Progress in Biophysics and Molecular Biology 128 (2017) 85–99

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

Basic units of protein structure, folding, and function

Igor N. Berezovsky [a, b, *], Enrico Guarnera [a], Zejun Zheng [a]

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

ARTICLE INFO

Article history:
Received 29 July 2016
Received in revised form
5 September 2016
Accepted 26 September 2016
Available online 30 September 2016

Keywords:
Protein structure
Evolution
Closed loops
Protein function
Hierarchy of domain structure
Protein folding

ABSTRACT

Study of the hierarchy of domain structure with alternative sets of domains and analysis of discontinuous domains, consisting of remote segments of the polypeptide chain, raised a question about the minimal structural unit of the protein domain. The hypothesis on the decisive role of the polypeptide backbone in determining the elementary units of globular proteins have led to the discovery of closed loops. It is reviewed here how closed loops form the loop-n-lock structure of proteins, providing the foundation for stability and designability of protein folds/domain and underlying their co-translational folding. Simplified protein sequences are considered here with the aim to explore the basic principles that presumably dominated the folding and stability of proteins in the early stages of structural evolution. Elementary functional loops (EFLs), closed loops with one or few catalytic residues, are, in turn, units of the protein function. They are apparent descendants of the prebiotic ring-like peptides, which gave rise to the first functional folds/domains being fused in the beginning of the evolution of protein structure. It is also shown how evolutionary relations between protein functional superfamilies and folds delineated with the help of EFLs can contribute to establishing the rules for design of desired enzymatic functions. Generalized descriptors of the elementary functions are proposed to be used as basic units in the future computational design.

Contents

1. Introduction

Despite nowadays wealth of structural data in the Protein Data Bank (Berman et al., 2000) and decades of protein studies, some of the very fundamentals of protein structure are still under intense discussion. The protein structure unit is one of the basic concepts that was first addressed by Svedberg in his seminal work “Mass and

* Corresponding author. Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, 138671, Singapore.
E-mail address: igorb@bii.a-star.edu.sg (I.N. Berezovsky).

http://dx.doi.org/10.1016/j.pbiomolbio.2016.09.009
0079-6107/

size of protein molecules" (Svedberg, 1929). After analysis of sedimentation fractions obtained in ultracentrifugation experiments, he postulated that there is a size increment in proteins of about 160 amino acid residues. Svedberg concluded that "proteins … can, with regards to molecular weight, be divided into four subgroups …. . The molecular characteristic of the three higher sub-groups are – as a first approximation – derived from molecular mass of the first subgroup by multiplying by the integers two, three, …". The evaluation of the optimal surface/volume ratio of hydrophilic and hydrophobic residues in the theoretical landmark work by Bresler and Talmud resulted in the first formulation of the "minimal condition" for the stable globular protein (Bresler and Talmud, 1944a, 1944b): (i) the hydrophobic nucleus should be covered by the hydrophobic envelope; (ii) van der Waals interactions are the major forces for globular protein formation. As a result, Bresler and Talmud also postulated that "sharply limited size" of about 130 residues (estimated on the basis of hydrophobic/hydrophilic balance) is the archetype for a stable globular protein (Bresler and Talmud, 1944a). Remarkably, the size of 130–160 amino acid residues is well within the range of typical protein domain sizes, from 100 to 200 residues, observed in the analysis of crystalized proteins (Gerstein, 1998; Jones et al., 1998; Wheelan et al., 2000) regardless of the domain/fold type. The exponential increase of protein designability (England and Shakhnovich, 2003) is best manifested in the range of protein chain length corresponding to the typical domain size (Zeldovich et al., 2006), indirectly corroborating the fundamental importance of the latter. Optimal ring closure size about 300–600 base pairs for double-stranded DNA (Shore et al., 1981; Berezovsky, 2002; Trifonov et al., 2001) and recombination experiments with bacterial insertion sequences (Goryshin et al., 1994) show that the advantage of ring's stability for protection of the gene ends and continuity of replication and transcription could be used at the DNA ring-closure stage of evolution, rendering, at the same time, the domain size to 100–200 amino acid residues (Berezovsky, 2002; Trifonov et al., 2001; Goncearenco and Berezovsky, 2015).

It has been shown that the formation and evolution of large proteins is chiefly driven by domain (re)combinations (Chothia et al., 2003; Koonin et al., 1998, 2002), and their structures and functions are shaped by mutations (Aharoni et al., 2005; Glasner et al., 2006; Roodveldt et al., 2005; Tokuriki and Tawfik, 2009; Romero and Arnold, 2009). Yet, protein domains themselves should be built from small and simple elementary units, because it is virtually impossible that evolution would have started from the large multidomain structures that perform multi-step biochemical transformations. Discontinuous domains and alteration of domain structure at different levels of energy hierarchy are universal inherent characteristics of protein structure (Berezovsky, 2003; Berezovsky et al., 1999, 2000a; Koczyk and Berezovsky, 2008) and energetics (Berezovskii et al., 1997; Berezovsky et al., 1997, 2000b) which corroborates an existence of elementary units from which all domains are universally built. Though three common structural patterns were described by Levitt and Chothia back in 1976 (Levitt and Chothia, 1976), protein modularity and architecture are still under intense discussion (Fernandez-Fuentes and Fiser, 2013; Hleap and Blouin, 2016; Rorick, 2012; Vallat et al., 2015).

This review is focused around common basic units of globular proteins, closed loops of nearly standard size of 25–30 residues, which were first discovered in the analysis of crystallized proteins (Berezovsky et al., 2000c). The physical origins and sequence/structure characteristics of closed loops, their role in formation of protein folds/domains, and potential involvement in co-translational protein folding are discussed in this work. Special attention is paid to the structural organization and folding of protein folds/domains. In particular, folding simulations and potential evolutionary implications obtained in the analysis of simplified proteins are reviewed here. Further, we consider loops that deliver one of few catalytic residues to the functional site, so-called elementary functional loops (EFLs). The computational framework for the derivation of the EFLs' evolutionary prototypes is described. We also discuss here the structure-function relations from an evolutionary perspective, obtained by using EFLs and their prototypes/profiles, their importance for the establishing rules for design of desired functions, and the "descriptor of elementary function". In conclusions of this work, an outline of the major future research directions is sketched, including the annotation/prediction of protein function on the whole-proteome level and computational protocol for derivation and usage of the descriptor of elementary function. The latter is planned to be used as the elementary building block in future computational design of protein function.

2. Discovery of closed loops

Rigorous study of the hierarchy of protein domain structure (Berezovsky et al., 1999, 2000a; Berezovskii et al., 1997) prompted one of the authors to raise a question about the size and shape of the elementary structural unit of protein domain (Berezovsky et al., 1999; Berezovskii et al., 1997). Since protein architecture and topology is determined by the protein backbone, it was assumed that the latter can be instrumental in detecting the protein partitioning. Indeed, the typical curve of a protein backbone revisits the densely packed parts of the molecule, "walking" back and forth between them and forming complex/discontinuous domains. It was hypothesized, therefore, that following the chain's trajectory one can delineate the highly packed and stable elementary units (sub-domains) of globular proteins. An exhaustive enumeration of subcurves of the protein backbones with close contacts (short three-dimensional distances) between their ends resulted in the discovery of common basic units of proteins - closed loops or returns of the polypeptide backbone with preferential contour length of 25–30 residues (Berezovsky et al., 2000c). It is important to note that these are not loops in the traditional definition as connectors between elements of secondary structure studied elsewhere (Kolinski et al., 1997; Kwasigroch et al., 1996; Leszczynski and Rose, 1986; Martin et al., 1995; Oliva et al., 1997; Panchenko and Madej, 2004, 2005). It was shown that the specific size of the closed loops originates from the polymer nature of polypeptide chains. First, according to Shimada-Yamakawa theory the maximal ring-closure probability of the polymer chain is 3–4 persistence lengths (Shimada and Yamakawa, 1984; Yamakawa and Stokmayer, 1972). Second, the available experimental data on the persistence length of homo- and heteropolymers of different amino acid compositions (Schimmel and Flory, 1967) and consideration of the average content of secondary structure elements in proteins resulted in an estimate of the typical size of closed loops in natural proteins - 20–50 residues (Berezovsky et al., 2000c). Thus, the preferential size of 27 ± 5 residues observed in the discovery of closed loops fairly agrees with the theoretically estimated interval. Closed loops are common in all proteins regardless of the superkingdom (see Fig. 1A and B with distributions of closed loops in prokaryotic and eukaryotic proteins), fold type (Berezovsky et al., 2000c; Berezovsky and Trifonov, 2001a; Berezovsky and Trifonov, 2001b), secondary structure content (Fig. 1C–F), as well as the protein size (Berezovsky, 2002). Noteworthy, elements of secondary structure have different rigidity compared to those of the non-structured polypeptide chain and they are involved into contacts with each other forming the scaffolds of folds/domains. The typical size for the elements of secondary structure is between five and fifteen residues (see for example a distribution of the α-helices'

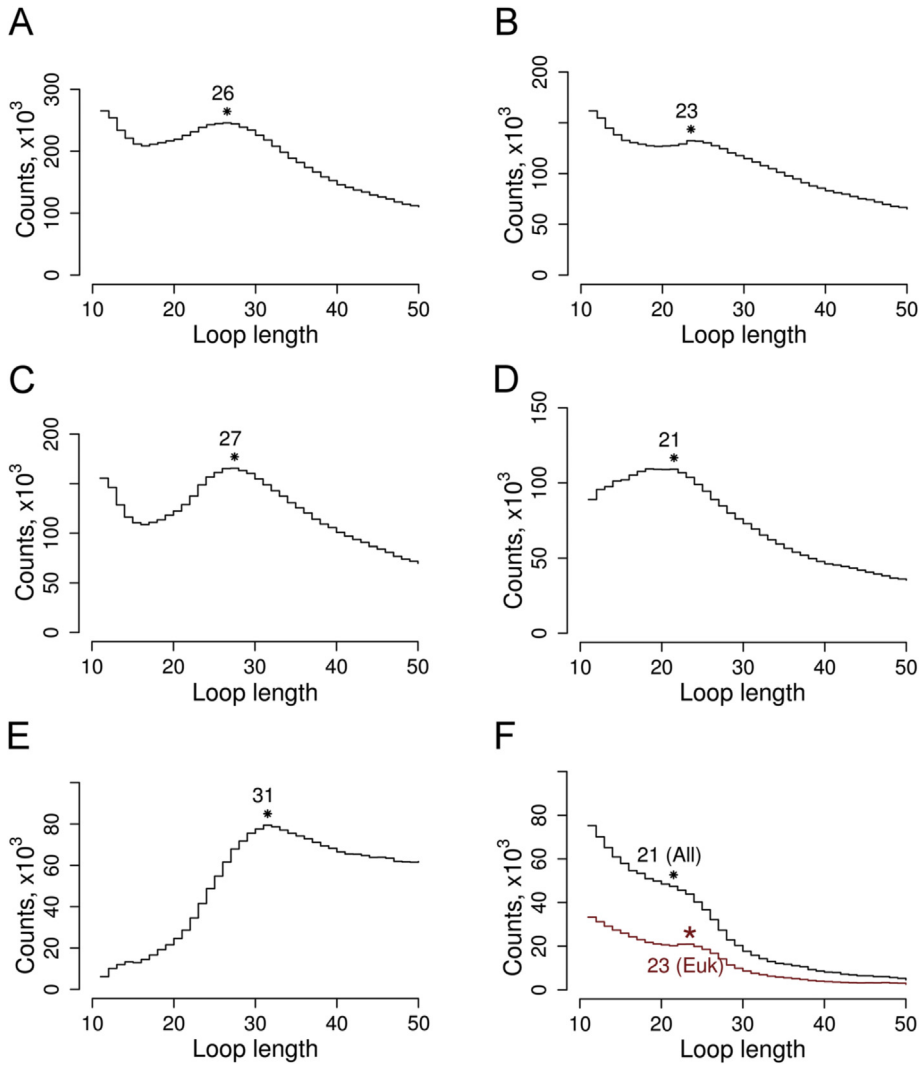

Fig. 1. The universal basic element of globular proteins - closed loops of 25–35 amino acid residues. Closed loops of nearly standard size are omnipresent in proteins of both prokaryotes (A) and eukaryotes (B). They can include different combinations of secondary structure elements, only α–helices (C), only β–strands (D), both α–helices and β–strands (E), or be completely unstructured (F). *Methodology used for producing* Fig. 1. Complete set of structures was downloaded from the Protein Data Bank, and CD-HIT was used to eliminate redundancy at 50% level. Protein chains longer than 600 residues were also excluded, as they can be dominated by the non-globular structures. In total, 14501 and 10732 protein chains of eukaryotes and prokaryotes, respectively, were analyzed. Closed loops are defined as subtrajectoires of the polypeptide backbone with the end-to-end (Cα-Cα) distance within 10 Å. Distributions were plotted for loops with the contour length longer than 10 residues. The secondary structure composition in the loops was analyzed in order to group them into three categories: loops that contain only α-helices (C), only β -strands (D), loops contain both elements of secondary structures (E), and neither of the two elements (F).

sizes (Fig. 2 in Berezovsky et al. (2015)) with a maximum at ten residues). Longer loops contain combination of α-helices and β-strands (maximum of the loop size distribution is at 31 residues Fig. 1E) or only α-helices (the most frequent size is 27 residues, Fig. 1C), whereas only β-containing structures and completely unstructured loops are slightly shorter with preferential loop size of 21 residues in both cases (Fig. 1 D and F). It may seem counterintuitive that closed loops with only α-helical segments have smaller preferential contour length (Fig. 1C) than loops containing combinations of the α-helices and β-strands, and the distribution of the latter is strongly skewed (Fig. 1E). The explanation is, however, pretty straightforward: helices' containing only loops can include few (sometimes disconnected) turns of the former, being, otherwise, chiefly unstructured and flexible. The "α-turn-β" loops are characteristic elements of α/β-barrels/sandwiches and similar architectures with high packing, where extended α and β segments are heavily involved in formation of the overall fold. Overall, while typical size of the closed loops can be slightly affected by the compositions of secondary structure elements, return of the chain to itself many times with formation of numerous chain-to-chain contacts is a common principle of structural organization in all types of globular proteins.

Structural and evolutionary relevance of closed loops was corroborated by different independent studies. Genetic foundation of closed loops as individual evolutionary units is revealed by the estimated size of ancestral exons (Roy et al., 1999) correlated, in turn, with centripetal structural modules of proteins (Sato et al., 1999). It is also corroborated by the size distribution of insertions and deletions in proteins (Qian and Goldstein, 2001) and by the analysis of the recombinatorial swapping of protein sequences (Voigt et al., 2002), which suggest segments of 20–30 amino acid residues as apparent protein building blocks. The probability distribution of any amino acid pair to be in contact (Dokholyan and Shakhnovich, 2001), break in the power-low pattern observed for

the protein fractal characteristics (Moret and Zebende, 2007), as well as Delaunay (Taylor and Vaisman, 2006) and Voronoi (Angelov et al., 2002) tessellations provide strong structural evidences in support of closed loops as basic units of proteins. A strategy for finding closed loops in protein pairs with the same fold but with the insertions/deletions and comparison of detected loops with experimentally determined foldons (Wetlaufer, 1973; Panchenko et al., 1996, 1997) further corroborate the fundamental role of closed loops (Chintapalli et al., 2010). The contact probability, i.e. probability that two monomers of a polymer chain separated by a particular contour distance are found in contact in three dimensions, was used in recent study of the organizing principles of the protein and RNA molecules (Liu and Hyeon, 2016). It was shown that closed loops of preferential size lie in the foundation of the equilibrium globule with "intermingling chain configurations" observed in proteins, contrary to the territories formed in a crumpled globule typical for RNA. The contact probability is a generic characteristic that can be used for distinguishing the phases of a polymer chain involved in the distinctive packing of nucleic acids and proteins (Grosberg, 2016). Sequence conservation observed from multiple sequence alignments built for proteins with mapped closed loops points to an evolutionary persistence of closed loops (Yew et al., 2007; Berezovsky et al., 2002). On the basis of the definition of closed loops (Berezovsky et al., 2000c) it was proposed to decompose proteins into supersecondary structure modules, so-called Smotifs (Fernandez-Fuentes and Fiser, 2013). It was also assumed that Smotifs can be used as basic elements in protein structure modeling and design. Coarse-grain simulations were used to illustrate the role of closed loops in the folding process (Chintapalli et al., 2014; Papandreou et al., 2004).

3. Closed loops and protein folds

Mapping the closed loops on sequences of natural proteins shows that the standard size loops almost completely cover the protein sequences in majority of the cases (Berezovsky and Trifonov, 2002a). It indirectly shows that design of modern proteins as consecutive arrays of the loops (chain returns) was apparently determined by the mutual work of physics and evolutionary selection. Likely, the former brought to the prebiotic scene ring-like peptides with primitive functions (Trifonov and Berezovsky, 2002), while the latter provided a fusion of those peptides together leading to the formation of functional protein folds/domains that perform multi-step biochemical transformations (Trifonov and Berezovsky, 2003). Fig. 2 illustrates that all protein folds are built from the closed loops of preferential size of 25–35 residues. Fig. 2 (charts A and B) show small 1eca (all alpha globin-like fold) and 1pht (all beta SH3-like barrel) folds of 56 and 67 residues, respectively. Beta-galactosidase (1023 residues, Fig. 2C) exemplifies the generic rule of splitting proteins into domains with closed loops of nearly standard size 25–30 residues playing a role of their elementary units (Fig. 2C). There are also examples of proteins, which are regularly packed structures with compactness achieved by the tight systematic packing of the standard size loops (Berezovsky, 2002; Berezovsky and Trifonov, 2001b). Altogether, Fig. 2 indicates that natural protein domains/folds did not evolve from the random compact globules, so-called Flory globule. The latter is characterized by the polypeptide chain that randomly walks between the globule's walls with the size of chain returns proportional to the linear size of the globule, hence short closed loops in small globules and long loops – in large ones. In natural proteins, however, loop size and linear succession of loops are the major invariants in folds, with loop ends serving as the punctuation marks which organize the overall structure of the protein fold/domain (Berezovsky and Trifonov, 2002a; Trifonov and Berezovsky, 2003; Lamarine et al., 2001; Berezovsky et al., 2001). Importantly, positional autocorrelation analysis of protein sequences of 23 fully sequenced bacterial genomes corroborates an existence of the "punctuation" on the sequence level, revealing the signal in the region 24–31 residues that vanishes in randomized sequences (Berezovsky et al., 2001). Another example is the long-range contact order, which suggests the presence of folding nucleus at an interval of approximately 25 residues (Gromiha and Selvaraj, 2001), whereas residues that make multiple contacts affect the protein folding rate (Gromiha, 2009). All the above supports a scenario where, firstly, closed loops (or the polypeptide chain returns) of invariant size originate from the ring-like prebiotic peptides and, secondly, selection pressure in protein evolution maintains the nearly standard loop size, and the linear arrangement along the protein sequence and tight packing of loops in the structures of modern protein folds/domains. Closed loops in contemporary proteins are structurally heterogeneous, virtually consisting of any combination of secondary structure elements within their contours (Fig. 1C–F).

Switching to the language of biological evolution, it appears that evolution had appropriated simple ring-like peptides with primitive functions that emerged in prebiotic world on the basis of the polypeptides' natural flexibility. Fusion of the corresponding short genes presumably resulted in the formation of first folds/domains with several tightly packed closed loops. It is worth to note that the size of domains could be reinforced by the same the law of polymer flexibility-based ring closure that is applied to the double-stranded DNA, according to which 300–600 base pairs optimal for dsDNA circularization (Shore et al., 1981; Shimada and Yamakawa, 1984; Yamakawa and Stokmayer, 1972) can be instrumental in maintaining the typical domain size of 100–200 residues (Berezovsky, 2002; Trifonov et al., 2001; Goncearenco and Berezovsky, 2015).

4. Contribution of Van der Waals interactions to protein stability

A spatial arrangement of the polypeptide chain returns in a 3D-structure of the fold/domain underlies the hierarchy in its structures (Berezovsky, 2002; Berezovsky, 2003; Koczyk and Berezovsky, 2008) and presumably its folding scenario (Papandreou et al., 2004; Trifonov and Berezovsky, 2003; Berezovsky et al., 2001; Berezovsky and Trifonov, 2001c, 2002b). The major interactions stabilizing protein globules include van der Waals, hydrogen bonds, ion pairs. Of note, "case study" of protein thermophilic adaptation had become a classical field that provided many fundamental insights in the contribution of different types of interactions and their combinations in molecular mechanisms and evolutionary strategies of adaptation (Berezovsky, 2011; Berezovsky et al., 2005; Berezovsky and Shakhnovich, 2005; Berezovsky et al., 2007; Goncearenco et al., 2014; Ma et al., 2010; Pucci et al., 2016; Pucci and Rooman, 2014; Tokuriki et al., 2009; Zeldovich et al., 2007; Gromiha et al., 2002, 2013; Kumar et al., 2006). Among all stabilizing interactions, only van der Waals is a shear interaction that involves all the atoms in the structure (Ponnuswamy and Gromiha, 1994; Pace et al., 1996; Dill, 1990). Contrary to hydrogen bonds which are always saturated (either by contact with water or within protein interior) and ion pairs that can be shielded by counter ions, interactions between fluctuating atomic dipoles – origin of the van der Waals forces – occur between all atoms, contributing to the enthalpy term of the protein stability. While the role of van der Waals interactions in formation of the protein-globule cores was first described by Bresler and Talmud already in 1944, an analytical approach for the calculation of van der Waals interaction energy in globular proteins was proposed only five decades later (Berezovsky et al., 2000b; Berezovskii et al., 1998a). In these works, van der

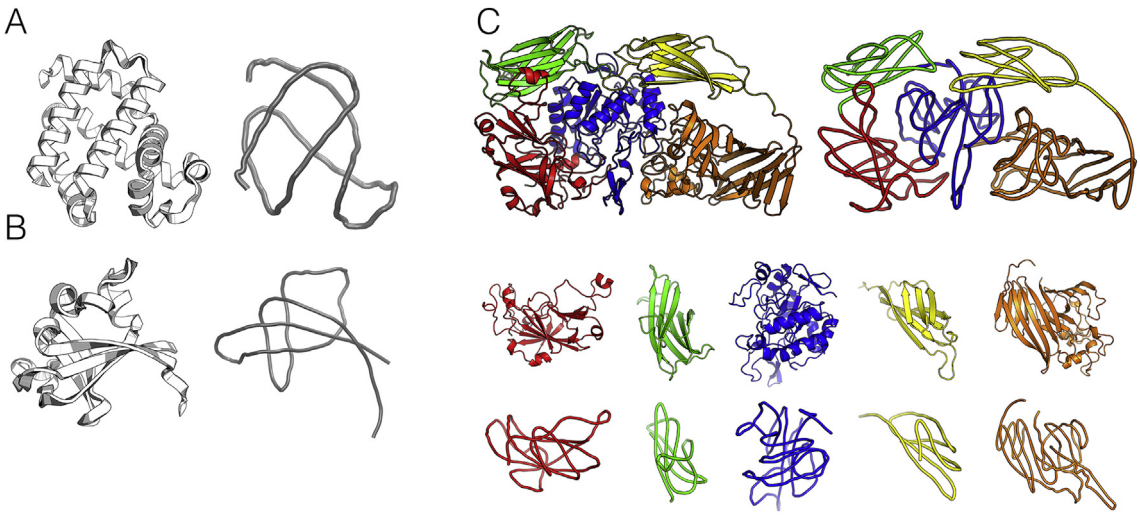

Fig. 2. Globular proteins are built from the closed loops of preferential size of 25–35 residues. A. Small all alpha globin-like fold, 56 residues (1eca). **B**. Small all beta SH3-like barrel fold, 67 residues (1pht). **C**. Beta-galactosidase is a big five-domain protein (1023 residues); each domain of the beta-galactosidase consists of several loops. Beta-galactosidase exemplifies the generic rule of splitting proteins into domains with closed loops of nearly standard size 25–30 residues playing a role of their elementary units (Berezovsky, 2002). Smoothing procedure replaces the coordinates of every Cα by averaged coordinates for seven Cα atoms centered at a given residue.

Waals interactions were calculated on the basis of the interaction of a variable electromagnetic field with a continuous media characterized by the dielectric permittivity. The latter is a function of the electromagnetic filed frequency and properties of the media. In view of the notion of local medium permittivity, Maxwell equations were solved using the perturbation theory with required level of accuracy. This approach allows one to take into account peculiarities of the continuous media for structural units of any scale and, therefore, to explore hierarchical relationships in non-homogenous macromolecular systems of globular protein. Limitation of the widely used pair-wise approximation was also demonstrated, as corrections due to high order interactions at distances longer than 3–5 Å have the same order of magnitude as the energies of pair-wise interactions (Berezovsky et al., 2000b; Berezovskii et al., 1998a).

5. Closed loops and the hierarchy of protein domain structure

Unraveling the hierarchy of protein domain structure have led to the very important implications for the problem of assigning structural domains in globular proteins. First, the hierarchical subdivision on the basis of the van der Waals energy contributions makes it possible to delineate structural domains consisting of any number of continuous and discontinuous segments of the polypeptide chain (Berezovsky, 2003; Berezovsky et al., 1999; Koczyk and Berezovsky, 2008; Berezovskii et al., 1997). Second, it allows one to reconcile traditional definitions of domains (Alden et al., 2010), where performance of human experts is still the best (10% disagreement) and automatic methods do not reach even this level (Holland et al., 2006; Veretnik et al., 2004; Ochoa et al., 2015; Kelley and Sternberg, 2015; Wieninger and Ullmann, 2015). Fig. 3 shows maltogenic amylase (1sma, chain A), where combinations of closed loops (Fig. 3A, left structure) underlie the hierarchy of the domain structure with two, three, and four domains at different levels of hierarchy (Fig. 3B). It exemplifies continuous (Fig. 3 B, central structure: residues 1–125 (blue), 126–537 (green), 538–588 (orange)) and discontinuous domains (Fig. 3 B, right structure: residues 126–242 and 376–499 (green)), and existence of the hierarchy of domain structure (Fig. 3 B). Fig. 3 C shows a match between domains at different levels of hierarchy and those determined by other methods, such as Domain Parser, NCBI-based server, and PDB classification (Koczyk and Berezovsky, 2008), representing the reconciliatory power of the hierarchy. Alternative domain structures include two continuous (Fig. 3 B, left) and three continuous domains (Fig. 3 B, center), and four-domain structure that consist of continuous (residues 1–125 (blue), 243–375 (red), and 500–588 (orange)) and discontinuous (residues 126–242 and 376–499, green) domains (Fig. 3 B, right). In the transition from the three-to four-domain structure, domain 126-537 (Fig. 3 B, center structure, green) is split. As a result, the complex domain 126-242 and 376–499 (Fig. 3 B, right, green) and domain 243-375 (Fig. 3 B, right, red) are formed. The subdomain 500-537 that consists of two sequential loops (508–524 (cyan) and 520–535 (yellow) in Fig. 3 A, left) extends domain 538–588 into 500–588 (Fig. 3 B, right, orange). The Domain Hierarchy and closed Loops (DHcL) web-server (Koczyk and Berezovsky, 2008) (currently located at: http://cropnet.pl/dhcl/) is an implementation of the hierarchical approach to domain decomposition. The DHcL server also allows one to determine the set of closed loops in the protein of interest and to obtain the corresponding primary and secondary van der Walls locks that characterize its overall loop-n-lock structure.

6. Loop-n-lock structure of globular proteins and co-translational protein folding

The formation of protein globule from a compact linear array of nearly standard size closed loops (Berezovsky et al., 2000c; Berezovsky and Trifonov, 2001b) is apparently supported by the primary and secondary van der Waals lock, forming the loop-n-lock structure of the protein globule (Koczyk and Berezovsky, 2008; Berezovsky and Trifonov, 2001a). Tightly packed locks consolidate the hydrophobic core, and mostly polar loop-heads form the hydrophilic surface of a protein globule as a result of the loop-and-lock structure formation (Koczyk and Berezovsky, 2008). The latter (Berezovsky and Trifonov, 2001a) facilitates discrete structure of protein domains (Berezovsky, 2003) and its hierarchy (Berezovsky et al., 1999; Koczyk and Berezovsky, 2008; Berezovskii et al., 1997). Specifically, good quantitative agreement between the boundaries of domains and loops (Berezovsky, 2003) and shifting the loops between domains with formation of alternative van der

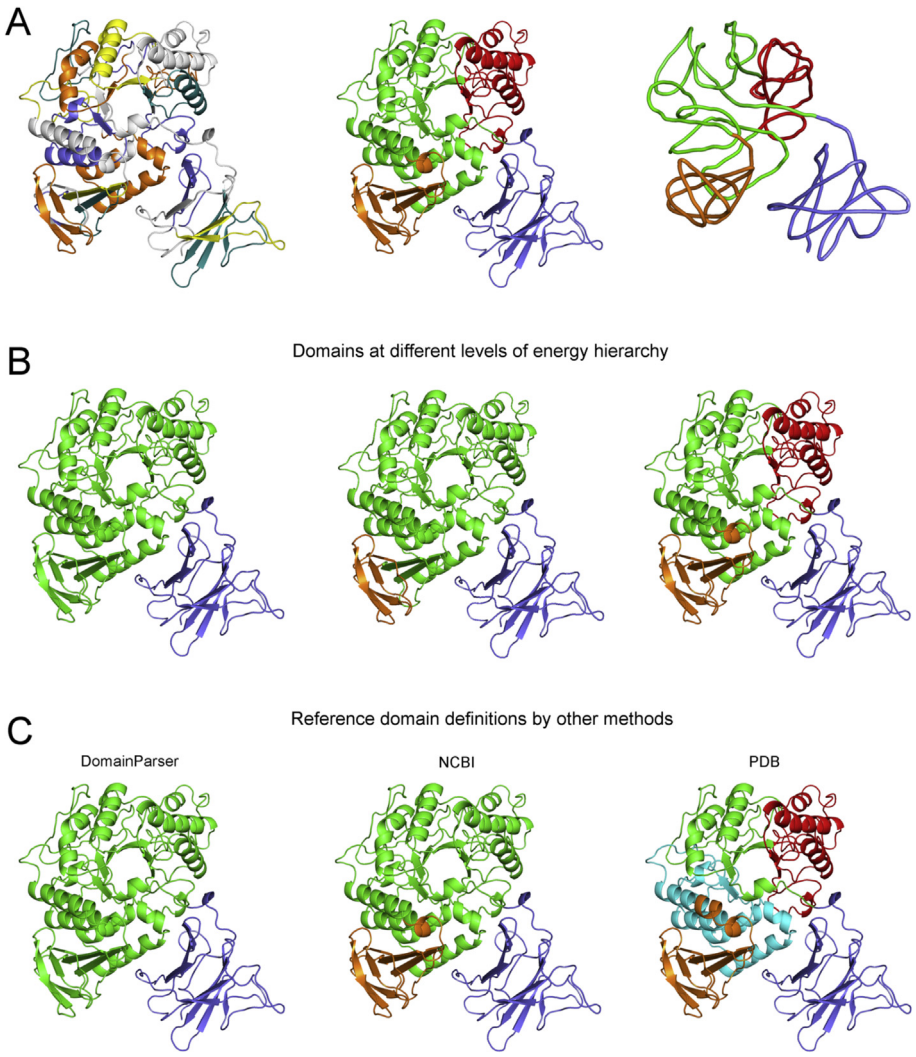

Fig. 3. Hierarchy of protein domain structure is based on the loop-n-lock structure and on the rearrangement of loops with formation of alternative sets of domains. A. Closed loops and domains mapped on the structure of maltogenic amylase (1sma, chain A). Loops are (left structure): 6–29 (blue), 38–58 (yellow), 83–116 (cyan), 146–177 (cyan), 175–198 (orange), 200–216 (blue), 212–244 (yellow), 271–298 (orange), 324–353 (cyan), 356–374 (blue), 380–420 (orange), 426–466 (blue), 462–483 (yellow), 508–524 (cyan), 520–535 (yellow), 532–546 (blue), and 545–572 (orange). Noteworthy, all the loops and linkers between them are of nearly standard size about 25 residues. Middle structure shows four-domains. Domain 1: residues 1–125 (blue); domain 2: 126–242 and 376–499 (green); domain 3: 243–375 (red); domain 4: 500–588 (orange). Right structure shows the same four domains in smoothed representation. **B**. Domain structure at different levels of energy hierarchy. Two-domain structure (left): domain 1: 1–125 (blue); domain 2: 126–588 (green). Three-domain structure (center): domain 1: 1–125 (blue); domain 2: 126–537 (green); domain 3: 538–588 (orange). Four-domain structure (right): domain 1: 1–125 (blue); domain 2: 126–242 and 376–499 (green); domain 3: 243–375 (red); domain 4: 500–588 (orange). **C**. Domains at different levels of energy hierarchy matches to those determined by other methods (see Koczyk and Berezovsky (2008); Alden et al. (2010); Holland et al. (2006) for details). *DomainParser* decomposition (left structure): domain 1: 1–125 (blue); domain 2: 126–588 (green). *NCBI* decomposition (center): domain 1: 1–125 (blue); domain 2: 126–500 (green); domain 3: 500–588 (orange). *PDB* decomposition: domain 1: 1–125 (blue); domain 2: 126–242 (green); domain 3: 243–375 (red); domain 4: 376–494 (cyan); domain 5: 495–588 (orange). The loops and domains are determined with the help of the Domain Hierarchy and closed Loops (DHcL) web-server, which is an implementation of the hierarchical approach to domain decomposition. It also allows one to survey representation of a protein as a set of closed loops and to obtain primary and secondary van der Walls locks that stabilize its overall structure. The DHcL web-server is currently located at: http://cropnet.pl/dhcl/.

Waals domains at different levels of energy hierarchy (Berezovsky et al., 1999; Koczyk and Berezovsky, 2008; Berezovskii et al., 1997) corroborates the role of closed loops as the elementary units of protein domains. While the overall loop-n-lock structure is chiefly stabilized by van der Waals interactions, the directed interactions (such as hydrogen bonds (Berezovskii et al., 1998b) or ion pairs) can modulate interactions between individual loops and (sub)domains depending on the environmental conditions (pH, hydration etc). This modulation causes loop regrouping that results in different conditions-dependent sets of cooperative units in microcalorimetric experiment (Protasevich et al., 1987), which, in turn, is a manifestation of the alternative sets of domains at different levels of hierarchy (Berezovsky, 2003; Berezovsky et al., 1999; Koczyk and Berezovsky, 2008; Berezovskii et al., 1997).

Loop-n-lock structure is directly linked to the co-translational scenario of the protein folding. Analysis of crystalized proteins, specifically, their Kyte and Doolitle hydropathicity plots, supported by the autocorrelation analysis of complete sets of protein sequences of 23 prokaryotic genomes (Berezovsky et al., 2001) showed that loops' ends are mostly hydrophobic. Noteworthy, “Levinthal-like” estimate (Berezovsky and Trifonov, 2002b) of the size of structural unit, *i.e.* substructures for which exhaustive conformational sampling would be feasible within typical folding times (from 10^{-1}–10^{3} s), is about 20–30 residues if three alternative conformations for each residue are considered in the estimate. Therefore, consecutive looping of the nascent polypeptide chain

with the loop closure provided by the strong van der Waals interactions between hydrophobic residues (that eventually form the protein core) followed by the arrangement of these loops into the final fold/domain seems to be a plausible scenario of the globule formation (Berezovsky et al., 2001; Berezovsky and Trifonov, 2002b). Assuming appearance of the linearly arranged loops (Berezovsky and Trifonov, 2001b) in the course of translation process, the folding time of the whole globule will be only several fold larger than the one required for a single unit (loop) The estimated time for the folding of a typical (say 150 residues) protein fold is in a fair agreement with the typical translation rate of 3–20 residues per second (Varenne et al., 1984). Co-translational scenario of the protein folding that starts on the ribosome (Gloge et al., 2014) also implicates polypeptide chain of the approximately closed loop's size. Specifically, the peptide exit tunnel (Nissen et al., 2000; Voss et al., 2006) of the 50S ribosomal subunit can accommodate a nascent chain of some 30–35 amino acids in extended conformation, apparently maintaining the environment for the initial formation of the “low entropy structures” (Kosolapov and Deutsch, 2009; O'Brien et al., 2011). On the experimental side, time resolved resonance excitation energy transfer (trFRET) implemented in the folding kinetics studies show that non-local interactions closing the loops are important for initiation of the folding transition in BPTI (Ittah and Haas, 1995) and *E. coli* adenilate kinase (Ben Ishay et al., 2012; Orevi et al., 2013, 2014).

7. Case study of simplified sequence proteins

Proteins with simplified sequences of the limited alphabet of amino acid types are characterized by the low sequence complexity (Clarke, 1995). While simplified sequences are mostly studied in the context of protein design, there are also different naturally occurring low complexity sequence regions, such as homopolymers, irregular combinations of two or a more residues, and regular/tandem repeats (Wootton, 1994). The theoretical prerequisite for a protein-like sequence that designates foldable structure is the existence of a sufficiently large stability gap between the energy of the native state and energies of the decoy conformations (Shakhnovich, 1998). There is an understanding and general agreement that homopolymers and sequences with very low complexity cannot be folded into the unique and stable structure. Of note, we consider in this section theoretical models that describe folding of protein sequences into individual single-domain structures. Intrinsically disordered protein chains that fold upon binding is a subject of separate studies (Berezovsky, 2011; Fong and Panchenko, 2010; Fong et al., 2012). The lower limit for the size of amino acid alphabet is still under discussion. It was proposed, for example, that the effective number of amino acid types must be larger than the number of expected conformations per residue in order to have a sufficient stability gap (Shakhnovich, 1998). The alternating polar and non-polar amino acids (Kamtekar et al., 1993) in certain symmetric folds do challenge the theoretical limits for the size of reduced amino acid alphabets, pointing to its possible dependence on the structure type.

The question about minimal number of residue types necessary for the protein to be foldable has been under experimental and theoretical scrutiny since several decades. Random libraries of sequences constructed by Davidson et al. from reduced amino acid alphabets is one of the first examples of the experiments on simplified sequence proteins. The most important result of Davidson's works is that structured proteins can be obtained from the libraries of protein sequences constructed from the three-letter amino acid code (Cordes et al., 1996; Davidson et al., 1995; Davidson and Sauer, 1994). The library of synthetic genes encoding 80–100 residues composed mainly of random combinations of glutamine, leucine, and arginine were expressed in *Escherichia coli*. Glutamine and leucine were putatively chosen on the basis of their hydrophobicity and hydrophilicity, respectively, and arginine – in order to improve protein solubility with the help of positively charged side chain. As a result, about one percent of the QLR-proteins were well expressed and characterized. Despite their high content of sequence randomness, these polypeptides possess high helical content observed in CD measurements. High helical content not only corroborated the importance of proper proportion (Goncearenco and Berezovsky, 2014) and placement (Berezovsky and Trifonov, 2001a; Berezovsky et al., 2015; Berezovsky et al., 2001) of hydrophobic and hydrophilic residues in the protein's sequence/structure, but also demonstrated for the first time that the three letter residue alphabet was sufficient to obtain stable folded polypeptides.

The SH3 domain was investigated by the phage-display technique on libraries constructed with a five-letter alphabet that included isoleucine, lysine, glutamic acid, alanine, and glycine (Riddle et al., 1997). This choice of residues was motivated by the conservatism of alanine and glycine, and the non-polar (isoleucine) and polar (lysine and glutamic acid) nature of amino acids that are crucial for the formation of the hydrophobic core and hydrophilic surface of a protein globule. Remarkably, the folding rates and stabilities of this oversimplified versions of SH3 protein resembled those of the wild type. Further NMR analysis revealed a well-packed core as a result of the selection procedure that eliminated molten globular structures in favor of the function. Baker's work on simplified sequences of the SH3 protein further corroborated that reduced alphabets of amino acids do encode complex protein topologies, providing the kinetic accessibility of the native state (Plaxco et al., 1998). Many other theoretical simplification schemes have been proposed to reduce the amino acid code. For example, amino acid alphabets were reduced by grouping amino acids according to their physicochemical properties, considering the pairwise residue-reside interaction matrices (Fan and Wang, 2003; Wang and Wang, 1999), and correlated mutations (Murphy et al., 2000). A direct evolution technique was used for design of the functional enzyme from a nine-letter amino acid alphabet (Jackel and Hilvert, 2010; Vamvaca et al., 2004; Walter et al., 2005). The obtained simplified protein was not only topologically equivalent to its natural analog, but it also resembled the typical molten globular behavior in terms of the decrease of stability coupled with the increase of structural flexibility.

The folding mechanism of a putatively primordial α/β protein with a low complexity sequence and relatively complex α/β topology was explored via molecular dynamics simulations (Guarnera et al., 2009). The protein studied is a 56-residue α/β structure, called ssG, which is a simplified version of the B1 domain of the protein G. In particular, the sequence alphabet of the ssG protein, consisting of only three residues, glycine, alanine, and threonine, was sufficient to provide conservation of the secondary structure pattern similar to that of the wild-type sequence. In addition to its mild β propensity, threonine was also chosen to counterbalance the hydrophobicity of alanine and the increased packing potential of glycine. The simplification of the wild-type sequence was achieved by placing alanine, threonine, and glycine in the α-helical, β-strand, and turn regions of the native crystal structure (PDB code 1pgb), respectively. As a result, the analyzed sequence is a combination of the four poly-T stretches with a long poly-A stretch punctuated by the four G-dipeptides: $T_9G_2T_9G_2A_{15}G_2T_8G_2T_8$. The goal of this study was multi-fold, and it is summarized in the following questions. Is three-letter amino acid alphabet sufficient to encode stable and complex α/β protein topology? How should the folded state of a hypothetical primordial protein look like? Is it possible to observe reversible folding for a

mid-size protein with low complexity sequence within the computationally accessible time scale? Does the decrease of the residue-residue interaction diversity induce an overall decrease of kinetic barriers in the protein free energy landscape, resulting in non-cooperative protein dynamics? The results on ssG protein were obtained from the 15-μs implicit solvent molecular dynamics simulation at 330 K from a fully extended polypeptide configuration as a starting conformation. The MD simulations were performed using CHARMM with the SASA implicit solvation model (Brooks et al., 2009). Multiple folding and unfolding events were observed along the 15-μs trajectory according to the fraction of native contacts and C_α-RMSD with the X-ray structure. The ssG protein assumes different folded topologies along the simulation trajectory with the fraction of native contacts between 0.6 and 0.9 and the C_α-RMSD with the native structure in the range between 2.5 and 5 Å. These values reflect a fluid-like phenomenology, suggesting that the folded state of the ssG protein is compatible with that of a molten globule (Fig. 4 A). The folding time scale of the ssG is within the order of 0.2 μs. On the other hand, three 1-μs simulations of the wild-type protein starting from the fully extended conformation did not achieve properly folded native state structure within this simulation time length, showing a smoother free energy landscape for the ssG protein in comparison to that of the natural protein G. This conclusion was further corroborated in another control 1-μs simulation (3 runs) performed on the wild-type protein (1pgb), which showed that protein G is also structurally stable on the 1-μs timescale. In particular, the root mean square fluctuations (RMSF) of the protein ssG, calculated using the portions of a trajectory where the protein is folded, showed significantly larger fluctuations of the simplified protein ssG compared to those of the protein G's native state (Fig. 4 C). The cluster analysis combined with the Markov State Modeling of the ssG's trajectory also showed that the most populated free energy basin (~22%) corresponds to the folded state, and it is characterized by the same secondary structure of the wild-type protein G (Fig. 4 B, D). Additionally, the folded basin includes protein conformers with one hairpin flipped, which indicates that slightly different topologies but same secondary structure can interconvert very rapidly within the folded free energy basin (Fig. 4 D). It was hypothesized, that this is a result of the strong bias of the putative primordial design that favors the secondary structure content of protein G rather than its tertiary contacts. The unfolded state appeared to be structurally very heterogeneous and composed of configurational states with diverse relative amounts of α-helical and β-sheet contents. In particular, α-helical rich regions of the configurational space play a role of the kinetic hub connecting different helical topologies. The β-sheet rich regions interconvert very slowly toward the folded state and, therefore, could promote protein aggregation. Dynamical correlation analysis of the secondary structure formation suggests that parallel routes characterize the folding pathways towards a molten globular state, which is consistent with a diffusion-collision mechanism (Karplus and Weaver, 1976). As a result, folding is achieved by the assembly of regular local elements of secondary structure that are energetically driven by the backbone-backbone

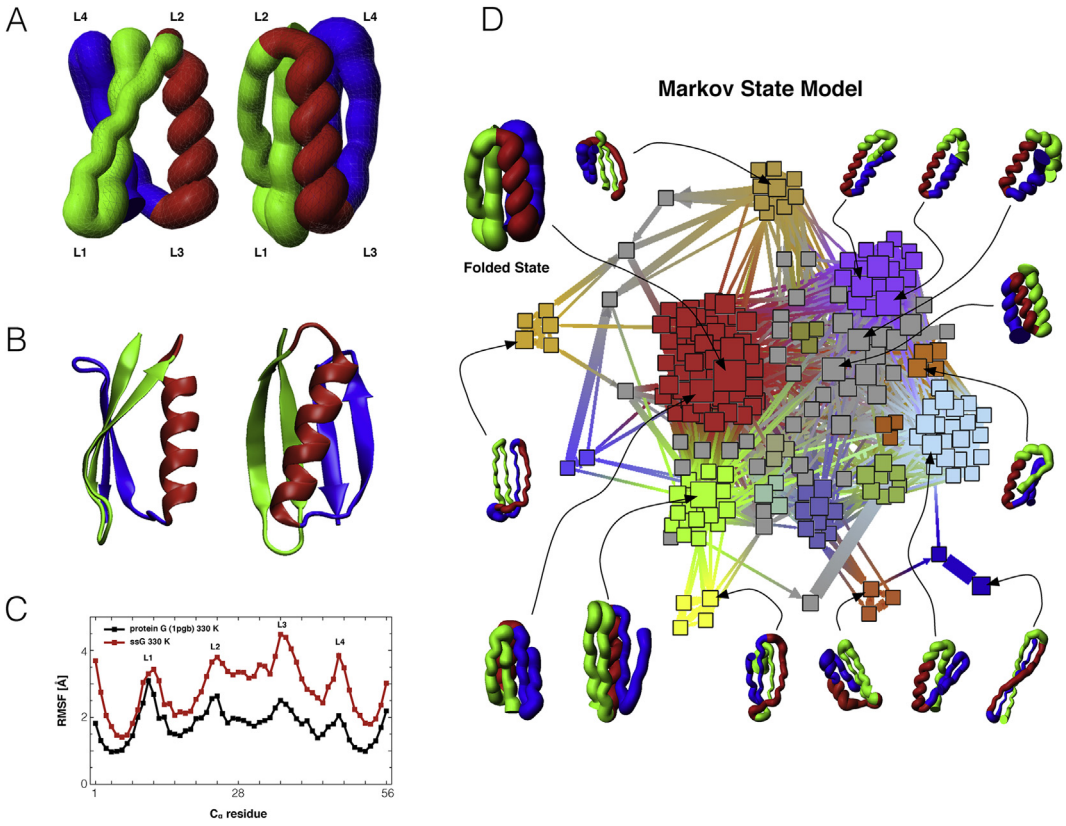

Fig. 4. Simplified-sequence protein ssG as a putative example of a primordial molten-globular folding. A. Comparison of the molten-globule state obtained from MD simulations of ssG protein and the X-ray structure of the protein G (**B**). The N-terminal β -hairpin, central α-helix, and C-terminal β -hairpin are in green, red, and blue, respectively. The tube-like rendering in (A) was generated using 100 snapshots from the most populated state. **C**. Comparison of the Cα root mean-square fluctuations (RMSF) of the ssG protein with the protein G (PDB ID: 1pgb) at 330K. **D**. Network representation of the Markov State Model (MSM) constructed from MD simulations of the ssG protein. MSM network contains 200 nodes and it is a synthetic representation of all possible pathways that the ssG protein can undertake. Each node corresponds to a configurational state of the protein, and its size is proportional to the state's population. Links in the network represent transition probabilities from one state-node to another. Nodes' colors are given according to the free energy basin of attraction, to which a specific node belongs. For instance, red is the color of the basin where a certain folded-state node can be found.

hydrogen bonding. Overall, the simulations studies of the ssG illustrated the tradeoff between large simplification schemes of the amino acid code and the corresponding loss of thermodynamical control (small energy gap) over the folded state (Guarnera et al., 2009), as well as a fluid-like (molten globular folded state) protein configurational space. The dramatic loss of stability in the folded state of a primordial-like protein does imply a highly disordered metastable dynamics, making the potentially dangerous aggregation-prone protein states to be kinetically accessible. This observation is consistent with the results of MD simulations on a coarse grained model that describes the phenomenology of amyloid aggregation. In the latter case, the intrinsic stability of the β-protected states (native states) favors the stability of the β-competent states (aggregation prone), dramatically increasing the propensity of the protein to form fiber-like aggregates (Pellarin and Caflisch, 2006; Pellarin et al., 2007).

8. Co-translational protein folding in crowded environment

We discussed here two basic scenarios of the protein folding. In the first case, the consecutive formation of closed loops (which are, presumably, insensitive to the secondary structure content) is followed by their arrangement in the final folds without paying much attention to times/mechanisms of the secondary structure formation. In the case of simplified proteins, the secondary structure stabilized by the hydrogen bonds plays an important role in the folding. These two scenarios complement each other, describing different scale of structural units ((i) loops: (Berezovsky et al., 2000c; Berezovsky and Trifonov, 2001a,b; Berezovsky et al., 2001) and (ii) secondary structure elements: (Eisenhaber et al., 1995)) and contacts (non-local interactions locking the loops (Berezovsky and Trifonov, 2001a; Berezovsky et al., 2001), secondary structure propensities (Berezovsky et al., 2015), and backbone-backbone hydrogen bonding (Berezovskii et al., 1998b; Rose et al., 2006)). A smooth transition between folding mechanisms is characterized by the adjustment of the balance between structural units of different scales and local-versus-global contacts (Daggett and Fersht, 2003). The unifying nucleation-condensation mechanism incorporates different folding pathways that continuously alter in the spectrum of the folding strategies with extremes represented by the hydrophobic collapse and the framework models (Daggett and Fersht, 2003; Karplus and Weaver, 1994).

Recent derivation of the predictor of α-helicity (on the basis of the amino acids' α-helical propensity) and analysis of the amino acid patterns in α-helices of natural proteins in relation to specific repertoire of aminoacyl-tRNA synthetases in the eukaryotic multiaminoacyl-tRNA synthetase (MARS) complex suggest that the very organization of MARS complex can be advantageous (Berezovsky et al., 2015) for the fast and efficient folding of the α–helical parts of proteins. While formation of the secondary structure is one of the determinants of the type of folding pathway (as it was pointed out in the previous paragraph), the latter can be also a result of many factors, such as the type of protein fold and its topology. It can also be influenced by the interactions with other cellular objects (Choi et al., 2013; Turner and Varshavsky, 2000; Gershenson and Gierasch, 2011; Ellis and Hartl, 1999), such as ribosome (Gloge et al., 2014; Nissen et al., 2000; Voss et al., 2006; Kosolapov and Deutsch, 2009; O'Brien et al., 2011), MARS complex (Berezovsky et al., 2015), and elements of the endoplasmic reticulum. Therefore, only proper consideration of co-translational folding within the framework of crowded cellular environment (Choi et al., 2013; Turner and Varshavsky, 2000; Gershenson and Gierasch, 2011; Ellis and Hartl, 1999) and interactions with all relevant objects (Berezovsky et al., 2015; Gloge et al., 2014; Nissen et al., 2000; Voss et al., 2006; Kosolapov and Deutsch, 2009; O'Brien et al., 2011) will allow one to finalize the folding scenario for each individual protein and to find the right balance between the involvement/role of the closed loops and secondary structure elements in the folding process.

9. From prebiotic ring-like peptides to elementary functional loops and contemporary enzymes

The final and the most challenging topic addressed in this review is the emergence and evolution of the protein function. While there are around 5000 currently known biochemical transformations (Bairoch, 2000) that process natural substrates and metabolites, the number of characterized mechanisms that provide these transformations is an order of magnitude less (Holliday et al., 2012). It is another order of magnitude reduction when chemical roles of the amino acid residues involved into catalysis, such as electron or protein donor/acceptor, electrostatic or hydrophobic stabilizer, activator etc., are considered (Holliday et al., 2005). As a result, thousands of multi-step biochemical transformations are actually built from the very small repertoire of elementary reactions, which is limited to several dozens (Holliday et al., 2012; Akiva et al., 2014; Andreini et al., 2009). Such striking reduction of functional complexity at the level of individual chemical reactions and residues' catalytic roles is in a fair agreement with the evolutionary scenario of the very emergence of enzymes from prebiotic ring-like peptides. These primitive protein-like molecules of the size of closed loops that possess one or few functional residues were presumably able to work in accord and to form primitive assemblies (Gazit, 2007) in the prebiotic world (Miller, 1987). They were eventually brought together by the fusion of corresponding ancient genes that were encoding them, forming the first functional folds/domains. Thus, closed loops with elementary functions, dubbed elementary functional loops (EFLs (Goncearenco and Berezovsky, 2010; 2011)), are the descendants of the ancient ring-like peptides that have eventually transformed into the returns of the protein backbones, bringing together the functional residues and forming active sites of domains/folds (Goncearenco and Berezovsky, 2012, 2015; Zheng et al., 2016).

Size distributions of the non-gapped functional signatures (Fig. 5) corroborate that closed loop can serve as a structural basis of the EFL. The length distribution of non-gapped Blocks (Pietrokovski et al., 1996) is shown in Fig. 5A. The length distributions for functional signature in CDD (Marchler-Bauer et al. (2015), Fig. 5B), Pfam (Finn et al. (2016), Fig. 5C), and PROSITE (Sigrist et al. (2013), Fig. 5D) were obtained by splitting multiples sequence alignment into non-gapped blocks (see figure legends for explanations). Preferential size of non-gapped functional signatures is about 15–20 residues, which together with segments of van der Waals locks (3–5 residues on each loop terminus) result in the closed loop's typical contour length. Noteworthy, back in 1969 Jacques Monod (according to Pierre-Gilles de Gennes (de Gennes, 1990; de Gennes, 1998)) envisioned that “it is of some interest to estimate the minimum size required for comparatively long loops of the polypeptide chain that linked together amino acids directly involved into the active site of a protein”. De Gennes estimated the minimal size of a loop and of a minimal functional domain (de Gennes, 1990; de Gennes, 1998), observing a fair agreement with numbers obtained for natural proteins and described in previous sections of this review (Gerstein, 1998; Jones et al., 1998; Wheelan et al., 2000; Berezovsky, 2003; Berezovsky et al., 2000c; Berezovsky and Trifonov, 2001a; Trifonov and Berezovsky, 2003). Current operational definition of the elementary functional loop (EFL) is (Goncearenco and Berezovsky, 2015): a structural-functional unit formed by the closed loop or return of the protein backbone, possessing functional residue(s) that perform a certain elementary

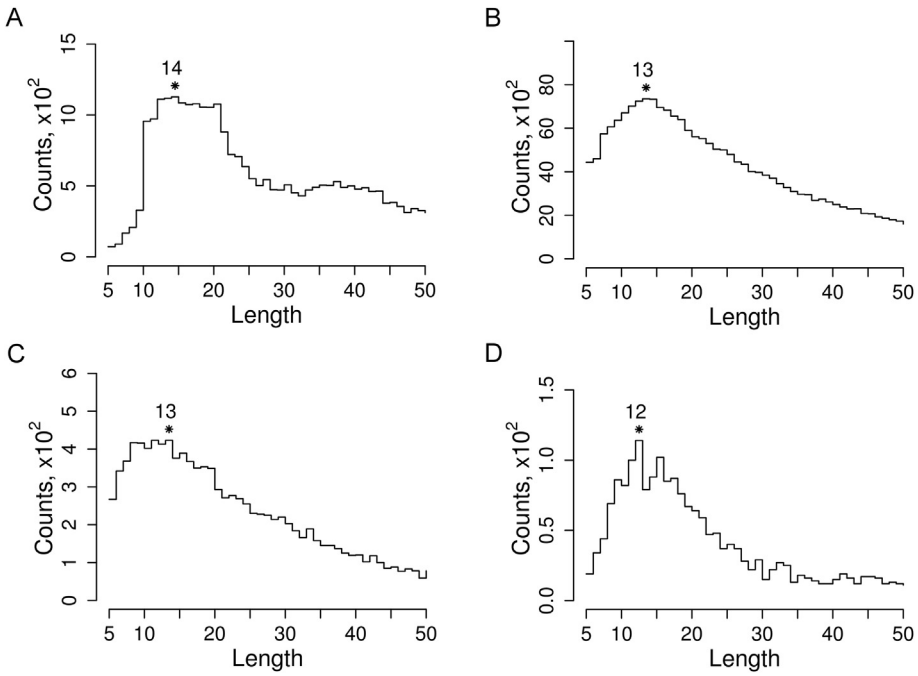

Fig. 5. Closed loop is a structural basis of the elementary functional loop (EFL). Length distributions of functional signatures show that they can be carried by the closed loops. *Methodology used for producing* Fig. 5. Data on the non-gapped functional signatures are obtained from CDD (52241 entities, chart B), Pfam (16295, chart C), Blocks (32125, chart A), and PROSITE (2416, chart D) databases. Distributions of the non-gapped Blocks's lengths are plotted directly. Multiple sequence alignments from CDD, Pfam, and PROSITE databases are split into the non-gapped blocks (single-residues gaps are allowed; in each block total number of small gaps in individual sequences should not exceed five per cent of the number of residues in the block). Length distributions of non-gapped blocks from CDD, Pfam, and PROSITE are plotted.

function (EF). The latter is an elementary reaction or binding interaction that provides the biochemical transformation or stabilization of the transition state. Elementary reaction (term E02035 in IUPAC Goldbook) has no intermediates, occurs in a single step and passes through a single transition state (McNaught, Wilkinson).

The first attempts to reconstruct evolutionary prototypes of elementary functions resulted in a small set of basic functional units (Berezovsky et al., 2003a), showing that globular protein can be "spelled" as words that are built from the letters of elementary functions (Berezovsky et al., 2003b). Later, a rigorous statistical approach for obtaining prototypes (Goncearenco and Berezovsky, 2010) and profiles (Goncearenco and Berezovsky, 2011) of elementary functions was developed. It is important to emphasize, that there is a fundamental difference between the ancient prototypes and the ancestral sequences that are obtained on the basis of phylogenetic relations. Ancient prototypes are entities that do not exist in modern proteins, they are represented by their descendants – EFLs. The ancestors typically correspond to one functional (super)family, having been represented in all or most of its members. Therefore, ancestors can be reconstructed by using a phylogenetic tree built from the alignments of related (super) family members and an evolutionary model with particular mutation and amino acid substitution rates (Cai et al., 2004; Harms and Thornton, 2010; Mirkin et al., 2003). The EFLs, on the contrary, are short sequence fragments belonging to the phylogenetically unrelated proteins from remote (super)families and even different protein folds (Goncearenco and Berezovsky, 2012, 2015), making it difficult to delineate hidden evolutionary connections. The specific nature of prototypes has led to developing a new computational approach for the prototypes' derivation (Goncearenco and Berezovsky, 2010, 2011), which is an iterative procedure outputting profile in form of position specific scoring matrix (PSSM). A scoring function weights profile positions according to their informational content expressed via Kullback-Leibler divergence, allowing thus to discriminate between the matches with specific signatures from the non-specific ones. Reshuffled profiles are used for the calculation of empirical distributions for the estimation of the scores' statistical significance. The elimination of redundancy and generalization of profiles is completed by the iterative hierarchical clustering of profiles (Goncearenco and Berezovsky, 2010).

The profile derivation procedure was used in subsequent works with the overall goal to explore the emergence of different functional folds/domains and evolutionary relations between them. First, the prototypes of EFLs were used for the analysis of distant evolutionary connections between the protein functions in archaeal kingdom (Goncearenco and Berezovsky, 2012). The presence of prototypes' descendants in different functional domains, as well as reutilization of EFLs and functional domains in the formation of multidomain structures and protein complexes were shown. Specific attention was paid to the methanogenesis pathway archetypal for Archaea. It was found that along with highly designable folds, such as β/α–barrel, Rossman, and ferredoxin, frequently employed in this pathway, new folds had emerged in response to demands on specific functions of the methanogenesis pathway (Goncearenco and Berezovsky, 2012). In both cases, certain EFLs were used as the building blocks of corresponding folds and functions. Involvement of different EFLs in one protein is exemplified by the formyl-methanofuran dehydrogenase (Fwd), work of the same EFL in different functional domains – by the heterosulfide reductase (Hdr), reutilization of both EFLs and functional domains – by the cofactor F_{430} binding in Methyl-coenzyme M reductase (Mcr) (Goncearenco and Berezovsky, 2012). The task for finding distant evolutionary connections that go beyond functional (super)families and folds was achieved in this work (Goncearenco and Berezovsky, 2012) by matching 525 profiles of elementary functions (derived on the sequences of archaeal proteomes) to sequences of archaeal clusters of orthologous genes –

arCOGs (Makarova et al., 2007). It was shown that arCOGs can be grouped around one or few elementary functions typical for the biochemical function of the group. Among the most representative group of functions are: amynoacyl-tRNA synthetases, transcriptional regulators, methylases/methyltransferases, ABC transporters, helicases.

Another comprehensive work on the evolution of protein function from its emergence to diversity in contemporary proteins was performed with the two-fold aim (Goncearenco and Berezovsky, 2015). First, the basic physics that lies in the foundation of the protein structure and function was discussed from the evolutionary perspective. In particular, it was surveyed how polymer nature of proteins and DNA work in accord, establishing the shape and size of basic units of proteins – closed loops (Berezovsky et al., 2000c) and typical size of protein folds/domains (Berezovsky, 2002; Berezovsky, 2003; Koczyk and Berezovsky, 2008). It was also pointed out that designability is another physics-based requirement (Zeldovich et al., 2006), which is crucial for: (i) structure stability and adaptation to extreme environments (Berezovsky et al., 2007) and (ii) for providing versatile structural scaffolds, which are able to adopt many different sequences that encode a wide diversity of functions (Zeldovich et al., 2006; Goncearenco and Berezovsky, 2012, 2015; Panchenko et al., 2005). Second, it was surveyed how contemporary proteins are built from the descendants of prebiotic ring-like peptides complemented by more specific EFLs emerged *en route* of evolution. Using the library of essential EFLs, intricate evolutionary relations were established between the different folds and functions. Fig. 6 contains an example of FAD/NAD-linked reductase (PDB ID: 1zmc, chain A), where elementary functional loops that work in binding/processing of FAD and NAD are shown. It also shows how the Gly-rich motif responsible for binding phosphate moiety in dinucleotide-containing ligands works in different protein superfamilies and folds (see the figure legend for details). The rules that govern the evolution of protein function learnt from the nature and described in this work (Goncearenco and Berezovsky, 2015) provided a foundation for introducing the generalized sequence-structure descriptor of the elementary function. The descriptor of elementary function is a supposed building block for future design of desired enzymatic functions (Goncearenco and Berezovsky, 2015).

On a side of practical implications, the NBDB database (http://nbdb.bii.a-star.edu.sg) motivated by the importance of nucleotide-containing ligands and other biologically relevant cofactors/coenzymes was recently created (Zheng et al., 2016). It contains detailed information on 249 EFLs involved into interactions with different chemical parts (nitrogen bases, phosphate groups, ribose sugar and other moieties such as flavin and nicotinamide) of 24 ligands/cofactors (ATP, AMP, ATP, GMP, GDP, GTP, CTP, PAP, PPS, FMN, FAD(H), NAD(H), NADP, cAMP, cGMP, c-di-AMP, and c-di-GMP, ThPP, THD, F-420, ACO, CoA, PLP, and SAM). Sequence profiles of the EFL motifs were derived *de novo* from the non-redundant UniProt (2008) proteome sequences. Each EFL profile in the database is characterized by the pattern of corresponding ligand–protein interactions found in crystallized ligand–protein complexes. A search routine allows one to detect fragments that match to profiles of particular EFLs in the protein sequence provided by the user. EF-Patterns routine for annotation/prediction of protein function as a combination of elementary functions on the basis of 294 profiles of EFLs derived in Goncearenco and Berezovsky (2015) is included in the ANNOTATOR software environment (Eisenhaber et al., 2016). Using the profile-derivation procedure, the key functional fragments in Zinc transporter (Lasry et al., 2012), ZnT-2 (SLC30A2) were derived, unraveling the conserved signature important for transporter's dimerization that provides its activity (Lasry et al., 2012). Table 1 summarizes relevant web resources discussed and/or developed

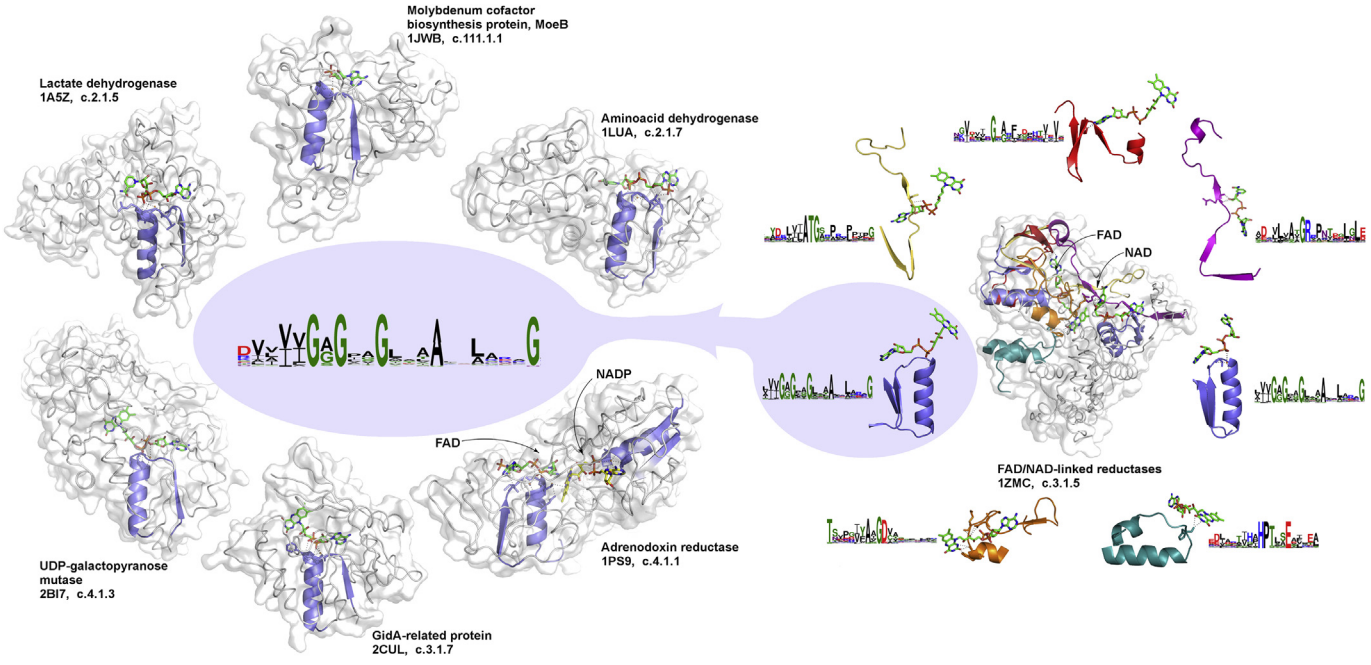

Fig. 6. Elementary functional loops are building blocks of the protein enzymatic function. Human dihydrolipoamide dehydrogenase (PDB ID: 1zmc, chain A; FAD/NAD(P)-binding domain fold) exemplifies combination of elementary functional loops that interact with two dinucleotide-containing ligands: flavin adenine dinucleotide (FAD) and nicotine adenine dinucleotide (NAD). Glycine-rich motif (with a characteristic signature GxGxxG) provides binding of the phosphate moiety in dinucleotide-containing ligands. Set of structures on the left side shows that this signature is present in different functional superfamilies and folds: c.111.1.1 is activating enzymes of the ubiquitin-like proteins fold (1jwb); c.2.1.5 and c.2.1.7 – NAD(P)-binding Rossmann-fold (1a5z and 1lua); c.4.1.1 and c.4.1.3 – nucleotide-binding domain (1ps9 and 2bi7); c.3.1.7 – FAD/NAD(P)-binding domain (2cul).

Table 1
Web resources used and recommended in the review.

Resource, reference, URL	Description
DHcL (Koczyk and Berezovsky, 2008) http://cropnet.pl/dhcl/	Identifies domain structures at different levels of energy hierarchy and elements of the loop-n-lock structure, closed loops and van der Waals locks
CDD (Marchler-Bauer et al., 2015) http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	A collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins
Pfam (Finn et al., 2016) http://pfam.xfam.org/	A large collection of protein families, each represented by multiple sequence alignments and hidden Markov models
PROSITE (Sigrist et al., 2013) http://prosite.expasy.org/	Describes protein domains, families and functional sites as well as associated patterns and profiles to identify them
Blocks (Pietrokovski et al., 1996) http://blocks.fhcrc.org/	Multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins
NBDB (Zheng et al., 2016) http://nbdb.bii.a-star.edu.sg/	A collection of profiles of Elementary Functional Loops involved in binding of nucleotide-containing ligands and biologically relevant cofactors/coenzymes
UniProt (UniProt, 2008) http://www.uniprot.org	A stable, comprehensive, central resource on protein sequences and functional annotation
ANNOTATOR (Eisenhaber et al., 2016) http://annotator.bii.a-star.edu.sg	An integration of tools for protein sequence analysis

in the works reviewed here.

The concept of EFLs as basic units of the enzymatic function is gaining support and is being actively implemented in a current research on evolution of protein structure and function. For example, it was shown that ancient fingerprint of the ribose binding is an indicator of the common ancestry of Rossmann-fold enzymes that utilize different ribose-containing cofactors (Laurino et al., 2016). On the basis of elementary functional loops, Caetano-Anolles et al. have recently built a model of the emergence and early history of molecular functions (Aziz et al., 2016). In some cases, closed loops and elementary functional loop are being serendipitously rediscovered (Alva et al., 2015), and their prototypes/profiles derived with slightly different procedures resemble those derived earlier (Goncearenco and Berezovsky, 2010, 2011, 2012, 2015). Specifically, "observable remnants of a primordial RNA-peptide world" observed in (Alva et al., 2015) yield a median size 24 amino acid residues and, in most of the cases, closed loop-like shape. Domination of the iron-sulfur- and nucleic acid-binding elementary function observed for 40 fragments (Alva et al., 2015) agrees with the earlier described abundance of "binding and metabolic processing of the metal- and nucleotide-containing cofactors and ligands" among the ancient elementary functions (Goncearenco and Berezovsky, 2010, 2011, 2012, 2015). Besides the evolutionary importance of nucleotide-base cofactors that are presumed to have preceded proteins (Laurino et al., 2016), they are crucial for various biochemical transformations taking place in the living cell (Holliday et al., 2012).

Many works on protein modularity and evolution of function via recombination of the modules provide additional, indirect evidences of the role of closed loops and EFLs. We will give here only one example for closed loops and functional motifs, respectively. First, it was shown, on the basis of sequence analysis, that $(\alpha\beta)_2$ structural motif is seemingly the minimal discernable unit that connects the families of the $(\beta\alpha)_8$ barrel and flavodoxin folds (Farias-Rico et al., 2014). Second, the comparison of signature-based active sites' profiles delineate the molecular functional details more accurately than the whole-sequence and structure-based comparisons (Leuthaeuser et al., 2015). Protein modularity on the basis of loop-like elements lies in the foundation of many current design approaches. Khersonsky and Fleishman reviewed recent design efforts (Khersonsky and Fleishman, 2016) advocating the synthetic strategy of building yet unsampled proteins from the fragments of natural proteins recombined and optimized in computational design calculations (Khersonsky and Fleishman, 2016). In another work, the space of folded structures was explored by generating the tandem repeats of a simple helix-loop-helix-loop structural motif (Brunette et al., 2015). The SEWING computational protocol allows one to *de novo* design proteins, using structural motifs of natural proteins (Jacobs et al., 2016).

10. Concluding remarks

Though first databases (Koczyk and Berezovsky, 2008; Zheng et al., 2016) and web-servers (Koczyk and Berezovsky, 2008; Eisenhaber et al., 2016) for the analysis and investigation of closed loops and elementary functional loops are already available, the exhaustive catalogs with detail description of all relevant characteristics are yet to be produced. The catalog will be instrumental in the high-throughput annotation/prediction of protein function on the basis of its building blocks – elementary functions (Goncearenco and Berezovsky, 2015; Zheng et al., 2016; Eisenhaber et al., 2016). Further development of the concept of descriptor of elementary function (Goncearenco and Berezovsky, 2015; Zheng et al., 2016), construction of the comprehensive library of descriptors, and implementation of the computational protocol for descriptor-based design of required catalytic functions are also the first-priority, future tasks.

References

Aharoni, A., Gaidukov, L., Khersonsky, O., Mc, Q.G.S., Roodveldt, C., Tawfik, D.S., 2005. The 'evolvability' of promiscuous protein functions. Nat. Genet. 37, 73–76.
Akiva, E., Brown, S., Almonacid, D.E., Barber 2nd, A.E., Custer, A.F., Hicks, M.A., Huang, C.C., Lauck, F., Mashiyama, S.T., Meng, E.C., et al., 2014. The structure-function linkage database. Nucleic Acids Res. 42, D521–D530.
Alden, K., Veretnik, S., Bourne, P.E., 2010. dConsensus: a tool for displaying domain assignments by multiple structure-based algorithms and for construction of a consensus assignment. BMC Bioinforma. 11, 310.
Alva, V., Soding, J., Lupas, A.N., 2015. A vocabulary of ancient peptides at the origin of folded proteins. Elife 4, e09410.
Andreini, C., Bertini, I., Cavallaro, G., Holliday, G.L., Thornton, J.M., 2009. MetalMACiE: a database of metals involved in biological catalysis. Bioinformatics 25, 2088–2089.
Angelov, B., Sadoc, J.F., Jullien, R., Soyer, A., Mornon, J.P., Chomilier, J., 2002. Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. Proteins 49, 446–456.
Aziz, M.F., Caetano-Anolles, K., Caetano-Anolles, G., 2016. The early history and emergence of molecular functions and modular scale-free network behavior. Sci. Rep. 6, 25058.
Bairoch, A., 2000. The ENZYME database in 2000. Nucleic Acids Res. 28, 304–305.
Ben Ishay, E., Rahamim, G., Orevi, T., Hazan, G., Amir, D., 2012. Haas E: fast subdomain folding prior to the global refolding transition of E. coli adenylate kinase: a double kinetics study. J. Mol. Biol. 423, 613–623.
Berezovskii, I.N., Esipova, N.G., Tumanian, V.G., 1997. Isolation of the energy-significant parts of the globule and the hierarchy of the domain structure of the protine macromolecule. Biophysics 42, 557–565.

Berezovskii, I.N., Esipova, N.G., Tumanian, V.G., Namiot, V.A., 1998a. A new approach to calculating van der Waals interaction energies in protein mcromolecules. Dielectric constant as a physical parameter. Biophysics 43, 909–916.
Berezovskii, I.N., Esipova, N.G., Tumanian, V.G., 1998b. Spatial Distribution of directed interactions in globular protein structures. Biophysics 43, 367–377.
Berezovsky, I.N., 2003. Discrete structure of van der Waals domains in globular proteins. Protein Eng. 16, 161–167.
Berezovsky, I.N., 2011. The diversity of physical forces and mechanisms in intermolecular interactions. Phys. Biol. 8, 035002.
Berezovsky, I.N., Shakhnovich, E.I., 2005. Physics and evolution of thermophilic adaptation. Proc. Natl. Acad. Sci. U. S. A. 102, 12742–12747.
Berezovsky, I.N., Trifonov, E.N., 2001a. Van der Waals locks: loop-n-lock structure of globular proteins. J. Mol. Biol. 307, 1419–1426.
Berezovsky, I.N., Trifonov, E.N., 2001b. Loop fold nature of globular proteins. Protein Eng. 14, 403–407.
Berezovsky, I.N., Trifonov, E.N., 2001c. Protein structure and folding: a new start. J. Biomol. Struct. Dyn. 19, 397–403.
Berezovsky, I.N., Trifonov, E.N., 2002a. Flowering buds of globular proteins: transpiring simplicity of protein organization. Comp. Funct. Genomics 3, 525–534.
Berezovsky, I.N., Trifonov, E.N., 2002b. Loop fold structure of proteins: resolution of Levinthals paradox. J. Biomol. Struct. Dyn. 20, 5–6.
Berezovsky, I.N., Tumanyan, V.G., Esipova, N.G., 1997. Representation of amino acid sequences in terms of interaction energy in protein globules. FEBS Lett. 418, 43–46.
Berezovsky, I.N., Namiot, V.A., Tumanyan, V.G., Esipova, N.G., 1999. Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition. J. Biomol. Struct. Dyn. 17, 133–155.
Berezovsky, I.N., Esipova, N.G., Tumanyan, V.G., 2000a. Hierarchy of regions of amino acid sequence with respect to their role in the protein spatial structure. J. Comput. Biol. 7, 183–192.
Berezovsky, I.N., Esipova, N.G., Tumanyan, V.G., Namiot, V.A., 2000b. A new approach for the calculation of the energy of van der Waals interactions in macromolecules of globular proteins. J. Biomol. Struct. Dyn. 17, 799–809.
Berezovsky, I.N., Grosberg, A.Y., Trifonov, E.N., 2000c. Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett. 466, 283–286.
Berezovsky, I.N., Kirzhner, V.M., Kirzhner, A., Trifonov, E.N., 2001. Protein folding: looping from hydrophobic nuclei. Proteins 45, 346–350.
Berezovsky, I.N., 2002. Protein structure: chapters which have been missed. In: Gromiha, M.M., Selvaraj, S. (Eds.), Recent Research Developments in Protein Folding, Stability and Design. research Signpost, pp. 1–23.
Berezovsky, I.N., Kirzhner, V.M., Kirzhner, A., Rosenfeld, V.R., Trifonov, E.N., 2002. Closed loops: persistence of the protein chain returns. Protein Eng. 15, 955–957.
Berezovsky, I.N., Kirzhner, A., Kirzhner, V.M., Rosenfeld, V.R., Trifonov, E.N., 2003a. Protein sequences yield a proteomic code. J. Biomol. Struct. Dyn. 21, 317–325.
Berezovsky, I.N., Kirzhner, A., Kirzhner, V.M., Trifonov, E.N., 2003b. Spelling protein structure. J. Biomol. Struct. Dyn. 21, 327–339.
Berezovsky, I.N., Chen, W.W., Choi, P.J., Shakhnovich, E.I., 2005. Entropic stabilization of proteins and its proteomic consequences. PLoS Comput. Biol. 1, e47.
Berezovsky, I.N., Zeldovich, K.B., Shakhnovich, E.I., 2007. Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput. Biol. 3, e52.
Berezovsky, I.N., Zheng, Z., Kurotani, A., Tokmakov, A.A., Kurochkin, I.V., 2015. Organization of the multiaminoacyl-tRNA synthetase complex and the cotranslational protein folding. Protein Sci. 24, 1475–1485.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein Data Bank. Nucleic Acids Res. 28, 235–242.
Bresler, S.E., Talmud, D.L., 1944a. On the nature of globular proteins. Compt Rend. Acad. Sci. URSS 43, 310–314.
Bresler, S.E., Talmud, D.L., 1944b. On the nature of globular proteins. II A few consequences of the new hypothesis. Compt Rend. Acad. Sci. URSS 43, 349–350.
Brooks, B.R., Brooks 3rd, C.L., Mackerell Jr., A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al., 2009. CHARMM: the biomolecular simulation program. J. Comput. Chem. 30, 1545–1614.
Brunette, T.J., Parmeggiani, F., Huang, P.S., Bhabha, G., Ekiert, D.C., Tsutakawa, S.E., Hura, G.L., Tainer, J.A., Baker, D., 2015. Exploring the repeat protein universe through computational protein design. Nature 528, 580–584.
Cai, W., Pei, J., Grishin, N.V., 2004. Reconstruction of ancestral protein sequences and its applications. BMC Evol. Biol. 4, 33.
Chintapalli, S.V., Yew, B.K., Illingworth, C.J., Upton, G.J., Reeves, P.J., Parkes, K.E., Snell, C.R., Reynolds, C.A., 2010. Closed loop folding units from structural alignments: experimental foldons revisited. J. Comput. Chem. 31, 2689–2701.
Chintapalli, S.V., Illingworth, C.J., Upton, G.J., Sacquin-Mora, S., Reeves, P.J., Mohammedali, H.S., Reynolds, C.A., 2014. Assessing the effect of dynamics on the closed-loop protein-folding hypothesis. J. R. Soc. Interface 11, 20130935.
Choi, S.I., Kwon, S., Son, A., Jeong, H., Kim, K.H., Seong, B.L., 2013. Protein folding in vivo revisited. Curr. Protein Pept. Sci. 14, 721–733.
Chothia, C., Gough, J., Vogel, C., Teichmann, S.A., 2003. Evolution of the protein repertoire. Science 300, 1701–1703.
Clarke, N.D., 1995. Sequence 'minimization': exploring the sequence landscape with simplified sequences. Curr. Opin. Biotechnol. 6, 467–472.
Cordes, M.H., Davidson, A.R., Sauer, R.T., 1996. Sequence space, folding and protein design. Curr. Opin. Struct. Biol. 6, 3–10.
Daggett, V., Fersht, A.R., 2003. Is there a unifying mechanism for protein folding? Trends Biochem. Sci. 28, 18–25.
Davidson, A.R., Sauer, R.T., 1994. Folded proteins occur frequently in libraries of random amino acid sequences. Proc. Natl. Acad. Sci. U. S. A. 91, 2146–2150.
Davidson, A.R., Lumb, K.J., Sauer, R.T., 1995. Cooperatively folded proteins in random sequence libraries. Nat. Struct. Biol. 2, 856–864.
de Gennes, P.G., 1998. In: de Gennes, P.G. (Ed.), Simple Views on Condensed Matter, vol. 8. World Scientific, Singapore, New Jersey, London, Hong Kong.
de Gennes, P.G., 1990. In: di Brozolo, L.A.R. (Ed.), Introduction to Polymer Dynamics. Cambridge University Press, Cambridge.
Dill, K.A., 1990. Dominant forces in protein folding. Biochemistry 29, 7133–7155.
Dokholyan, N.V., Shakhnovich, E.I., 2001. Understanding hierarchical protein evolution from first principles. J. Mol. Biol. 312, 289–307.
Eisenhaber, F., Persson, B., Argos, P., 1995. Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. Crit. Rev. Biochem. Mol. Biol. 30, 1–94.
Eisenhaber, B., Kuchibhatla, D., Sherman, W., Sirota, F.L., Berezovsky, I.N., Wong, W.C., Eisenhaber, F., 2016. The recipe for protein sequence-based function prediction and its implementation in the ANNOTATOR software environment. Methods Mol. Biol. 1415, 477–506.
Ellis, R.J., Hartl, F.U., 1999. Principles of protein folding in the cellular environment. Curr. Opin. Struct. Biol. 9, 102–110.
England, J.L., Shakhnovich, E.I., 2003. Structural determinant of protein designability. Phys. Rev. Lett. 90, 218101.
Fan, K., Wang, W., 2003. What is the minimum number of letters required to fold a protein? J. Mol. Biol. 328, 921–926.
Farias-Rico, J.A., Schmidt, S., Hocker, B., 2014. Evolutionary relationship of two ancient protein superfolds. Nat. Chem. Biol. 10, 710–715.
Fernandez-Fuentes, N., Fiser, A., 2013. A modular perspective of protein structures: application to fragment based loop modeling. Methods Mol. Biol. 932, 141–158.
Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285.
Fong, J.H., Panchenko, A.R., 2010. Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. Mol. Biosyst. 6, 1821–1828.
Fong, J.H., Shoemaker, B.A., Panchenko, A.R., 2012. Intrinsic protein disorder in human pathways. Mol. Biosyst. 8, 320–326.
Gazit, E., 2007. Self-assembled peptide nanostructures: the design of molecular building blocks and their technological utilization. Chem. Soc. Rev. 36, 1263–1269.
Gershenson, A., Gierasch, L.M., 2011. Protein folding in the cell: challenges and progress. Curr. Opin. Struct. Biol. 21, 32–41.
Gerstein, M., 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold. Des. 3, 497–512.
Glasner, M.E., Gerlt, J.A., Babbitt, P.C., 2006. Evolution of enzyme superfamilies. Curr. Opin. Chem. Biol. 10, 492–497.
Gloge, F., Becker, A.H., Kramer, G., Bukau, B., 2014. Co-translational mechanisms of protein maturation. Curr. Opin. Struct. Biol. 24, 24–33.
Goncearenco, A., Berezovsky, I.N., 2010. Prototypes of elementary functional loops unravel evolutionary connections between protein functions. Bioinformatics 26, i497–503.
Goncearenco, A., Berezovsky, I.N., 2011. Computational reconstruction of primordial prototypes of elementary functional loops in modern proteins. Bioinformatics 27, 2368–2375.
Goncearenco, A., Berezovsky, I.N., 2012. Exploring the evolution of protein function in Archaea. BMC Evol. Biol. 12, 75.
Goncearenco, A., Berezovsky, I.N., 2014. The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. Biol. Direct 9, 29.
Goncearenco, A., Berezovsky, I.N., 2015. Protein function from its emergence to diversity in contemporary proteins. Phys. Biol. 12, 045002.
Goncearenco, A., Ma, B.G., Berezovsky, I.N., 2014. Molecular mechanisms of adaptation emerging from the physics and evolution of nucleic acids and proteins. Nucleic Acids Res. 42, 2879–2892.
Goryshin, I., Kil, Y.V., Reznikoff, W.S., 1994. DNA length, bending, and twisting constraints on IS50 transposition. Proc. Natl. Acad. Sci. U. S. A. 91, 10834–10838.
Gromiha, M.M., 2009. Multiple contact network is a key determinant to protein folding rates. J. Chem. Inf. Model 49, 1130–1135.
Gromiha, M.M., Selvaraj, S., 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J. Mol. Biol. 310, 27–32.
Gromiha, M.M., Uedaira, H., An, J., Selvaraj, S., Prabakaran, P., Sarai, A., 2002. ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0. Nucleic Acids Res. 30, 301–302.
Gromiha, M.M., Pathak, M.C., Saraboji, K., Ortlund, E.A., Gaucher, E.A., 2013. Hydrophobic environment is a key factor for the stability of thermophilic proteins. Proteins 81, 715–721.
Grosberg, A.Y., 2016. Ensemble view of RNAs and proteins: loops, knots, territories, and evolution. Biophys. J. 110, 2289–2290.
Guarnera, E., Pellarin, R., Caflisch, A., 2009. How does a simplified-sequence protein fold? Biophys. J. 97, 1737–1746.
Harms, M.J., Thornton, J.W., 2010. Analyzing protein structure and function using ancestral gene reconstruction. Curr. Opin. Struct. Biol. 20, 360–366.
Hleap, J.S., Blouin, C., 2016. The semantics of the modular architecture of protein structures. Curr. Protein Pept. Sci. 17, 62–71.
Holland, T.A., Veretnik, S., Shindyalov, I.N., Bourne, P.E., 2006. Partitioning protein

structures into domains: why is it so difficult? J. Mol. Biol. 361, 562–590.
Holliday, G.L., Bartlett, G.J., Almonacid, D.E., O'Boyle, N.M., Murray-Rust, P., Thornton, J.M., Mitchell, J.B., 2005. MACiE: a database of enzyme reaction mechanisms. Bioinformatics 21, 4315–4316.
Holliday, G.L., Andreini, C., Fischer, J.D., Rahman, S.A., Almonacid, D.E., Williams, S.T., Pearson, W.R., 2012. MACiE: exploring the diversity of biochemical reactions. Nucleic Acids Res. 40, D783–D789.
Ittah, V., Haas, E., 1995. Nonlocal interactions stabilize long range loops in the initial folding intermediates of reduced bovine pancreatic trypsin inhibitor. Biochemistry 34, 4493–4506.
Jackel, C., Hilvert, D., 2010. Biocatalysts by evolution. Curr. Opin. Biotechnol. 21, 753–759.
Jacobs, T.M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J.F., Szyperski, T., Kuhlman, B., 2016. Design of structurally distinct proteins using strategies inspired by evolution. Science 352, 687–690.
Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C., Thornton, J.M., 1998. Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci. 7, 233–242.
Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., Hecht, M.H., 1993. Protein design by binary patterning of polar and nonpolar amino acids. Science 262, 1680–1685.
Karplus, M., Weaver, D.L., 1976. Protein-folding dynamics. Nature 260, 404–406.
Karplus, M., Weaver, D.L., 1994. Protein folding dynamics: the diffusion-collision model and experimental data. Protein Sci. 3, 650–668.
Kelley, L.A., Sternberg, M.J., 2015. Partial protein domains: evolutionary insights and bioinformatics challenges. Genome Biol. 16, 100.
Khersonsky, O., Fleishman, S.J., 2016. Why reinvent the wheel? Building new proteins based on ready-made parts. Protein Sci. 25, 1179–1187.
Koczyk, G., Berezovsky, I.N., 2008. Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure. Nucleic Acids Res. 36, W239–W245.
Kolinski, A., Skolnick, J., Godzik, A., Hu, W.P., 1997. A method for the prediction of surface "U"-turns and transglobular connections in small proteins. Proteins 27, 290–308.
Koonin, E.V., Tatusov, R.L., Galperin, M.Y., 1998. Beyond complete genomes: from sequence to structure and function. Curr. Opin. Struct. Biol. 8, 355–363.
Koonin, E.V., Wolf, Y.I., Karev, G.P., 2002. The structure of the protein universe and genome evolution. Nature 420, 218–223.
Kosolapov, A., Deutsch, C., 2009. Tertiary interactions within the ribosomal exit tunnel. Nat. Struct. Mol. Biol. 16, 405–411.
Kumar, M.D., Bava, K.A., Gromiha, M.M., Prabakaran, P., Kitajima, K., Uedaira, H., Sarai, A., 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 34, D204–D206.
Kwasigroch, J.M., Chomilier, J., Mornon, J.P., 1996. A global taxonomy of loops in globular proteins. J. Mol. Biol. 259, 855–872.
Lamarine, M., Mornon, J.P., Berezovsky, I.N., Chomilier, J., 2001. Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? Cell Mol. Life Sci. 58, 492–498.
Lasry, I., Seo, Y.A., Ityel, H., Shalva, N., Pode-Shakked, B., Glaser, F., Berman, B., Berezovsky, I., Goncearenco, A., Klar, A., et al., 2012. A dominant negative heterozygous G87R mutation in the zinc transporter, ZnT-2 (SLC30A2), results in transient neonatal zinc deficiency. J. Biol. Chem. 287, 29348–29361.
Laurino, P., Toth-Petroczy, A., Meana-Paneda, R., Lin, W., Truhlar, D.G., Tawfik, D.S., 2016. An ancient fingerprint indicates the common ancestry of rossmann-fold enzymes utilizing different ribose-based cofactors. PLoS Biol. 14, e1002396.
Leszczynski, J.F., Rose, G.D., 1986. Loops in globular proteins: a novel category of secondary structure. Science 234, 849–855.
Leuthaeuser, J.B., Knutson, S.T., Kumar, K., Babbitt, P.C., Fetrow, J.S., 2015. Comparison of topological clustering within protein networks using edge metrics that evaluate full sequence, full structure, and active site microenvironment similarity. Protein Sci. 24, 1423–1439.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.
Liu, L., Hyeon, C., 2016. Contact statistics highlight distinct organizing principles of proteins and RNA. Biophys. J. 110, 2320–2327.
Ma, B.G., Goncearenco, A., Berezovsky, I.N., 2010. Thermophilic adaptation of protein complexes inferred from proteomic homology modeling. Structure 18, 819–828.
Makarova, K.S., Sorokin, A.V., Novichkov, P.S., Wolf, Y.I., Koonin, E.V., 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol. Direct 2, 33.
Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al., 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43, D222–D226.
Martin, A.C., Toda, K., Stirk, H.J., Thornton, J.M., 1995. Long loops in proteins. Protein Eng. 8, 1093–1101.
McNaught AD, Wilkinson A: I{UPAC}. Compendium of Chemical Terminology, 2nd Ed. (The "Gold Book"): WileyBlackwell; 2nd Revised edition edition.
Miller, S.L., 1987. Which organic compounds could have occurred on the prebiotic earth? Cold Spring Harb. Symp. Quant. Biol. 52, 17–27.
Mirkin, B.G., Fenner, T.I., Galperin, M.Y., Koonin, E.V., 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. 3, 2.
Moret, M.A., Zebende, G.F., 2007. Amino acid hydrophobicity and accessible surface area. Phys. Rev. E Stat. Nonlin Soft Matter Phys. 75, 011920.
Murphy, L.R., Wallqvist, A., Levy, R.M., 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng. 13, 149–152.
Nissen, P., Hansen, J., Ban, N., Moore, P.B., Steitz, T.A., 2000. The structural basis of ribosome activity in peptide bond synthesis. Science 289, 920–930.
O'Brien, E.P., Christodoulou, J., Vendruscolo, M., Dobson, C.M., 2011. New scenarios of protein folding can occur on the ribosome. J. Am. Chem. Soc. 133, 513–526.
Ochoa, A., Storey, J.D., Llinas, M., Singh, M., 2015. Beyond the E-value: stratified statistics for protein domain prediction. PLoS Comput. Biol. 11, e1004509.
Oliva, B., Bates, P.A., Querol, E., Aviles, F.X., Sternberg, M.J., 1997. An automated classification of the structure of protein loops. J. Mol. Biol. 266, 814–830.
Orevi, T., Rahamim, G., Hazan, G., Amir, D., Haas, E., 2013. The loop hypothesis: contribution of early formed specific non-local interactions to determination of protein folding pathways. Biophys. Rev. 5, 85–98.
Orevi, T., Rahamim, G., Shemesh, S., Ben Ishay, E., Amir, D., Haas, E., 2014. Fast closure of long loops at the initiation of the folding transition of globular proteins studied by time-resolved FRET-based methods. Bio-Algorithms Med-Systems 10, 169–193.
Pace, C.N., Shirley, B.A., McNutt, M., Gajiwala, K., 1996. Forces contributing to the conformational stability of proteins. FASEB J. 10, 75–83.
Panchenko, A.R., Madej, T., 2004. Analysis of protein homology by assessing the (dis)similarity in protein loop regions. Proteins 57, 539–547.
Panchenko, A.R., Madej, T., 2005. Structural similarity of loops in protein families: toward the understanding of protein evolution. BMC Evol. Biol. 5, 10.
Panchenko, A.R., Luthey-Schulten, Z., Wolynes, P.G., 1996. Foldons, protein structural modules, and exons. Proc. Natl. Acad. Sci. U. S. A. 93, 2008–2013.
Panchenko, A.R., Luthey-Schulten, Z., Cole, R., Wolynes, P.G., 1997. The foldon universe: a survey of structural similarity and self-recognition of independently folding units. J. Mol. Biol. 272, 95–105.
Panchenko, A.R., Wolf, Y.I., Panchenko, L.A., Madej, T., 2005. Evolutionary plasticity of protein families: coupling between sequence and structure variation. Proteins 61, 535–544.
Papandreou, N., Berezovsky, I.N., Lopes, A., Eliopoulos, E., Chomilier, J., 2004. Universal positions in globular proteins. Eur. J. Biochem. 271, 4762–4768.
Pellarin, R., Caflisch, A., 2006. Interpreting the aggregation kinetics of amyloid peptides. J. Mol. Biol. 360, 882–892.
Pellarin, R., Guarnera, E., Caflisch, A., 2007. Pathways and intermediates of amyloid fibril formation. J. Mol. Biol. 374, 917–924.
Pietrokovski, S., Henikoff, J.G., Henikoff, S., 1996. The Blocks database–a system for protein classification. Nucleic Acids Res. 24, 197–200.
Plaxco, K.W., Riddle, D.S., Grantcharova, V., Baker, D., 1998. Simplified proteins: minimalist solutions to the 'protein folding problem'. Curr. Opin. Struct. Biol. 8, 80–85.
Ponnuswamy, P.K., Gromiha, M.M., 1994. On the conformational stability of folded proteins. J. Theor. Biol. 166, 63–74.
Protasevich, I.I., Platonov, A.L., Pavlovsky, A.G., 1987. Esipova NG: distribution of charges in Bacillus intermedius 7P ribonuclease determines the number of cooperatively melting regions of the globule. J. Biomol. Struct. Dyn. 4, 885–893.
Pucci, F., Rooman, M., 2014. Stability curve prediction of homologous proteins using temperature-dependent statistical potentials. PLoS Comput. Biol. 10, e1003689.
Pucci, F., Bourgeas, R., Rooman, M., 2016. Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing HoT-MuSiC. Sci. Rep. 6, 23257.
Qian, B., Goldstein, R.A., 2001. Distribution of indel lengths. Proteins 45, 102–104.
Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q., Baker, D., 1997. Functional rapidly folding proteins from simplified amino acid sequences. Nat. Struct. Biol. 4, 805–809.
Romero, P.A., Arnold, F.H., 2009. Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. 10, 866–876.
Roodveldt, C., Aharoni, A., Tawfik, D.S., 2005. Directed evolution of proteins for heterologous expression and stability. Curr. Opin. Struct. Biol. 15, 50–56.
Rorick, M., 2012. Quantifying protein modularity and evolvability: a comparison of different techniques. Biosystems 110, 22–33.
Rose, G.D., Fleming, P.J., Banavar, J.R., Maritan, A., 2006. A backbone-based theory of protein folding. Proc. Natl. Acad. Sci. U. S. A. 103, 16623–16633.
Roy, S.W., Nosaka, M., de Souza, S.J., Gilbert, W., 1999. Centripetal modules and ancient introns. Gene 238, 85–91.
Sato, Y., Niimura, Y., Yura, K., Go, M., 1999. Module-intron correlation and intron sliding in family F/10 xylanase genes. Gene 238, 93–101.
Schimmel, P.R., Flory, P.J., 1967. Conformational energy and configurational statistics of poly-L-proline. Proc. Natl. Acad. Sci. U. S. A. 58, 52–59.
Shakhnovich, E.I., 1998. Protein design: a perspective from simple tractable models. Fold. Des. 3, R45–R58.
Shimada, J., Yamakawa, H., 1984. Ring-closure probabilities for twisted wormlike chains. Application to DNA. Macromolecules 17, 689–698.
Shore, D., Langowski, J., Baldwin, R.L., 1981. DNA flexibility studied by covalent closure of short fragments into circles. Proc. Natl. Acad. Sci. U. S. A. 78, 4833–4837.
Sigrist, C.J., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I., 2013. New and continuing developments at PROSITE. Nucleic Acids Res. 41, D344–D347.
Svedberg, T., 1929. Mass and Size of protein molecules. Nature 123, 871.
Taylor, T.J., Vaisman, I.I., 2006. Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures. Phys. Rev. E Stat. Nonlin Soft

Matter Phys. 73, 041925.
Tokuriki, N., Tawfik, D.S., 2009. Protein dynamism and evolvability. Science 324, 203–207.
Tokuriki, N., Oldfield, C.J., Uversky, V.N., Berezovsky, I.N., Tawfik, D.S., 2009. Do viral proteins possess unique biophysical features? Trends Biochem. Sci. 34, 53–59.
Trifonov, E.N., Berezovsky, I.N., 2002. Molecular evolution from abiotic scratch. FEBS Lett. 527, 1–4.
Trifonov, E.N., Berezovsky, I.N., 2003. Evolutionary aspects of protein structure and folding. Curr. Opin. Struct. Biol. 13, 110–114.
Trifonov, E.N., Kirzhner, A., Kirzhner, V.M., Berezovsky, I.N., 2001. Distinct stages of protein evolution as suggested by protein sequence analysis. J. Mol. Evol. 53, 394–401.
Turner, G.C., Varshavsky, A., 2000. Detecting and measuring cotranslational protein degradation in vivo. Science 289, 2117–2120.
UniProt, C., 2008. The universal protein resource (UniProt). Nucleic Acids Res. 36, D190–D195.
Vallat, B., Madrid-Aliste, C., Fiser, A., 2015. Modularity of protein folds as a tool for template-free modeling of structures. PLoS Comput. Biol. 11, e1004419.
Vamvaca, K., Vogeli, B., Kast, P., Pervushin, K., Hilvert, D., 2004. An enzymatic molten globule: efficient coupling of folding and catalysis. Proc. Natl. Acad. Sci. U. S. A. 101, 12860–12864.
Varenne, S., Buc, J., Lloubes, R., Lazdunski, C., 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. 180, 549–576.
Veretnik, S., Bourne, P.E., Alexandrov, N.N., Shindyalov, I.N., 2004. Toward consistent assignment of structural domains in proteins. J. Mol. Biol. 339, 647–678.
Voigt, C.A., Martinez, C., Wang, Z.G., Mayo, S.L., Arnold, F.H., 2002. Protein building blocks preserved by recombination. Nat. Struct. Biol. 9, 553–558.
Voss, N.R., Gerstein, M., Steitz, T.A., Moore, P.B., 2006. The geometry of the ribosomal polypeptide exit tunnel. J. Mol. Biol. 360, 893–906.
Walter, K.U., Vamvaca, K., Hilvert, D., 2005. An active enzyme constructed from a 9-amino acid alphabet. J. Biol. Chem. 280, 37742–37746.
Wang, J., Wang, W., 1999. A computational approach to simplifying the protein folding alphabet. Nat. Struct. Biol. 6, 1033–1038.
Wetlaufer, D.B., 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. U. S. A. 70, 697–701.
Wheelan, S.J., Marchler-Bauer, A., Bryant, S.H., 2000. Domain size distributions can predict domain boundaries. Bioinformatics 16, 613–618.
Wieninger, S.A., Ullmann, G.M., 2015. CoMoDo: identifying dynamic protein domains based on covariances of motion. J. Chem. Theory Comput. 11, 2841–2854.
Wootton, J.C., 1994. Sequences with 'unusual' amino acid compositions. Curr. Opin. Struct. Biol. 4, 413–421.
Yamakawa, H., Stokmayer, W.H., 1972. Statistical mechanics ofwormlike chains. 2. Excluded volume effects. J. Chem. Phys. Biol. 57, 2843–2854.
Yew, B.K., Chintapalli, S.V., Upton, G.G., Reynolds, C.A., 2007. Conservation of closed loops. J. Mol. Graph Model 26, 652–655.
Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I., 2006. Physical origins of protein superfamilies. J. Mol. Biol. 357, 1335–1343.
Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I., 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput. Biol. 3, e5.
Zheng, Z., Goncearenco, A., Berezovsky, I.N., 2016. Nucleotide binding database NBDB–a collection of sequence motifs with specific protein-ligand interactions. Nucleic Acids Res. 44, D301–D307.