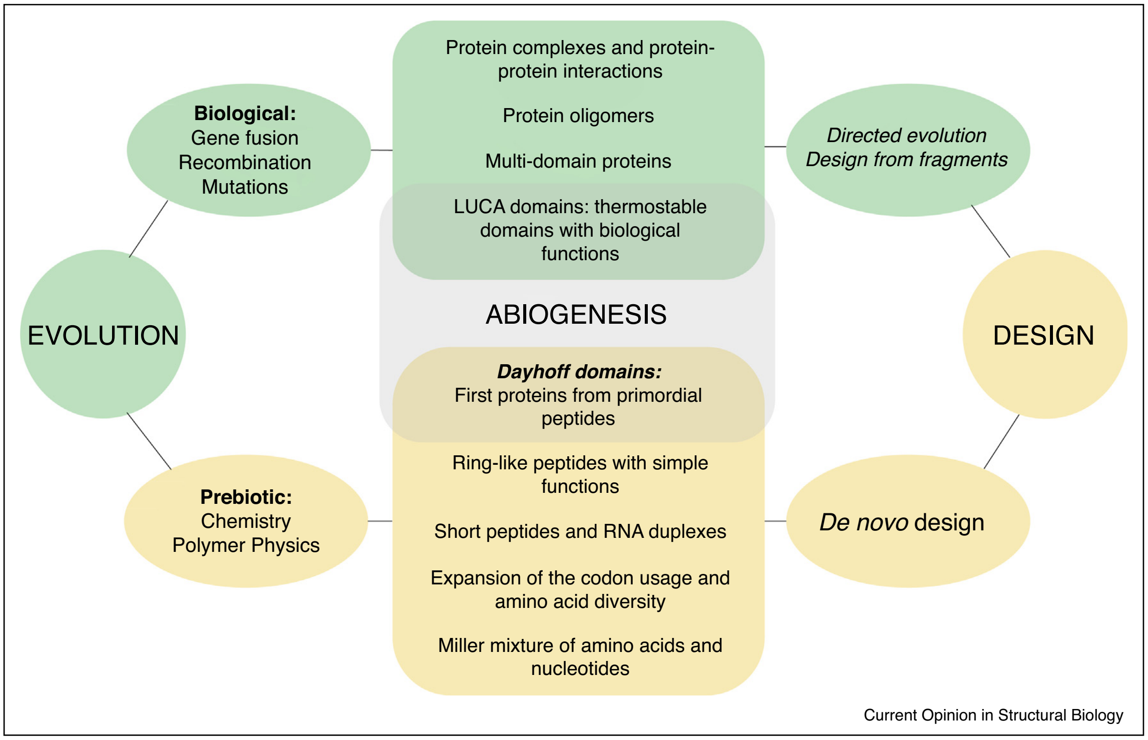
https://doi.org/10.1016/j.sbi.2019.06.010

Introduction

A wealth of sequence information [1] combined with new high-throughput structural and biophysical approaches [2] fuel ambitious goals in protein research, shifting its focus from tedious annotation-characterization tasks to the laboratory evolution [3,4] and design efforts [5•,6]. Diversity of tasks spans from the engineering of modified functions [4,7] to *de novo* design of structures with required activities [8], invoking respective range of experimental methods and theoretical approaches [2–4,9]. While laboratory evolution can be associated with mechanisms of protein evolution since emergence of LUCA [4,10,11], lessons from the very emergence of folds and functions in the Origin of Life can be instructive for *de novo* design [12••,13,14••,15,16]. Figure 1 illustrates a correspondence between events and underlying mechanisms in the evolution and approaches in protein engineering and design that can provide similar outcomes. Factors of success in the enzyme engineering are similar to those that work in natural evolution, including availability of appropriate starting point (wild-type protein), its designability [11,14••] or evolvability [4], and a potential for the acquisition of a new function [4]. The cornerstone method, which allows to obtain robust tailor-made enzymes, is directed evolution [3,4]. One of the major challenges in the protein modification process – correct consideration of epistatic interactions – is typically addressed via ancestral reconstruction [4,17,18•,19], which also helps to facilitate and broaden the substrate specificity [20–22]. Promiscuity of enzymes can be achieved via neutral drift, which can evolve them into generalists with additional activities [23,24]. Among the most interesting recent achievements in laboratory evolution of proteins is the emergence of catalysis in a noncatalytic protein scaffold, such as evolution of chalcone isomerase from a noncatalytic ancestor [19] and of cyclohexadienyl dehydratase from an ancestral solute-binding protein [17]. Extensive computational analysis showed that multiple functions can evolve in nitroreductase superfamily as a result of insertions at key positions in the flavin-binding scaffold [25]. Versatility of promiscuous sulfatase was also used in the evolution of catalysis via repurposing on the basis of high reactivity of its permissive active-site architecture that allow multiple substrate binding modes [24]. Advances in understanding of universality [26] and power of allosteric regulatory mechanisms [27] prompt us to start using them as well in protein design [28].

While ancestral reconstruction can drive one up the evolutionary road to LUCA domains and can potentially provide correct starting point for directed evolution, it cannot help *de novo* protein design. The latter is obviously linked to the question how first folds and functions emerged. Here, we will discuss rules of protein design that nature teaches us, starting from the seminal Margaret Dayhoff's work [15,29] and showing how her provision on the basis of few available sequences found multiple confirmations in current times of big data and provided theoretical foundation for the protein design from fragments. We will show evidences that natural proteins are built from descendant of ancient peptides, will review developments in engineering of modified and chimeric structures from parts of natural proteins, and efforts on the *de novo* design on the basis of elementary structural-functional units. We will argue that basic elementary units of

Figure 1


Protein complexes and protein-protein interactions
Protein oligomers
Multi-domain proteins
LUCA domains: thermostable domains with biological functions
ABIOGENESIS
Dayhoff domains: First proteins from primordial peptides
Ring-like peptides with simple functions
Short peptides and RNA duplexes
Expansion of the codon usage and amino acid diversity
Miller mixture of amino acids and nucleotides
Biological: Gene fusion Recombination Mutations
EVOLUTION
Prebiotic: Chemistry Polymer Physics
Directed evolution Design from fragments
DESIGN
De novo design
Current Opinion in Structural Biology

The evolution-design connectivity scheme.
Major events in the prebiotic and biological evolutions are shown in the middle in relation to relevant chemistry, physics, and biology that determined these outcomes. Different methods in protein engineering and *de novo* design should be used depending on the task.

proteins – closed loops or returns of the protein backbone [30] – were determined by the polymer nature of the polypeptide chain and, then, provided a foundation for elementary units of protein function – elementary functional loops (EFLs, [14••]). We will show that closed loops and EFLs are apparently descendants of prebiotic ring-like peptides that merged into first functional folds, in agreement with what was hypothesized by Eck and Dayhoff yet in 1966 [29]. Finally, we will consider a notion of descriptor of the elementary function, which we proposed as the basic unit for future *de novo* protein design and engineering [13,14••].

Modular structure of proteins from the evolutionary perspective

A number of recent works support Dayhoff's idea of the evolution of modern proteins through the fusion of fragments, which itself may have emerged through the merger of even shorter and simpler ancient peptides [16]. Noteworthy, high-throughput (anti)-correlation analysis of complete proteomes allowed to obtain the most probable sizes (five–six residues) and compositional trends of these peptides [16]. The 'Dayhoff-fragment' stage of the protein evolution left many more marks in contemporary proteins than the previous, short-peptide stage of evolution. For example, analysis of the sequence-fragment families revealed common ancestry of two ancient $(\beta\alpha)_8$-barrel and flavodoxin-like folds, which is characterized by the conserved $(\alpha\beta)_2$ motif of about 40 residues [31]. Another ancient fingerprint [14••] indicates the common ancestry of Rossman-fold enzymes [32]. Symmetric β-trefoil fold was shown to be a result of the oligomerization of simple peptide motifs [33]. Single β-hairpin structure has been identified as a structural unit of different all-β structures, such as Immunoglobulin, β-trefoil, β-roll, β-prism, β-propeller, β-solenoid folds (see Figure 4 in [34] for illustration). Since most of all-β proteins are involved in binding of other proteins, corresponding β-hairpin motifs typically have structural or 'binder' functional signatures discussed in [35]. The set of 40 fragments was proposed as a vocabulary of ancient peptides that existed at the origin of proteins [36]. Numerous evolutionary footprints were found in the analysis of reused protein segments in the large representative set of protein domain [37]. The Evolutionary Classification Of protein Domains (ECOD) database provides an overview of how duplication and divergence of small motifs worked in the emergence and evolution of protein function [38•].

From protein engineering and fragment-based design to *de novo* design

Evolutionary emergence of folds and functions via the fragment duplication, fusion, and recombination also finds a strong support in numerous engineering and design efforts. Designs of different topologies followed by the furnishing them with functions [39] were performed on the βα-barrels (using duplication and recombination of fragments), β-trefoils (duplication) and β-propellers (duplication). Recombination between the $(\beta\alpha)_8$-barrel and flavodoxin-like folds yielded a chimeric protein with modified functional characteristics [40]. Investigation of repeat structures designed from tandems of helix-loop-helix-loop structural motifs showed that natural repeat proteins occupy only a small fraction of possible sequence-structure space, concluding that many novel repeat proteins with specified geometries can be designed [41]. Assembly of complex β-sheet topologies from the *de novo* designed building blocks corroborated an assumption [34,35] that β-sheet fusion mechanism may have worked in the emergence of complex β-sheets during natural protein evolution [42]. An automated SEWING framework was used for design of helical proteins showing some successful outputs and justifying its further development and extension to other types of protein scaffolds [43].

One of the most important goals in protein research is *de novo* design, which is currently an area of rapid developments and great achievements [5•,6]. Recent designs include α-helical bundles with 3–5 [44] and 5–7 [45] helices obtained in Thomson *et al.* labs, respectively. Baker lab also produced completely new TIM-barrel sequence-scaffold designs guided purely by geometrical and chemical principles [46]. Despite challenges in the design of the all-β-sheet proteins, because of the large fraction of non-local interactions that lead to slower folding rates and potential aggregation, the same group designed double-stranded β-helix on the basis of rules describing the geometry of β-arch loops and their interactions in β-arcades [47]. The binding sites of natural cytokines were recapitulated by the otherwise unrelated in topology or amino acid sequence *de novo* globular structures [48•]. Strategies for designing non-natural enzymes and binders were recently reviewed in [8], highlighting developments in computational methods and applications of these techniques in design of receptors, sensors, and enzymes. One of the most recent automated methods, FuncLib, is based on the phylogenetic analysis and Rosetta design calculations, which allows one to improve the enzyme activity via multipoint mutations at the enzyme active site [7]. Khersonsky and Fleishman, who are among the authors of FuncLib, also advocate a synthesis of engineering by fragment substitutions/exchanges and *de novo* design strategies in order to build new proteins on the basis of 'ready-made parts' [49]. They propose that future design method combining phylogenetic analysis, structure/sequence bioinformatics, and atomistic modelling will be more successful rather than above methods used individually.

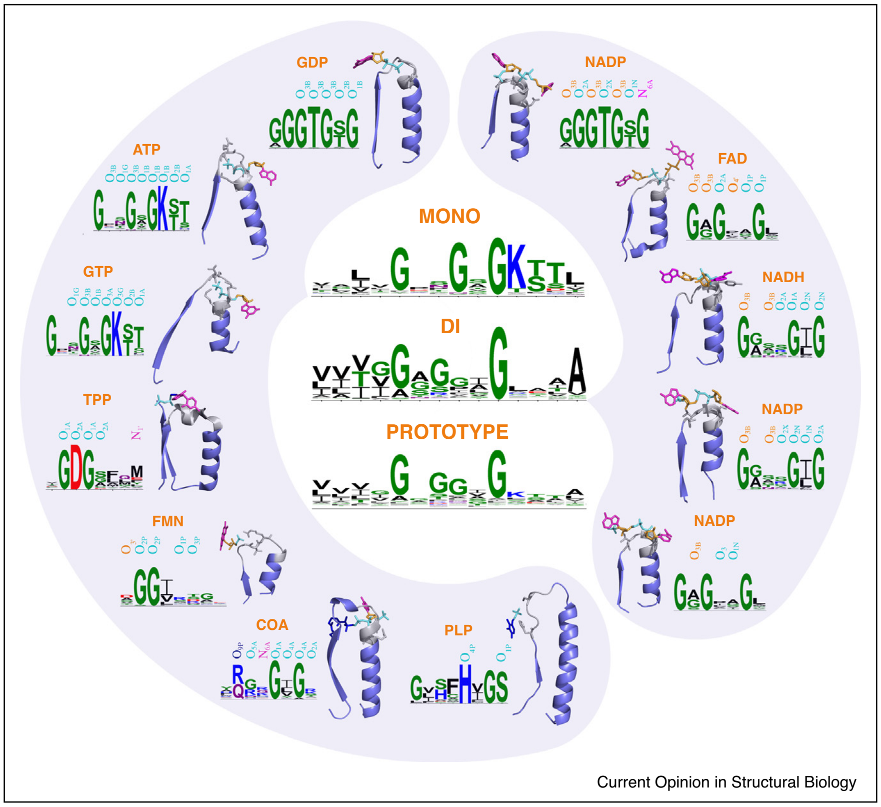
Closed loops and elementary functional loops in the emergence and evolution of protein function

Accepting the strategy of learning from nature and referring to the previous knowledge of proteins built form fragments, one may ask a question about optimal nature-driven unit of proteins [13] that determine and facilitate their structural-functional characteristics. Below, we will share our experience in finding and investigating basic units of globular proteins, will describe the most recent experimental and theoretical evidences of their existence, including the very recent successful *de novo* design of functional folds on the basis of these units, and, finally, will discuss the notion of the descriptor of the elementary function and ways of using it in future design efforts.

Exploring the hierarchy of protein domain structure [50,51] we arrived to the question about basic unit of globular proteins, which would allow one to build a consistent picture of the protein structure organization, folding, and function [13]. It was hypothesized that, because polypeptide backbone chiefly determines protein architecture and topology, its analysis could be instrumental in finding the most basic and universal protein units. An exhaustive enumeration of subsections of protein backbones with contacting ends revealed the universal basic unit of soluble proteins, closed loops or returns of the polypeptide backbone with preferential contour length of 25–30 residues [30]. The loop-n-lock structure [52] of globular proteins [30,34,53] provides a foundation for the hierarchy of protein domain structure [54,55] and co-translational scenario of the folding process [56–58,59•] demonstrated in recent FRET/CQ experiments [60] on the basis of loop hypothesis [59•] and supported by the Sequential Collapse Model [58]. Many other observations of the structural and evolutionary relevance of closed loops were reviewed in detail in Ref. [13].

We further proposed the notion of Elementary Functional Loops (EFLs), which are formed by closed loops that carry one or few functional catalytic or 'binder' residues and serve as a minimal building block in functional mechanisms. We have developed a rigorous computational procedure for the derivation of EFL evolutionary prototypes [61,62]. These prototypes are represented in modern proteins by EFLs, which are presumably descendants of simple ring-like prebiotic peptides that, according to Eck and Dayhoff [29], fused into first protein domains/folds (Figure 1). Therefore, EFLs of a certain prototype can belong to phylogenetically unrelated proteins from remote superfamilies or different folds. At the same time, intermediate sequence profiles obtained in the process of the prototype derivation can be associated with ancestors typical for certain functional (super)families. Figure 2, the P-loop prototype circle, illustrates the relationship between the profiles and prototype of the phosphate-binding signature

Figure 2



Current Opinion in Structural Biology

The P-loop prototype circle.
The group of representative glycine-rich elementary functional loops that bind different (di)nucleotide-containing ligands is shown. 'Mono' and 'Di' signatures in the center show the generalized profiles for the phosphate binding in the nucleotide-containing and dinucleotide-containing ligands. The 'Prototype' logo describes presumable ancient prototype of all glycine-rich phosphate-binding signatures that exist in contemporary proteins.

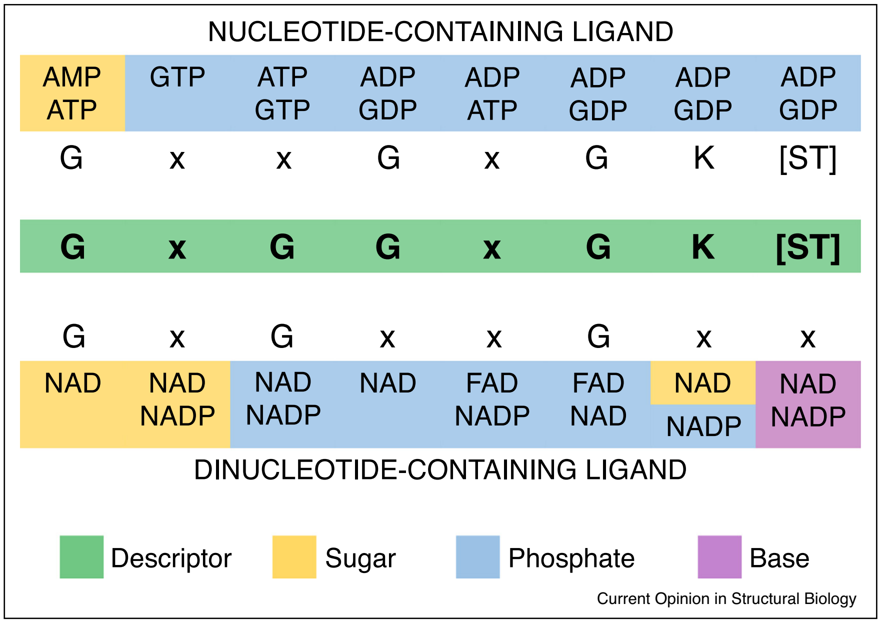
obtained on the basis of NBDB [63] that work in the binding of (di)nucleotide-containing ligands. The functional motifs typical for the binding of phosphates n different nucleotide-containing and dinucleotide-containing ligands correspond to 'MONO' and 'DI' prototypes with GxxGxG and GxGxxG signatures, respectively. The ancient prototype that presumably gave rise to all phosphate-binding fragments yields a Gly-rich signature – GxGGxG [14••,63]. Using collection of prototypes of different elementary functions we showed how contemporary proteins are built from a limited number of elementary functions and investigated intricate relations between different folds and functional superfamilies [14••]. Tracing domain history onto a bipatriate network of elementary functional loop sequences, Aziz *et al.* provided a consistent picture of the emergence and early history of the molecular function [12••,14••]. Importantly, the closed-loop/return shape and characteristic size is determined by the polymer nature of polypeptide chains [64,65], and it is corroborated in many recent studies of the protein evolution [13]. For example, the median length of 40 fragments proposed as a reference set of ancient peptides [36] is 24 residues in agreement with the preferential size of closed loops [13,30]. Detection of reused evolutionary footprints in the representative set of protein domains was shown to be more extensive when the length threshold was lower than 35 residues [37]. The priority of the loop closure over the exact secondary structure content was shown through the persistence of the protein chain returns [66]. Ignoring secondary structure elements reveals connection between the TIM-barrel and flavodoxin-like folds stronger than those between superfamilies of the flavodoxin-like folds superfamilies [31]. Indications that a simple β-hairpin served as an origin of different all-β proteins prompts us to assume that the β-hairpin itself can be just another way of the loops closure [13,52] used in the evolution followed by their fusion and recombination into distinct all-β folds. The NBDB database of profiles involved into the (di)nucleotide binding [63] further supports the conclusion on the primary importance of the loop closure and secondary role of the context-dependent secondary structure [13,30,34,53,56] with important implications for protein design.

Defining and deriving the descriptor of elementary function for protein design

A principal difference between the prototypes [14••,35,61,62] and reconstructed ancestors [17,18•,19,20] can be also associated with the different modes of action in protein engineering and design: while ancestors can be used in modifications within (super)families [18•] and design from fragments [39–41,49], prototypes should be instrumental in *de novo* design of proteins (Figure 1). For example, prototype of the phosphate-biding loop was recently used in *de novo* design of the β-α repeat P-loop protein, which yielded stable and soluble molecule that binds ATP in a magnesium-independent manner, polynucleotides, RNA, and single-strand DNA [67]. Anticipating importance of using functional signatures obtained from natural proteins corroborated in the above design of the P-loop protein [67], we concluded that it is necessary to develop comprehensive description of all features of EFLs that will be used in *de novo* design. To this end, the notion of the descriptor of elementary function was introduced using an example of nucleophile as the key element of the protease catalytic triad [14••]. We proposed that the descriptor of elementary function should contain exhaustive information on all possible sequences, structures, functional signatures, interactions inside the EFL and with the rest of the protein and so on, which are present in different realizations of this elementary function observed in natural proteins. The probabilistic model should be applied to the data collected for every descriptor, where realization of all parameters will depend on the requirements to the targeted protein structure and function. Figure 3 shows schematic representation of realizations of the phosphate-loop descriptor, and corresponding signatures and their interactions with the nucleotide-containing and dinucleotide-containing ligands.

Although there is no doubt that future design efforts should be based on the greatly advanced and successful Rosetta-based approach and its developments [68], we propose that it can be complemented by the library of descriptors of elementary functions, which will contain a comprehensive information on the functions of fragments and possible realizations of their characteristics in designed structures. As a result, while the major guiding principle of the design procedure, finding the energy minimum in the sequence-structure ensemble, should be strictly and universally followed, the probabilistic adjustments of fragments in the folds/domains obtained from realizations of corresponding descriptors would allow one to depart from the perfectly stable but frequently functionally incapable protein and to achieve flexibility and dynamics necessary for its function.

Figure 3



A scheme of the P-loop descriptor's realizations in the interactions with the nucleotide-and dinucleotide-containing ligands.

The central (green) signature shows the major descriptor's motif, which is realized in different signatures interacting with the nucleotide-containing (top) and dinucleotide-containing (bottom) ligands. Colors show atoms of which groups of corresponding ligands are in contact with the protein loop: yellow – sugar, blue – phosphate, magenta – base. Names of ligands (e.g. AMP, ATP, GDP, NAD, FAD) indicate that their atoms are in contact with corresponding position of the descriptor's realization.

Conflict of interest statement

Nothing declared.

Acknowledgement

This work is dedicated to my parents, Riva Z. L'vina and Naum D. Berezovsky, to all their love, support, and inspiration that they gave me. The work was funded by the Biomedical Research Council (BMRC) of the Agency for Science, Technology, and Research (A*STAR), Singapore.

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. Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL: **Confronting the catalytic dark matter encoded by sequenced genomes**. *Nucleic Acids Res* 2017, **45**:11495-11514.

2. Bolon DN, Baker D, Tawfik DS: **Editorial**. *Protein Sci* 2016, **25**:1164-1167.

3. Arnold FH: **Directed evolution: bringing new chemistry to life**. *Angew Chem Int Ed Engl* 2018, **57**:4143-4148.

4. Trudeau DL, Tawfik DS: **Protein engineers turned evolutionists-the quest for the optimal starting point**. *Curr Opin Biotechnol* 2019, **60**:46-52.

5. • Baker D: **What has de novo protein design taught us about protein folding and biophysics?** *Protein Sci* 2019, **28**:678-683.

A brief review describing what protein design teaches us about protein kinetics and thermodynamics, balance of forces in the folding process, control of protein thermostability, and other critical characteristics of natural proteins.

6. Huang PS, Boyken SE, Baker D: **The coming of age of de novo protein design**. *Nature* 2016, **537**:320-327.

7. Khersonsky O, Lipsh R, Avizemer Z, Ashani Y, Goldsmith M, Leader H, Dym O, Rogotner S, Trudeau DL, Prilusky J *et al.*: **Automated design of efficient and functionally diverse enzyme repertoires**. *Mol Cell* 2018, **72**:178-186 e175.

8. Lechner H, Ferruz N, Hocker B: **Strategies for designing non-natural enzymes and binders**. *Curr Opin Chem Biol* 2018, **47**:67-76.

9. Gainza P, Nisonoff HM, Donald BR: **Algorithms for protein design**. *Curr Opin Struct Biol* 2016, **39**:16-26.

10. Nath N, Mitchell JB, Caetano-Anolles G: **The natural history of biocatalytic mechanisms**. *PLoS Comput Biol* 2014, **10**: e1003642.

11. Berezovsky IN, Guarnera E, Zheng Z, Eisenhaber B, Eisenhaber F: **Protein function machinery: from basic structural units to modulation of activity**. *Curr Opin Struct Biol* 2017, **42**:67-74.

12. •• Aziz MF, Caetano-Anolles K, Caetano-Anolles G: **The early history and emergence of molecular functions and modular scale-free network behavior**. *Sci Rep* 2016, **6**:25058.

On the basis the mapping of loop motifs and domains that are likely of very ancient origin, it was shown that emergence and early history of molecular functions is involved in the molecular scale free network of 'elementary functions'. Authors propose that this 'elementary functionome' network including loop motif and structural domains create evolutionary 'waterfalls' that describe the emergence of primordial functions.

13. Berezovsky IN, Guarnera E, Zheng Z: **Basic units of protein structure, folding, and function**. *Prog Biophys Mol Biol* 2017, **128**:85-99.

14. •• Goncearenco A, Berezovsky IN: **Protein function from its emergence to diversity in contemporary proteins**. *Phys Biol* 2015, **12** 045002.

This work presents a comprehensive analysis of the emergence of proteins and evolutionary relations between different folds and functions on the basis of the concept of elementary functional loops (EFLs) and their evolutionary prototypes. Many other important aspects relevant to the emergence and evolution of protein structure and function, such as protein designability, stages of evolution, role of polymer physics in the emergence of basic units of proteins and others are also discussed.

15. Romero Romero ML, Rabin A, Tawfik DS: **Functional proteins from short peptides: Dayhoff's hypothesis turns 50**. *Angew Chem Int Ed Engl* 2016, **55**:15966-15971.

16. Trifonov EN, Kirzhner A, Kirzhner VM, Berezovsky IN: **Distinct stages of protein evolution as suggested by protein sequence analysis**. *J Mol Evol* 2001, **53**:394-401.

17. Clifton BE, Kaczmarski JA, Carr PD, Gerth ML, Tokuriki N, Jackson CJ: **Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein**. *Nat Chem Biol* 2018, **14**:542-547.

18. • Harms MJ: **Enzymes emerge by upcycling**. *Nat Chem Biol* 2018, **14**:526-527.

A news and views underscoring an importance of the ancestral sequence reconstruction (ASR) for revealing the evolutionary interval over which protein activity evolved. Several works in which it was possible to find mutations responsible for gaining activity and their biochemical mechanisms that turn non-enzymatic ancestors to the modern enzymes are discussed.

19. Kaltenbach M, Burke JR, Dindo M, Pabis A, Munsberg FS, Rabin A, Kamerlin SCL, Noel JP, Tawfik DS: **Evolution of chalcone isomerase from a noncatalytic ancestor**. *Nat Chem Biol* 2018, **14**:548-555.

20. Clifton BE, Jackson CJ: **Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins**. *Cell Chem Biol* 2016, **23**:236-245.

21. Gumulya Y, Gillam EM: **Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering**. *Biochem J* 2017, **474**:1-19.

22. Risso VA, Martinez-Rodriguez S, Candel AM, Kruger DM, Pantoja-Uceda D, Ortega-Munoz M, Santoyo-Gonzalez F, Gaucher EA, Kamerlin SCL, Bruix M *et al.*: **De novo active sites for resurrected Precambrian enzymes**. *Nat Commun* 2017, **8**:16113.

23. Martin-Diaz J, Paret C, Garcia-Ruiz E, Molina-Espeja P, Alcalde M: **Shuffling the neutral drift of unspecific peroxygenase in *Saccharomyces cerevisiae***. *Appl Environ Microbiol* 2018, **84**.

24. Miton CM, Jonas S, Fischer G, Duarte F, Mohamed MF, van Loo B, Kintses B, Kamerlin SCL, Tokuriki N, Hyvonen M *et al.*: **Evolutionary repurposing of a sulfatase: a new Michaelis complex leads to efficient transition state charge offset**. *Proc Natl Acad Sci U S A* 2018, **115**:E7293-E7302.

25. Akiva E, Copp JN, Tokuriki N, Babbitt PC: **Evolutionary and molecular foundations of multiple contemporary functions of the nitroreductase superfamily**. *Proc Natl Acad Sci U S A* 2017, **114**:E9549-E9558.

26. Guarnera E, Berezovsky IN: **On the perturbation nature of allostery: sites, mutations, and signal modulation**. *Curr Opin Struct Biol* 2019, **56**:18-27.

27. Guarnera E, Berezovsky IN: **Toward comprehensive allosteric control over protein activity**. *Structure* 2019, **27**:866-878.

28. Khersonsky O, Fleishman SJ: **Incorporating an allosteric regulatory site in an antibody through backbone design**. *Protein Sci* 2017, **26**:807-813.

29. Eck RV, Dayhoff MO: **Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences**. *Science* 1966, **152**:363-366.

30. Berezovsky IN, Grosberg AY, Trifonov EN: **Closed loops of nearly standard size: common basic element of protein structure**. *FEBS Lett* 2000, **466**:283-286.

31. Farias-Rico JA, Schmidt S, Hocker B: **Evolutionary relationship of two ancient protein superfolds**. *Nat Chem Biol* 2014, **10**:710-715.

32. Laurino P, Toth-Petroczy A, Meana-Paneda R, Lin W, Truhlar DG, Tawfik DS: **An ancient fingerprint indicates the common ancestry of Rossmann-fold enzymes utilizing different ribose-based cofactors**. *PLoS Biol* 2016, **14**:e1002396.

33. Lee J, Blaber M: **Experimental support for the evolution of symmetric protein architecture from a simple peptide motif**. *Proc Natl Acad Sci U S A* 2011, **108**:126-130.

34. Berezovsky IN, Trifonov EN: **Loop fold nature of globular proteins**. *Protein Eng* 2001, **14**:403-407.

35. Goncearenco A, Berezovsky IN: **Exploring the evolution of protein function in archaea**. *BMC Evol Biol* 2012, **12**:75.

36. Alva V, Soding J, Lupas AN: **A vocabulary of ancient peptides at the origin of folded proteins**. *eLife* 2015, **4**:e09410.

37. Nepomnyachiy S, Ben-Tal N, Kolodny R: **Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths**. *Proc Natl Acad Sci U S A* 2017, **114**:11703-11708.

38. • Schaeffer RD, Kinch LN, Liao Y, Grishin NV: **Classification of proteins with shared motifs and internal repeats in the ECOD database**. *Protein Sci* 2016, **25**:1188-1203.

The Evolutionary Classification Of protein Domains (ECOD) was developed with the goal to classify proteins that contain short repetitive fragments in domains emerged as a result of the duplication and divergence of small motifs. The paper explains how ECOD classifies proteins with small internal repeats, generic functional motifs, and assemblies of small domain-like fragments in the evolutionary schema.

39. Hocker B: **Design of proteins from smaller fragments-learning from evolution**. *Curr Opin Struct Biol* 2014, **27**:56-62.

40. Bharat TA, Eisenbeis S, Zeth K, Hocker B: **A beta alpha-barrel built by the combination of fragments from different folds**. *Proc Natl Acad Sci U S A* 2008, **105**:9942-9947.

41. Brunette TJ, Parmeggiani F, Huang PS, Bhabha G, Ekiert DC, Tsutakawa SE, Hura GL, Tainer JA, Baker D: **Exploring the repeat protein universe through computational protein design**. *Nature* 2015, **528**:580-584.

42. King IC, Gleixner J, Doyle L, Kuzin A, Hunt JF, Xiao R, Montelione GT, Stoddard BL, DiMaio F, Baker D: **Precise assembly of complex beta sheet topologies from de novo designed building blocks**. *eLife* 2015, **4**.

43. Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T, Kuhlman B: **Design of structurally distinct proteins using strategies inspired by evolution**. *Science* 2016, **352**:687-690.

44. Huang PS, Oberdorfer G, Xu C, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, Baker D: **High thermodynamic stability of parametrically designed helical bundles**. *Science* 2014, **346**:481-485.

45. Thomson AR, Wood CW, Burton AJ, Bartlett GJ, Sessions RB, Brady RL, Woolfson DN: **Computational design of water-soluble alpha-helical barrels**. *Science* 2014, **346**:485-488.

46. Huang PS, Feldmeier K, Parmeggiani F, Velasco DAF, Hocker B, Baker D: **De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy**. *Nat Chem Biol* 2016, **12**:29-34.

47. Marcos E, Chidyausiku TM, McShan AC, Evangelidis T, Nerli S, Carter L, Nivon LG, Davis A, Oberdorfer G, Tripsianes K *et al.*: **De novo design of a non-local beta-sheet protein with high stability and accuracy**. *Nat Struct Mol Biol* 2018, **25**:1028-1034.

48. • Silva DA, Yu S, Ulge UY, Spangler JB, Jude KM, Labao-Almeida C, Ali LR, Quijano-Rubio A, Ruterbusch M, Leung I *et al.*: **De novo design of potent and selective mimics of IL-2 and IL-15**. *Nature* 2019, **565**:186-191.

The most recent computational methods developed in this lab allowed them to design folds with natural binding sites of cytokines, but otherwise unrelated in topology or amino acid sequence. These designs are hyperstable, and they bind human and mouse IL-2R$\beta\gamma_c$ receptor with higher affinity than the natural ones.

49. Khersonsky O, Fleishman SJ: **Why reinvent the wheel? Building new proteins based on ready-made parts**. *Protein Sci* 2016, **25**:1179-1187.

50. Berezovskii IN, Esipova NG, Tumanian VG: **Isolation of the energy-significant parts of the globule and the hierarchy of the domain structure of the protein macromolecule**. *Biophysics* 1997, **42**:557-565.

51. Berezovsky IN, Namiot VA, Tumanyan VG, Esipova NG: **Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition**. *J Biomol Struct Dyn* 1999, **17**:133-155.

52. Berezovsky IN, Trifonov EN: **Van der Waals locks: loop-n-lock structure of globular proteins**. *J Mol Biol* 2001, **307**:1419-1426.

53. Berezovsky IN, Trifonov EN: **Flowering buds of globular proteins: transpiring simplicity of protein organization**. *Comp Funct Genomics* 2002, **3**:525-534.

54. Berezovsky IN: **Discrete structure of van der Waals domains in globular proteins**. *Protein Eng* 2003, **16**:161-167.

55. Koczyk G, Berezovsky IN: **Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure**. *Nucleic Acids Res* 2008, **36**:W239-W245.

56. Berezovsky IN, Kirzhner VM, Kirzhner A, Trifonov EN: **Protein folding: looping from hydrophobic nuclei**. *Proteins* 2001, **45**:346-350.

57. Berezovsky IN, Trifonov EN: **Loop fold structure of proteins: resolution of Levinthas paradox**. *J Biomol Struct Dyn* 2002, **20**:5-6.

58. Bergasa-Caceres F, Rabitz HA: **Predicting the location of the non-local contacts in alpha-synuclein**. *Biochim Biophys Acta Proteins Proteom* 2018, **1866**:1201-1208.

59. • Orevi T, Rahamim G, Hazan G, Amir D, Haas E: **The loop hypothesis: contribution of early formed specific non-local interactions to the determination of protein folding pathways**. *Biophys Rev* 2013, **5**:85-98.

Combination of Förster resonance energy transfer (FRET) and contact quenching (CQ) provided a unified physical model, showing that loop formation in folding varies not only with the type and number of loop-forming residues, but also with the size and properties of the loop-closing amino acids.

60. Jacob MH, D'Souza RN, Schwarzlose T, Wang X, Huang F, Haas E, Nau WM: **Method-unifying view of loop-formation kinetics in peptide and protein folding**. *J Phys Chem B* 2018, **122**:4445-4456.

61. Goncearenco A, Berezovsky IN: **Prototypes of elementary functional loops unravel evolutionary connections between protein functions**. *Bioinformatics* 2010, **26**:i497-i503.

62. Goncearenco A, Berezovsky IN: **Computational reconstruction of primordial prototypes of elementary functional loops in modern proteins**. *Bioinformatics* 2011, **27**:2368-2375.

63. Zheng Z, Goncearenco A, Berezovsky IN: **Nucleotide binding database NBDB—a collection of sequence motifs with specific protein-ligand interactions**. *Nucleic Acids Res* 2016, **44**:D301-D307.

64. Shimada J, Yamakawa H: **Ring-closure probabilities for twisted wormlike chains. Application to DNA**. *Macromolecules* 1984, **17**:689-698.

65. Yamakawa H, Stockmayer WH: **Statistical mechanics of wormlike chains. II. Excluded volume effects**. *J Chem Phys* 1972, **57**:2843-2854.

66. Berezovsky IN, Kirzhner VM, Kirzhner A, Rosenfeld VR, Trifonov EN: **Closed loops: persistence of the protein chain returns**. *Protein Eng* 2002, **15**:955-957.

67. Romero Romero ML, Yang F, Lin YR, Toth-Petroczy A, Berezovsky IN, Goncearenco A, Yang W, Wellner A, Kumar-Deshmukh F, Sharon M *et al.*: **Simple yet functional phosphate-loop proteins**. *Proc Natl Acad Sci U S A* 2018, **115**:E11943-E11950.

68. Geiger-Schuller K, Sforza K, Yuhas M, Parmeggiani F, Baker D, Barrick D: **Extreme stability in de novo-designed repeat arrays is determined by unusually stable short-range interactions**. *Proc Natl Acad Sci U S A* 2018, **115**:7539-7544.