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'