IOP Publishing Phys. Biol. 12 (2015) 045002 doi:10.1088/1478-3975/12/4/045002

Physical Biology

RECEIVED
30 September 2014

REVISED
15 February 2015

ACCEPTED FOR PUBLICATION
11 March 2015

PUBLISHED
9 June 2015

PAPER

Protein function from its emergence to diversity in contemporary proteins

Alexander Goncearenco[1,4] **and Igor N Berezovsky**[2,3,5]

[1] Computational Biology Unit and Department of Informatics, University of Bergen, N-5008 Bergen, Norway
[2] Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, Singapore 138671
[3] Department of Biological Sciences (DBS), National University of Singapore (NUS), 8 Medical Drive, 117597, Singapore
[4] Current address: Computational Biology Branch of the National Center for Biotechnology Information in Bethesda, MD, United States of America
[5] Author to whom any correspondence should be addressed.

E-mail: gonceare@ncbi.nlm.nih.gov **and** igorb@bii.a-star.edu.sg

Keywords: evolution of protein function, elementary function, protein folds and functions, origin of life, prebiotic peptides, enzymatic funciton, design of protein function

Supplementary material for this article is available online

Abstract

The goal of this work is to learn from nature the rules that govern evolution and the design of protein function. The fundamental laws of physics lie in the foundation of the protein structure and all stages of the protein evolution, determining optimal sizes and shapes at different levels of structural hierarchy. We looked back into the very onset of the protein evolution with a goal to find elementary functions (EFs) that came from the prebiotic world and served as building blocks of the first enzymes. We defined the basic structural and functional units of biochemical reactions—elementary functional loops. The diversity of contemporary enzymes can be described via combinations of a limited number of elementary chemical reactions, many of which are performed by the descendants of primitive prebiotic peptides/proteins. By analyzing protein sequences we were able to identify EFs shared by seemingly unrelated protein superfamilies and folds and to unravel evolutionary relations between them. Binding and metabolic processing of the metal- and nucleotide-containing cofactors and ligands are among the most abundant ancient EFs that became indispensable in many natural enzymes. Highly designable folds provide structural scaffolds for many different biochemical reactions. We show that contemporary proteins are built from a limited number of EFs, making their analysis instrumental for establishing the rules for protein design. Evolutionary studies help us to accumulate the library of essential EFs and to establish intricate relations between different folds and functional superfamilies. Generalized sequence-structure descriptors of the EF will become useful in future design and engineering of desired enzymatic functions.

Introduction

Enzymes are biological catalysts that maintain chemical transformations of the substrate molecules in living organisms, thus supporting their functioning and survival in different environments. The number of natural substrates and metabolites is limited, and there are around 5000 currently known biochemical transformations (Enzyme Commission (EC) numbers [1]) that process natural substrates. The number of characterized mechanisms that provide these transformations is only about 330 [2]. Further, there are only about 30 chemical roles for the amino acid residues involved in catalysis, such as electron donor/acceptor, proton donor/acceptor, electrostatic stabilizer, activator, etc [3]. Thus, thousands of multistep biochemical transformations are actually built from the several dozens of distinct elementary reactions [2, 4, 5]. This striking reduction of functional complexity at the level of individual chemical roles and their actions in biochemical transformations prompts one to think about the very emergence of the enzymes

and their elementary functions (EFs). Existence of the conserved functional motifs in different protein superfamilies [6–8] hints to their possible origin from ancestral peptides with primitive functions [9], which presumably were able to work as catalysts [10, 11] and to form primitive assemblies [12] in the prebiotic world [13]. Fusion of the prebiotic functional peptides into enzymatic domains/folds followed by the recombination of the latter into the multidomain structures and oligomeric complexes constitute the essence of protein evolution from its emergence to contemporary complexity and diversity [9, 14].

Fundamental physics [15–18] lies in the foundation of protein structure and drives its evolution, determining optimal sizes [19] and shapes [20, 21] at different levels of structural hierarchy [14, 22, 23], providing specific traits necessary for the work of natural selection [24–26], and introducing requirements on sequences and structures that secure their fitness and evolvability [27–29]. Specifically, polymer nature of the polypeptide chains [18, 30–32] established the shape and size [33, 34] of the basic structural elements of soluble proteins—returns of the protein backbone (or closed loops) with a preferential size of 25–30 residues [35, 36]. Circularization of the double-stranded DNA, governed by the same polymer law [33, 34] of the flexibility-based ring closure, apparently contributed to a selection of the typical size of a protein domain—100–200 amino acid residues (300–600 base pairs is an optimum for the ring closure of double-stranded DNA [14]). Designability of protein folds/domains—a characteristic of the packing density and the balance between the short- and long-range non-bonded contacts [25]—is a determinant of the structure stability [24] and its ability to adopt many different sequences [24–26]. Based on the fold's designability, the original ancestral set of thermostable and highly versatile protein folds could have been selected in a hot primordial environment. Highly designable folds would, in turn, provide scaffolds for numerous functional families and superfamilies of contemporary proteins [37]. Primitive catalysts with a potential for self-assembly [10–12] have presumably fused into the first folds [9, 38, 39] thus forming the catalytic domains, and they are now represented in modern proteins by their descendants—elementary functional loops (EFLs) [6, 7].

Despite a significant progress in understanding the evolution of protein function [14, 28, 40–45], its early history and the consequences for contemporary proteins remain enigmatic [31]. One of the most important unsolved questions in the studies of emergence and early protein evolution is to determine the building blocks of the first enzymatic domains and to characterize the critical demands on their sequences, structures, and catalytic activities. In particular, we would like to obtain a set of the most common EFs and their characteristic signatures in the form of protein sequence profiles [6, 7]. Next, we would need to prove their ancient character by detecting the corresponding representatives in unrelated folds and protein superfamilies. We would also like to determine the folds that could have provided the structural scaffolds for these functions, opening thus a road for further protein evolution [8, 28, 40–44]. While pursuing these tasks, we survey how physics determines the structural characteristics of natural proteins and the versatility of their design. We show that the number of natural catalytic sites is much smaller, and there is a limited number of combinations of dyads/triads/tetrads that are favorable for the majority of catalytic sites. The limited repertoire of the natural catalytic sites and strong restrictions on the sizes and structures of the functional folds imply that many of the contemporary proteins are either descendants and recombinants of a handful of ancient primitive proteins/peptides, or they incorporate some newly evolved EFs following the same rules. Therefore, consideration of enzymatic function as a combination of elementary ones not only helps in understanding the contemporary enzymes, their structures and functions, but it can also help to establish a set of rules for the design of desired functions. As an outlook, given the complexity and diversity of the sequence-structure relationship, we propose integrating the structural and sequence features of the EFLs into a probabilistic model—*descriptor of EF*. We discuss here a case study of the 'nucleophile' EF, its complexity and diversity, and consider how features of this EF can be integrated into a descriptor for future use in protein design.

Materials and methods

Important definitions

In order to define the *EF* and *EFL* we rely upon the following IUPAC definitions.

Biochemical transformation—conversion of a substrate into a particular product irrespective of the mechanisms involved. The EC nomenclature of enzymes classifies the biochemical transformations [1].

Biochemical mechanism—is a detailed description of the process leading from the reactants to the products of a reaction, including characterization of reaction intermediates and the corresponding steps. The databases Mechanism, Annotation, and Classification in Enzymes (MACiE), Metal-MACiE, and Structure-Function Linkage Database create a compendium of enzymatic mechanisms and reactions [3–5].

Elementary reaction (term E02035 in IUPAC Goldbook) has no intermediates, occurs in a single step and passes through a single transition state [32].

Based on the above definitions and extending it on to the cases of binding of common ligands, such as metal- and nucleotide-containing cofactors and their parts (sugar, base, phosphate, etc) we define the EF.

EF is an elementary reaction or binding interaction that provides stabilization of the transition state [46] *en route* of the biochemical transformation. We use MACiE's catalytic roles of amino acid residues for standardizing the glossary of EFs [47].

Closed loop or return of the polypeptide backbone with a typical size of 25–30 amino acid residues is a basic universal structural element of globular proteins originating from the polymer nature of polypeptide chains [35, 36].

EFL is defined as a structural-functional unit formed by the closed loop [35, 36], carrying one or a few functional residues responsible for the corresponding *elementary reaction* or binding interactions [6, 7]. EFL serves as a minimal functional building block in biochemical mechanisms.

Functional site is defined according to annotations in the Conserved Domains Database (CDD) [48, 49] and is classified into active, polypeptide binding, nucleic acid binding, ion binding, chemical binding, post-translational modification and other. We rely on CDD definition of the functional site for identification of the residues involved in binding and catalysis and annotation of EFLs. Functional site is represented by a set of residues. SwissProt site annotations overlap well with the CDD annotations especially for active, binding, and metal-binding sites [49].

Active site (*or catalytically active site*) is a type of functional sites that is directly involved in biochemical transformations. The active site consists of the residues involved in substrate and cofactor binding, stabilization of the transition state and the chemically active residues that facilitate catalysis (*catalytic residues*). The Catalytic Site Atlas (CSA [50]) annotates the catalytic residues in structures of enzymes.

Fold is defined according to the second level of classification of protein structures in SCOP database [37]. We refer to folds by a standard (*class.fold*) notation. For example, *c*.1 is TIM $ß/\alpha$-barrel fold in the $\alpha/ß$ class (*c*) of proteins.

Profile is a position specific scoring matrix (PSSM) representing an ensemble of multiple sequences of EFLs [6]. Profiles are assigned numeric identifiers and their sequence *signatures* may be shown as PROSITE-like patterns [51] or graphically as sequence logos [52].

Prototype of EF is a generalized and simplified representation of a set of corresponding sequence profiles, reflecting diversity of sequences and conservation of the major signature. Noteworthy, sequences representing the prototype in contemporary proteins should not necessarily have detectable similarity, however they all have to be detected by the profiles constituting the prototype [7].

Descriptor of EF is a combination of the sequence and structure characteristics (e.g. local contacts, distances and dihedral angles) of the EF, derived in the form of empirical distributions from the whole diversity of the representing sequences and structures.

On a more technical side, we use the following important notions.

COGs denote Clusters of Orthologous Groups at NCBI [53]. Proteins belonging to one COG presumably have the same function in different species. The most ubiquitous COGs represented in all of the taxonomic branches form the core, or the most ancient set of proteins that could be attributed to the Last Universal Common Ancestor [42].

CDD [48]—Conserved Domains Database at NCBI. Proteins and sequence alignments in the CDD are manually curated using the structures when available. CDD arranges proteins into families and superfamilies. Some families contain annotations of functional sites, however there is no common glossary or ontology. The CDD also incorporates imported domains from Pfam, SMART, TIGR and other largely overlapping resources; however here we only rely on the annotations performed at NCBI.

SCOP [37]—Structural Classification of Proteins database defines structural classes based on the secondary structure content and structural *folds* based on the architectures formed by the secondary structures. Further down the hierarchy, SCOP folds are classified into the superfamilies and families. Families within a superfamily presumably have a common ancestor. However, the superfamilies are expected to be evolutionary independent.

E-value is a measure of the error rate in a sequence search, which can be directly interpreted as the number of false positive hits expected given a scoring threshold. Typically in BLAST, PSI-BLAST or HMM-based (Hidden Markov Model) search, the *E*-value is approximated based on a theoretical distribution of hit scores and the size of the sequence database. In our pipeline, however, we sacrifice performance to precision and empirically calculate the distributions of the hit scores for each sequence profile. *E*-value, thus, is the number of false positive hits observed given the natural profile and a PSSM with reshuffled positions [6]. An *E*-value of 1 means that among all the obtained hits we expect at most one sequence hit unrelated to the query sequence profile. *E*-value is the only parameter that controls the sensitivity of our search. Precision of profile-superfamily search is further improved by controlling the coverage, or the fraction of the superfamily sequences found by the profile. Thus, even at relatively high *E*-values the search results will not be compromised by spurious hits due to the insignificant coverage they provide.

Data sources

We use several databases in our analysis. Uniprot/Swissprot is the source of the proteomic sequences [51]. Uniprot PROSITE is a collection of multiple sequence alignments of curated motifs and patterns [51]. SISYPHUS is a collection of structural alignments, non-trivial relationships between different

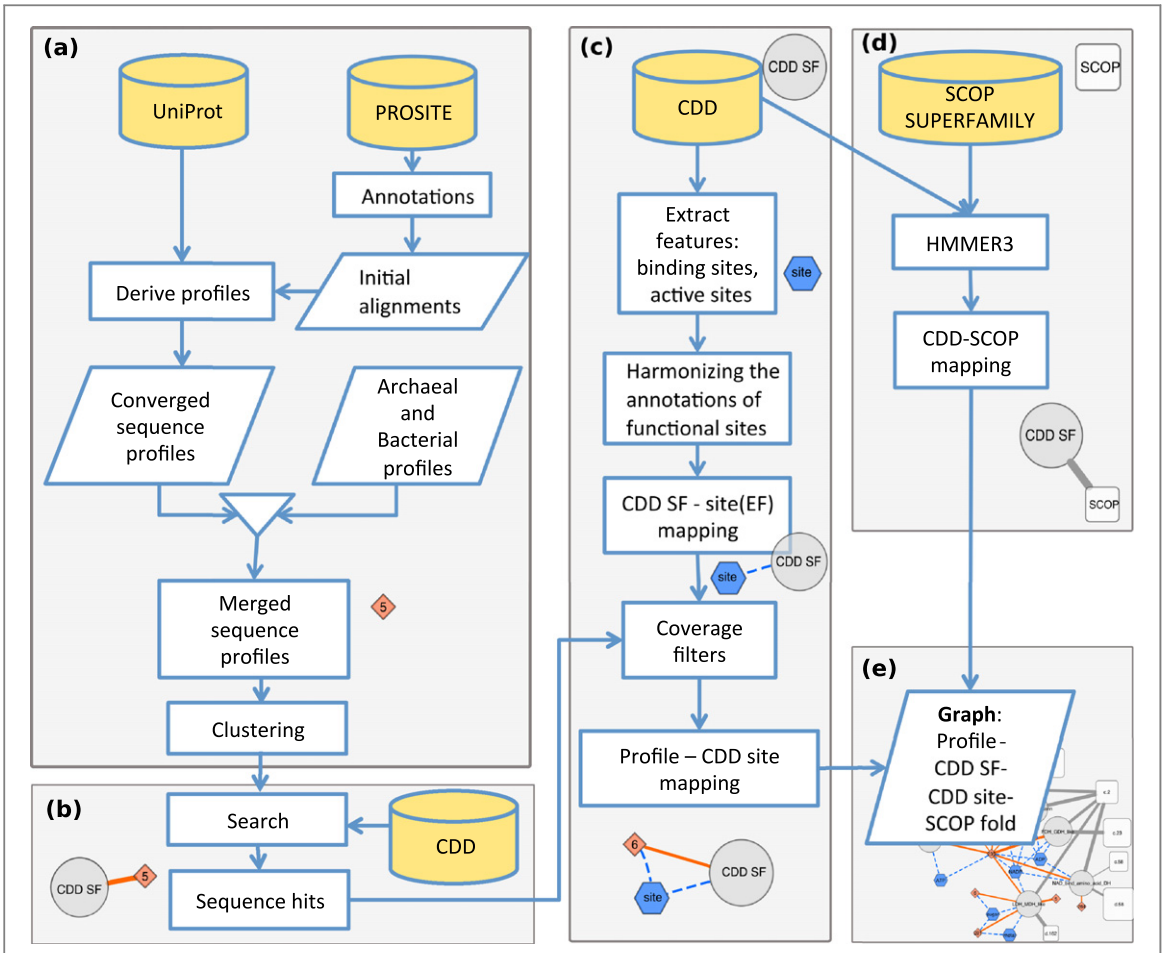

Figure 1. Flowchart of the analysis pipeline. The main components of the pipeline include: (a) derivation of sequence profiles from different sources; the graph is based on the annotation of active sites in the CDD database; (b) profile-sequence search in the CDD database with the obtained profiles; (c) mapping of functional profiles based on the annotations of binding- and active sites in the CDD database; (d) mapping of structural SCOP folds to CDD superfamilies; (e) assembling the graph.

superfamilies, and folds in SCOP [54]. CDD [48] is the source of functional annotations of domain families and superfamilies. We only use NCBI-annotated CDD models in this study. CDD can contain conserved domains that incorporate several structural domains, therefore we use SCOP as a reference for protein folds [37]. The SUPERFAMILY database conveniently provides HMM-models for SCOP [55] that we used for CDD-SCOP mapping.

Analysis pipeline

The flowchart of our analysis pipeline is shown in figure 1. Details of the matching procedure, corresponding parameters, and data sets are described in the following sections of materials and methods: data sources; obtaining sequence profiles of EFLs; assigning the EF. The main components of the pipeline include:

(a) *Obtaining the sequence profiles from different sources.* In this block, we have accumulated the most complete set of sequence profiles, including those from the previous works [6, 7, 9] and those derived *de novo* here using the earlier developed procedure [6] with the origins obtained from Uniprot and Prosite [51]. All obtained profiles were clustered [7] and merged in order to eliminate the redundancy.

(b) *Identifying the sequence matches corresponding to profiles in the CDD domains.* We found the sequence matches of profiles in the CDD domains, establishing thus connections between the sequence profiles and their representatives (EFLs) in functionally annotated conserved domains.

(c) *Delineating the EFs of EFLs by analyzing the binding and active sites in the CDD.* We annotated the EFs of the sequence profiles by determining the overlap between the corresponding EFLs and the functional sites annotated in the CDD.

(d) *Mapping the SCOP folds to CDD superfamilies.* CDD classifies conserved protein domains, while SCOP is a structural classification. We identified SCOP structural domains in the CDD superfamilies by using HMMER3 [56] for finding matches between the representative sequences of CDD superfamilies and sequences of SCOP folds provided by the SUPERFAMILY database [57].

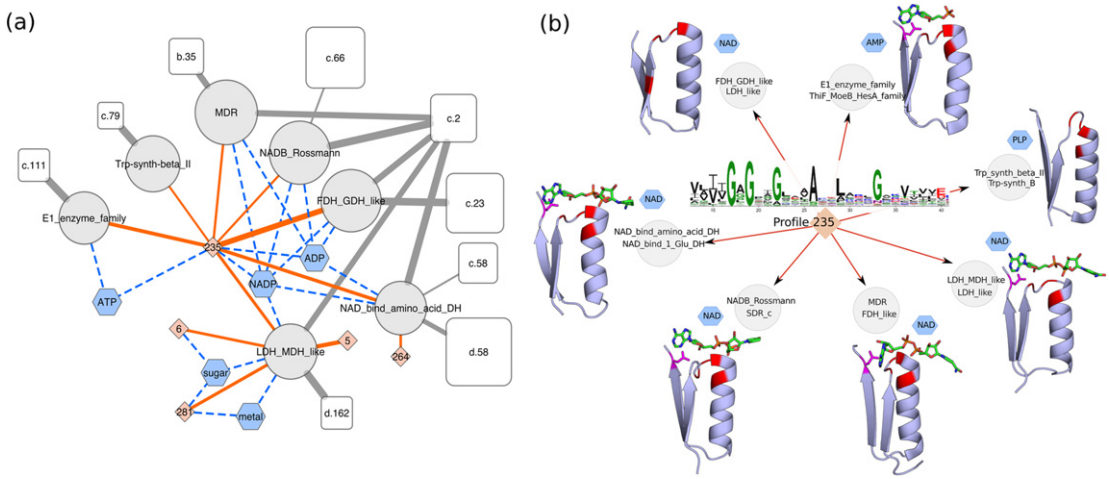

Figure 2. Elementary function—phosphate binding in dinucleotides. (a) The subgraph of connections between the CDD superfamilies and the profile 235 representing the nucleotide binding elementary function. Circles are superfamilies, rounded squares represent folds (with SCOP identifiers), orange diamonds denote profiles (with profile numbers), and the sites annotations are shown as hexagons. Following SCOP folds are represented in the chart (a): b.35—Smp-1-like; c.2—NAD(P)-binding Rossmann-fold domains; c.23—methylesterase CheB, C-terminal domain; c.58—subtilisin-like; c.66—(phosphotyrosine protein) phosphatases II; c.79—molybdenum cofactor biosynthesis proteins; c.111—SurE-like; d.58—NSFL1 (p97 ATPase) cofactor p47, SEP domain; and d.162—PG1857-like. (b) Examples of elementary functional loops in PDB structures from different CDD superfamilies. The nucleotide is shown in sticks, the conserved glycines are colored in red on the ribbon, whereas adenine-polarizing glutamate or aspartate is shown by pink sticks. The PDB codes of the proteins from which we obtained structural examples are: 1HYH, 1JBW, 1V9L, 1X1Q, 1XDW, 2BGM, 2DPH.

(e) *Assembling the graph of relationships between the CDD superfamilies, profiles of EFLs, functional sites in CDD, and the SCOP folds.* We first built a general graph, which contains multiple types of nodes and edges. The node types include: (1) diamonds—the profiles of EFLs, (2) hexagons—EFs represented by the type of the functional site according to CDD, (3) circles—superfamilies of conserved domains in CDD, and (4) rounded squares—SCOP folds. The graph encodes exhaustive information on the relationships between the aforementioned nodes and is provided in the supplementary materials available at stacks.iop.org/PB/12/045002/mmedia in form of the interactive Cytoscape sessions and their HTML snapshots. Different types of relationships are represented by different types of edges: (1) orange solid line—profile-CDD superfamily link denotes a sequence match; (2) blue dashed line —shows that the functional site belongs to a CDD superfamily and that the profile describes one of the functional loops forming the functional site; (3) gray solid line—indicates that the SCOP folds comprising the CDD domain.

The general graph is then split into two parts and used for in-depth analysis of the relations between profiles of EFLs in CDD superfamilies and EFs that form enzymatic activities (figure 3), and the second part—profiles of EFLs connecting CDD superfamilies and SCOP domains (figure 4). All the examples displayed in figures 2–5 are sub-graphs extracted from the general graph. We use Cytoscape 3.1.1 for visualizing the graphs. All the sub-graphs are also accessible as interactive Cytoscape sessions and as HTML snapshots for the web (https://github.com/neksa/elementary-functions). CDD superfamilies are represented by the corresponding short names, these abbreviations can be directly looked up in the CDD search (http://www.ncbi.nlm.nih.gov/cdd/). Long superfamily names will show up as pop-ups in the Cytoscape when the mouse cursor is placed over the corresponding node, for instance 'medium chain reductase (MDR)/dehydrogenase/zinc-dependent alcohol dehydrogenase-like family' is abbreviated in CDD as MDR. Profiles are referred to by their numbers and contain sequence signature and a short functional annotation. The sites are automatically named based on the type of cofactor/ligand involved or the type of the functional site. Edges encode the information about the *E*-value of the profile-sequence match and coverage of the corresponding superfamily (shown by edge thickness, the values are accessible as edge attributes in Cytoscape). SCOP–CDD edges are also characterized by their coverage.

Obtaining sequence profiles of EFLs

We obtained the sequences profiles from several sources. The first source is the archaeal proteomes. The 30- and 50-residue long segments from the sequences of representatives of archaeal COGs [42] were used here as origins for deriving the profiles. The origins were iteratively matched against 68 non-redundant (less than 70% sequence identity) archaeal proteomes until they converged into sequence profiles (figure 1(a)). The profiles were clustered and merged in order to remove any remaining redundancy. The

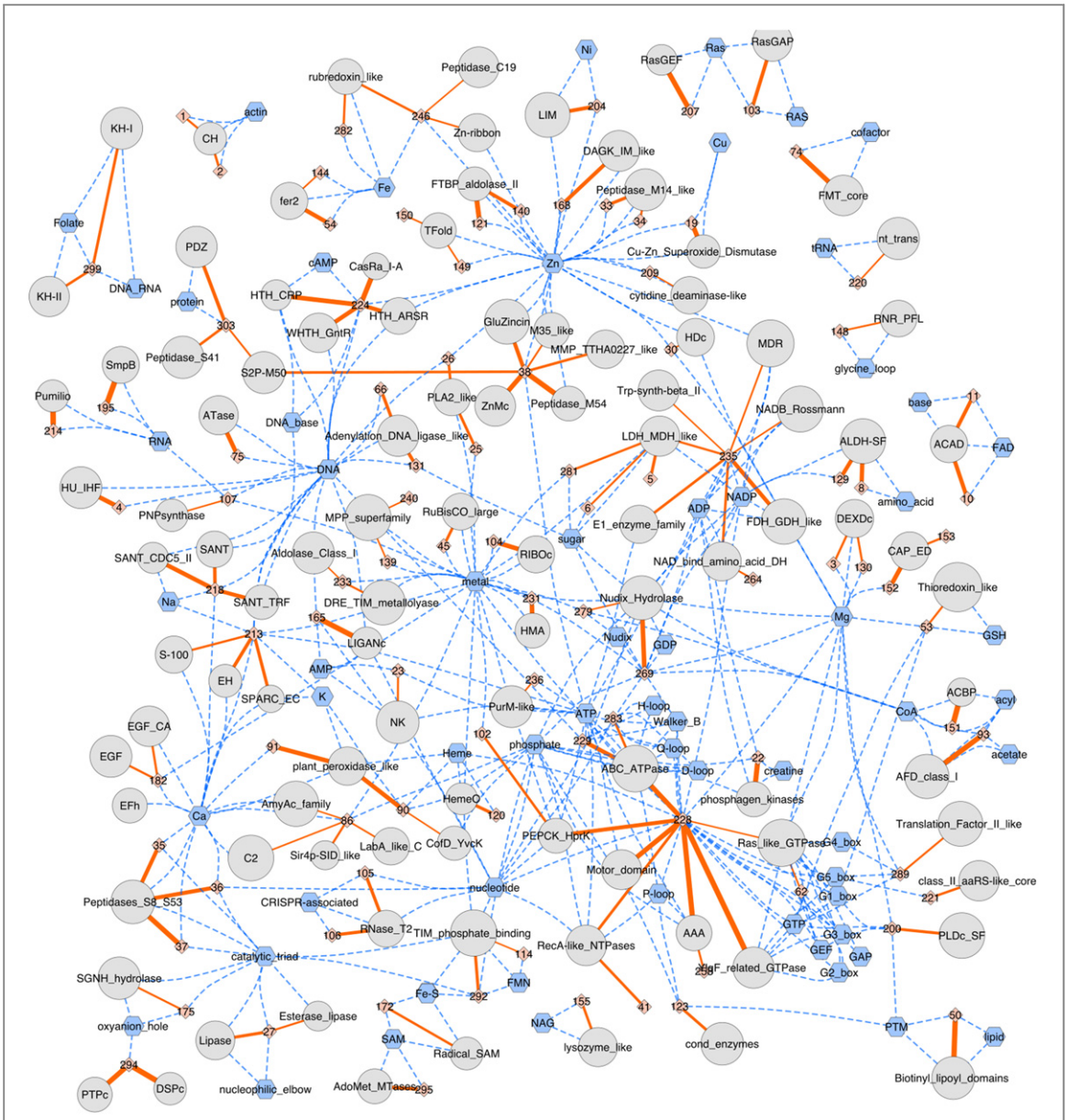

Figure 3. Graph of the evolutionary connections between superfamilies, elementary functions, and their sequence profiles. CDD superfamilies are shown as circles, where the size of the circle represents the number of families inside the superfamily. The number of families in which profile of the elementary function (orange diamond) finds a match is indicated by the thickness of the solid orange line. The total number of families in the CDD superfamily is indicated by the size of the circle. The presence of certain elementary function (blue hexagon) in the active sites of members of the CDD superfamily, and in the profile signature is shown by the dashed blue line. The sizes of the superfamilies are proportional to the number of families within each superfamily. Thickness of the edges represents the confidence in the relationship, in case of superfamily-profile link it denotes the coverage of the superfamily by sequence hits of the corresponding profile. We provide the graphs in Cytoscape format and suggest exploring the graph with Cytoscape software. All objects in the graph are linked to the corresponding websites via the URL attribute and full-length labels will be shown as tooltips when mouse cursor is pointing to them. The graph is also available as an HTML snapshot (see https://github.com/neksa/elementary-functions/).

procedures for converging and clustering/merging the profiles are described in detail elsewhere [6, 7]. Additionally, we extracted 49 multiple sequence alignments from the PROSITE [51] database and 21 alignments from the SISYPHUS database [54]. In this case, the alignments have been used as origins and have been iteratively matched to non-redundant Uniprot (less than 40% sequence identity) until convergence.

Totally, we obtained 294 sequence profiles with distinct functional signatures, which are represented by their serial numbers or by PROSITE-like patterns uniquely identifying their signatures. The list of 124 profile logos with annotated functional sites is provided in the supplementary file 1 available at stacks.iop.org/PB/12/045002/mmedia.

Assigning the EFs to sequence profiles

It is only possible to annotate the EF if the mechanism of biochemical transformation is known or at least the active or binding site is clearly defined. We rely on functionally-characterized domain families, where the functional sites are annotated as features in the CDD

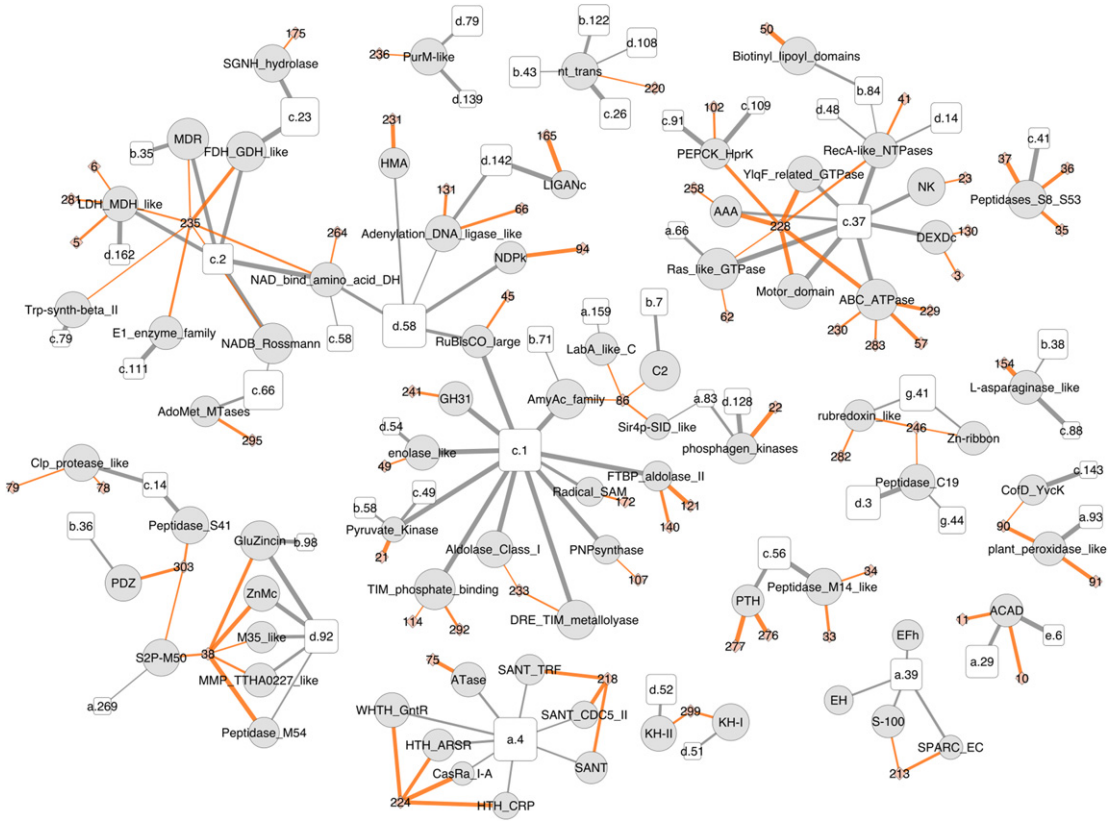

Figure 4. Graph of the evolutionary connections between superfamilies, folds, and sequence profiles of elementary functions. SCOP folds are mapped onto the CDD domain superfamilies. Some CDD functional domains contain multiple structural domains. The shapes and colors of nodes and edges are consistent with the previous figures. The size of the nodes representing the folds shows the number of SCOP superfamilies within the fold, which is a measure of its designability. The interactive graph is available as a Cytoscape session and as an HTML snapshot (https://github.com/neksa/elementary-functions/).

database [48] and use the functional annotations of the sequence signatures from InterPro Prosite [51] and SISYPHUS [54].

First, since there is no glossary or functional ontology in the CDD we harmonized the annotation of functional sites in CDD in order to merge synonyms and different spellings. We also oversimplified the original annotation in order to highlight the key EF associated (such as calcium ion coordination, a general base, acid, a nucleophile, phosphate group binding) and reduce it to a controlled glossary (e.g. Ca, base, acid, nucleophile, phosphate). We provide our simplified glossary in the supplementary Github repository (file input/site_keywords.tab). However, a large number of sites were annotated as the 'active site' in the CDD without any indication of their specific function. We added a unique superfamily identifier and the serial number of the site in order to identify the ambiguously annotated sites correctly. The annotation of the functional sites has been performed automatically with a condition that at least 30% of the functional site residues should be covered by a profile. We consider that a profile that matches a functional site represents an EF. Thereby, we were able to select the profiles that represent EFLs from the list of 294 sequence profiles that we obtained.

We mapped SCOP folds onto the CDD domains by using HMMER3 with the default parameters (*E*-value threshold 0.01). Each HMM model represented a SCOP domain superfamily, whereas the sequences were taken from the CDD, one sequence representing one CDD model. The CDD models and their matching SCOP folds have been aggregated to show all the structural folds within the CDD superfamily.

The resulting general graph of profiles-sites-superfamilies-folds required additional filtering for visualization. Filtering of matches has been performed based on the number of matches and the superfamily coverage at different *E*-value thresholds. The python script that assembles and filters the graph along with the necessary input files (in CSV format) and a sample output (in SIF, CSV and EDO formats) is provided in the Github repository (https://github.com/neksa/elementary-functions). All the thresholds that may affect the results are listed in the file header in order to make the results reproducible. The number of hits in a superfamily at *E*-value <1 had to be at least 2. The coverage of superfamilies with *E*-value <100 was set to be at least 5%. We only kept what we considered as functional profiles covering the annotated functional sites, while all other profiles are not shown in the graph. The CDD superfamilies without the annotated sites have

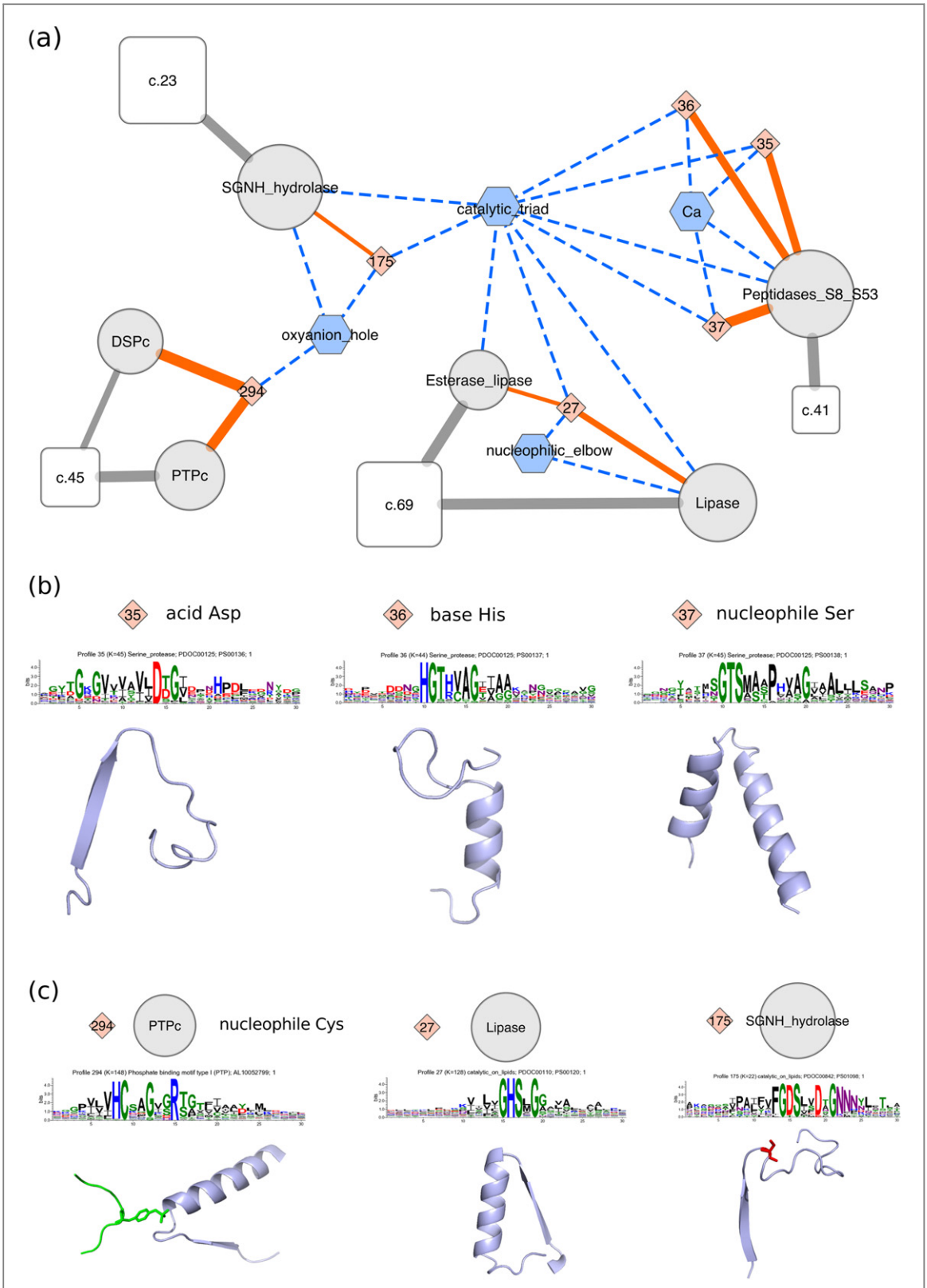

Figure 5. Elementary functions involved in hydrolysis: (a) subgraph of elementary function including a catalytic triad, oxyanion hole and a nucleophilie; the following SCOP folds are represented in the chart a: c.23—methylesterase CheB, C-terminal domain; c.41—isochorismatase-like hydrolases; c.45—NagB/RpiA/CoA transferase-like; and c.69–peptidyl-tRNA hydrolase II. (b) The elementary functions, sequences, and their structural representatives in the catalytic triad, which includes an acid, a base, and a nucleophile. (c) Distinct signatures of the nucleophile elementary function in different enzymatic functions. The substrate of phosphatases is a phosphorylated serine, threonine or tyrosine residue. The substrate peptide and phosphorylated tyrosyl side chain is shown in green in the left panel (profile 294). The phosphatases do not have a triad, but they can have an acid residue forming a dyad. The phosphoryl intermediate is stabilized by the backbone amides, which follows the nucleophile forming an oxyanion hole. The PDB codes of the structural examples are: 1THM (for profiles 35, 36 and 37), 1EEO (293), 1BU8 (27), and 1ESE (175).

also been excluded from the graph. The redundant profiles with slightly different signatures, thus very similar connectivity in the graph have also been removed for clarity. As a result of filtering we were left with 124 profiles, 126 superfamilies, 64 EFs/sites, and 104 folds. All graphical examples in the paper are showing parts of this graph. Importantly, filtering does not affect the results or conclusions; the main reason for filtering is clarity of presentation of the graphics in figures.

Results

Many questions should be answered in order to shed light on the emergence of modern proteins, and in order to understand how early events in the protein evolution determined their structures and functions. Specifically, we would like to find out: what are the rules for assembling the enzymatic functions from the elementary ones? What is the role of physics in determining the realm of natural enzymes and how the evolution contributed into establishing the structure-function relationships? Clear and transferable terminology is very important for understanding results and making conclusions, therefore we explain each term in the definitions section of materials and methods. Below we will sequentially consider the EFs, EFLs, and their roles in building the functional domains. We will also tackle the question of the impact of the fold designability on the diversification in functional superfamilies. Finally, we will formulate the requirements for the integrated representation of EF necessary for computational protein design—descriptor of the EF. We expect that answers on these questions will help us in understanding the evolution of protein function and in formulating the basic rules for computational *de novo* design of desired protein functions. In order to identify the signatures of ubiquitous EFLs we use the earlier developed sequence analysis procedure [6, 7]. Overall, we obtained 294 profiles with distinct signatures, and for 124 profiles we have found functional annotation.

Glycine-rich signature of the dinucleotide binding is one of the ancient EFs

Figure 2 contains an example of the very common Gly-reach signature with the most conserved part [ILV][ALV][ILV]xxGxGxxGxx[ALV]A and the key motif GxGxxG. The latter represents a generalized signature of the dinucleotide binding of substrates such as flavin adenine dinucleotide (FAD), nicotinamide adenine dinucleoide (NAD), NADP, as well as mononucleotides and phosphate-containing cofactors, such as pyridoxal phosphate (PLP) and AMP. It is important to note that while the commonly accepted designation of this signature is dinucleotide binding [6, 9, 58], the Gly-rich elbow actually binds the phosphate(s).

The corresponding sequence profile (number 235) finds matching representatives in different entities of the Conserved Domain Database (CDD [48], figure 2), among which are: Rossmann-fold nicotinamide-adenine dinucleotide (phosphate) (NAD(P)) (+)-binding proteins (NADB_Rossmann); NAD-dependent, lactate dehydrogenase-like, 2-hydroxycarboxylate dehydrogenase family (LDH_MDH_like); tryptophan synthase beta superfamily, fold type II (Trp_synth_beta_II); and medium chain reductase/dehydrogenase (MDR). (Di)nucleotide binding performed by these domains is important in biochemical reactions that are carried out by different enzymes. For example, NAD(P)-binding domains work in a number of dehydrogenases (e.g. alcohol dehydrogenases (ADHs)) and tyrosine-dependent oxidoreductases. The zinc-dependent ADHs, which represent the MDR, catalyze the NAD(P)(H)-dependent interconversion of alcohols to aldehydes or ketones. Tryptophan synthase beta superfamily (fold type II) utilizes the PLP for catalyzing beta-replacement and beta-elimination reactions. AMP binding signature presumably reflects utilization of the ATP in the superfamily of activating enzymes (E1). The EFLs described by the profile 235 bind phosphates in dinucleotide-containing (FAD/NAD/NADP), as well as in nucleotide- and phosphate-containing (PLP/ATP/AMP) ligands (figure 2). The repertoire of folds in which this EF is present includes mostly α/β folds (folds c.2, c.23, c.58, c.66, c.79, and c.111 belong to the *c* class in SCOP representing the α/β arrangement of secondary structures). In many cases (di)nucleotide-binding domains are coupled with other functional domains, such as β GroES-like fold (b.35), and $\alpha+\beta$ folds (d.58 and d.162) that form a biochemical function together with the α/β folds. The key motif, GxGxxG, of the dinucleotide binding EF (profile 235, see figure 1 for details) is very similar to the one (GxxGxG) of the mononucleotide binding (profile 228, see figure 3) dominated by the so-called P-loop that provides ATP/GTP binding. The glycine-rich elbows induce backbone conformations that provide preferential interactions with the phosphate groups in both profiles of the mono- and dinucleotide binding EFs (profiles 228 and 235, respectively). The spacers (xx and x) between the glycines are important for specificity to mono- and dinucleotides. Based on the diversity of folds and functional superfamilies that contain EF of the (di) nucleotide binding, one can conclude that the generalized signature GxGxxG of the profile 235 is probably a descendant of the unique prototype with the same 'poly-glycine-like' signature (we have recently found that GxGGxG is a common prototype; data not shown) and the EF of the phosphate binding [59, 60] in nucleotides. Phosphate binding, which, in turn, is part of the (di)nucleotide binding, was presumably one of the first EFs that provided the basic links between nucleic acids and proteins in the origin of life. Undoubtedly, there should be some other similarly

basic and important EFs [60, 61] that should have left traces in numerous folds and functional superfamilies of modern proteins [58, 60]. Therefore, the next question we would like to ask is what was the repertoire of EFs available for building the first enzymatic domains/folds?

Abundant EFs unravel functional relations between the enzymatic functions of CDD superfamilies

Figure 3 describes EFs (light blue hexagons) and sequence profiles (orange diamonds) in relation to CDD superfamilies (gray circles). The major EFs are described in the figure 3 via the corresponding substrates (e.g. DNA, RNA, actin, peptide, base, sugar, phosphate etc) and cofactors/metals (FAD, flavin mononucleotide (FMN), NADP, Fe, Ca, Cu, K, Mg, Na, Ni, Zn etc). Some EFs are named by the function it is part of (e.g. the catalytic triad in hydrolysis) or by the name of corresponding EFL (glycine loop, D-/H-/Q-loop, Walker B, nucleophilic elbow). We resorted to showing only the most abundant EFs present in CDD superfamilies in figure 3. The complete graph of the evolutionary relations that includes both figures 3 and 4 is given in the supplementary information available at stacks.iop.org/PB/12/045002/mmedia and can be interactively explored via the web pages in the supplementary Github repository (https://github.com/neksa/elementary-functions/).

The most common functions are metal ion binding and various nucleotide-derived ligand binding; these are also the most common site annotations among SwissProt proteins [49]. Metals and metal-containing hemes and clusters are indispensable elements in redox reactions that are involved into many biochemical functions (figure 3). For example, the signatures containing the CxxC pattern are very frequent in the binding of metals and metal-containing cofactors (profiles 17, 55, 113, 182, 204, and 246; sequence logos available in 124_profiles_logo.pdf in supplementary information). Earlier, we suggested that the ancient prototypes with a simple CxxC signature may have given rise to the diversity of contemporary EFs of metal-/metal-containing cofactor binding and redox reactions [7], contributing thus to the transition from the prebiotic catalysis [62, 63]. In addition to the CxxC motif, aspartate-rich signatures (profiles 90, 91, 149, 169) and histidine-containing signatures (profiles 19, 33, 34, and 38, figure 3 and Cytoscape File 1 in the supplementary information) are also frequent providers of the metal binding EFs.

The EFs involving binding of nucleotide-containing ligands are part of some of the most basic biochemical transformations: (i) hydrolysis of the phosphate groups as an energy source in many enzymatic reactions; (ii) transfer of phosphoryl groups from high energy donors such as ATP in phosphorylation of proteins. Additionally, pyridoxal 5′-phosphate (PLP), NAD(P), coenzyme A (Co-A), S-adenosylmethionine (SAM) work as cofactors/coenzymes [18, 64, 65] in more complex biochemical transformation. Some signatures of the nucleotide-containing ligand binding such as profiles 228, 235 are similar to the extent of their major functional motif, showing that they are possibly originated from one ancient prototype [6]. The key motif, GxGxxG, of the dinucleotide binding EF (profile 235, see figure 1 for details) is very similar to the one (GxxGxG) of the mononucleotide binding (profile 228). The latter is dominated by the so-called P-loop that provides ATP/GTP binding, and the most conserved part of the P-loop profile is: [ILV][ALV][ILV]xxGxxGxGK[ST]xxLLxxL. The glycine-rich elbows induce backbone conformations that provide preferential (depending on the placement of spacers xx and x between glycines) interactions with the phosphate groups in both profiles of the mono- and dinucleotide binding EFs (profiles 228 and 235, respectively). The similarities between GxGxxG (works in dinucleotide binding) and GxxGxG (P-loop, works in nucleotide binding) signature leads to a hypothesis that both signatures may have had a common glycine-rich prototype presumably with the GxGGxG signature (data not shown). Below we discuss several examples of the abundant functional superfamilies that work with nucleotide-containing ligands and cofactors. The motor domain and P-loop NTPase superfamilies (profile 228, figure 3, [48]) include ATPases that provide the driving force in myosin and kinesin processes. The ATP and GTP EFs are present in motor domain and NTPase superfamilies, respectively (profile 228, figure 3). Members of the Ras_like_GTPase superfamily (GTP EF is shown to be present) regulate a wide variety of cellular functions such as gene expression (in particular elongation, termination, and release in translation), DNA replication, cytoskeletal reorganization, vesicle trafficking, nucleo-cytoplasmic transport, microtubule organization, cell division and sporulation. All members of this superfamily possess the GTP-binding site (profile 228, GTP EF, figure 3) assisted by the Mg^{2+} binding (Mg EF, figure 3). MDR family (profile 235) contains the Zn-dependent alcohol dehydroganase with a broad range of activities, such as alcohol, sorbitol, formaldehyde, butanediol dehydrogenase, quinone, ketose, cinnamyl reductase, and numerous others. The Zn-dependent (Zn EF, figure 3) ADHs catalyze NAD(P)(H)-dependent (NAD EF, profile 235, figure 3) interconversion of alcohols to aldehydes or ketons. They are typically dimers or tetramers with two Zn atoms bound to each subunit—one is the catalytic Zn in the active site, the other one is the structural Zn, which is optional in some enzymes in the MDR family. Some EFs work in many different superfamilies. Nucleic acid (RNA/DNA) binding is present in SANT, HU_IHF, Adenylation_DNA_ligase_like, KH-I, SmpB, HTH_XRE and other superfamilies (all the provided abbreviations and short names for the superfamily can be looked up in the CDD database,

http://www.ncbi.nlm.nih.gov/cdd/). Ca and Zn EFs are also connected to multiple CDD superfamilies as representatives of the major cofactor ions involved in enzymatic reactions. Profiles of more specific EFs may be connected to many families within the superfamily but in figure 3 we are mainly interested in exploring the profiles and EFs connecting different superfamilies.

Biochemical function as a combination of the elementary ones

As a result of the graphical representation in figure 3, it becomes possible to analyze the biochemical functions as combinations of the profiles of EFs. EFs defined as phosphate, D-/H-/Q-loops, ATP, ADP, metal, and Walker B work in different combinations in families of the ABC_ATPase superfamily. For example, profiles 281 and 6 of the carbohydrate (sugar) binding, profile 235 of the FAD/NAD/NADP binding, and profile 5 of the ATP binding provide formation of functions in LDH_MDH_like superfamily. Profiles 35, 36, and 37 form a catalytic triad of proteases belonging to the Peptidases_S8_S53 superfamily. In many cases there is only one EF or profile archetypical to the corresponding superfamily. The EFs related to binding of the some abundant substrates, such as DNA, Ca, Zn are present in many superfamilies. Profiles 21, 24, 31, 110, 113, 154 are detected as signatures of different catalytic activities in corresponding superfamilies (figure 3). Many other specific profiles can be found in figure 3 and explored in the Cytoscape session file supplementary materials or with its HTML snapshot (see https://github.com/neksa/elementary-functions/). The complexity in the relationships between the EFs, folds, and functional superfamilies leads to a question about the requirements on the structural scaffolds for different enzymatic functions.

The importance of fold designability for enzymatic functions

Figure 4 depicts relationships between the CDD superfamilies of conserved domains (gray circles), their characteristic profiles of EFs (orange diamonds), and protein folds according to SCOP classification (white squares with SCOP classes). The folds provide structural scaffolds for different functional demands. The most designable folds, i.e. folds that can adopt many different sequences/functions [24–26] form the main connected components on the graph in figure 4: TIM-barrel (c.1), Rossmann fold (c.2) and Rossmann-like folds flavodoxin (c.23), ferredoxin (d.58), P-loop containing NTP hydrolase (c.37), and α-helical bundle (a.4). TIM-barrel and Rossmann folds form the largest hubs in the graph, because many different superfamilies and families are hosted by these most versatile folds (see also supplementary figure S5 available at stacks.iop.org/PB/12/045002/mmedia and supplementary table S3). We refer to folds by their SCOP codes (*class.fold*), complete information about the fold can be looked up in SCOP database online by entering the code in the search box (http://scop.mrc-lmb.cam.ac.uk/scop/search.cgi). Other folds with many superfamilies and families include all-α (four-helical bundle and α-α superhelix, and three-helical bundle), all-β (SH3 barrel and Ig-β sandwich), and $\alpha+\beta$ (Ferredoxin-lke) folds (supplementary figure S5 and supplementary table S3). The cluster formed by the 3-helical bundle fold (a.4) encompasses the DNA-binding superfamilies SANT, HTH etc, and DNA binding and nucleotide hydrolysis EFs (244, 218, 75). Another example is the EF-hand fold (a.39) with a DxDxxG calcium binding signature and structural motif found in superfamilies EH, S-100, SPARC_EC, EFh (these short names represent CDD superfamilies and can be looked up online http://www.ncbi.nlm.nih.gov/cdd/). Highly designable folds are also frequently encountered in functional superfamilies where they work in combinations with other folds. For example, Rossmann fold is coupled with Ferredoxin in amino acid dehydrogenases, and with Flavodoxin fold in formate dehydrogenases. TIM-barrel together with Ferredoxin is characteristic for the Rubisco enzyme superfamily. TIM-barrel also forms a canonical structure of Enolase superfamily [66] in combination with a small capping domain. Another example of the protein function that requires the combination of two folds is Formyltransferase superfamily incorporating N-domain (c.65) and C-domain (b.46) folds. Both folds are necessary for this function, therefore they are considered as a single entity in the CDD database [48]. In some cases several copies of the same fold can be employed by oligomerization thus providing the inter-domain pockets and cavities for binding small-molecule ligands [9, 67]. There are folds that are apparently fit certain functions better. For example, Ferredoxin, Hydrolase, and Rossmann folds are typical scaffolds for the redox and hydrolysis reactions. Some superfamilies recruit specific folds depending on the functional requirements of corresponding families, such as nucleotidyl transferase superfamily (nt_trans), which contains many folds (b.43, b.122, d.108, c.26).

Annotation of the enzymatic function

Existence of the complex relationships between the functions of single domains/folds and multidomain (or multifold) proteins in which they are involved emphasizes on the need for proper functional annotation. The major challenge is to avoid erroneous annotation of the whole protein by the function of an annotated domain, whereas the latter is only a part of the biochemical transformation performed by the protein [68]. The annotation should consider the function of a multidomain protein as a combination of its domains/folds' functions. Domains themselves should be characterized as combinations of EFs. It appears that the requirement for a larger number of

catalytic residues does not necessarily lead to an increase in the number of domains—rather one domain can frequently accommodate large number of active residues (figure S4 in the supplementary file 1). Rigorous domain annotation is also important for understanding how the single domain/fold functions have been formed and for determining the evolutionary relations between different folds and their functional superfamilies. For example, the P-loop EF (profile 228) is found in many superfamilies with the NTP hydrolase fold (c.37) where it binds phosphate in nucleotides and drives the hydrolysis. This EFL is also present in PEPCK fold where it binds the phosphate group of the Phosphonolpyruvate (PEP). Similar to the P-loop the EFL with the dinucleotide binding function (profile 235) is observed in several folds: tryptophan synthase (c.79), E1 ubiquitin-activating enzymes (c.111). Superfamilies DRE_TIM and Aldolase carried by the TIM-barrel fold are joined by the profile 233 (phosphate binding), indicating evolutionary connections between these functions. Common EFs allow detecting the evolutionary connections between protein domain families even when their structural folds are different, as in the case of eukaryotic RNA-binding KH domain and homologous bacterial KH-2 domain, where the profile 299 with the RNA-binding signature GxxIGxxG connects them.

Towards the descriptor of EF for protein design. Case study of the nucleophile EF

Figures 2–4 illuminate a key role of EFs as the basic building blocks of enzymes, which unravel evolutionary links between different folds and seemingly unrelated protein domain families. These figures also point to an important feature of EFs that can be used in the design of enzymes with desired functions. Specifically, the same EFs can have different sequence signatures and can be carried by different structural elements depending on the superfamily and fold in which they are observed. An obvious question, therefore, is: to what extent and how would it be possible to obtain a generalized description of a certain EF? Probabilistic approach based on the local features of structural segments of proteins has been applied to protein structure prediction and modeling [1, 69], however to our knowledge, the structural restraints of sub-domain-sized segment and sequence conservation profile has never been integrated within the context of protein function. We explored below one of the basic and very common EFs, the nucleophile. Abundance of this function and diversity of the representative sequences and structures will hopefully help us to suggest a unifying description of the EF, which can be used in future design efforts.

A well-known example of the active site that involves the EF of nucleophile is the catalytic triad in proteases [70]. The key player in the triad is the loop carrying a nucleophile, typically a serine, cysteine or threonine residue located in a sharp turn called nucleophilic elbow. The nucleophile side chain has to be activated in order to be reactive. A base, most often represented by a histidine residue deprotonates the nucleophile. The third element of the triad, the acid, is necessary for polarizing the base thus orienting it towards the nucleophile. The anionic intermediate is formed as a result of a nucleophilic attack, and it is stabilized by the oxyanion hole that typically involves the backbone amides. Components of the catalytic triad are characterized by specific structural motifs, and corresponding EFLs have distinctive signatures in their sequences. Figure 5(a) contains profiles representing the signatures of subtilisin and kexin peptidases in peptidases_S8_S53 superfamily (gray circle). Profiles 35, 36, and 37 detect EFLs (orange diamonds) that form the catalytic triad site (blue hexagon in the graph). Optional calcium binding provides additional structural stabilization of the enzyme. Figure 5(b) shows the sequence logos and structures of the EFLs in the catalytic triad of peptidases. Profile 35 with a characteristic D[TSD]G signature represents the catalytic aspartate and several conserved glycines playing structural roles. The base EF is described by the profile 36 with HG[TS][HR][VC]AG[IETV] characteristic signature. The GHSxG signature of the profile 37 contains serine—the nucleophile of the catalytic triad. In general, the catalytic triad is widely spread among protein functions and families. In addition to proteases it can be found in lipases, beta-lactamases, esterases, amidases and acetylases. Despite distinct chemical nature of the hydrolyzed covalent bond in different enzymes, the key EFs—acid, base, and nucleophile—are the same. The comparison of active sites of different hydrolases and transferases becomes possible by using the annotations of their actives sites in the CDD database (figure 5(a)). The EF of the nucleophile in phosphatases is very similar to other hydrolases, although the overall mechanism and the fold are quite different. For example, Protein tyrosine phosphotase (PTPc) and dual specificity phosphatases superfamilies have a characteristic motif HCSAGxGRxG in the active site (described by profile 294), where the cysteine residue executes nucleophilic attack on substrate's phosphate group (figure 5(c)). Esterase_lipase and lipase superfamily possess a hydrolysis active site arranged in a catalytic triad: Ser, His, Asp/Glu. Profile 27 with the signature GxSxG has a nucleophile serine, which is located in the nucleophilic elbow and is characterized by nonstandard backbone torsion angles (figure 5(c)). SGNH_hydrolase is the superfamily of lipases and esterases. However, the active site triad in SGNH_hydrolases resembles the catalytic triad in serine proteases. The nucleophile serine in signature FGDSxxDxG is represented by the profile 175 (figure 5(c)).

The folds of different hydrolases and transferases discussed above, flavodoxin fold (c.23), phosphotyrosine protein phosphatases fold (c.45), α/β hydrolases

fold (c.69) and subtilisin-like fold (c.41) have a three-layered $\alpha/\beta/\alpha$ architecture. However, the numbers of secondary structure elements in these folds are different and the ways the secondary structure elements are connected (i.e. topologies) also differ. The sequence profiles of nucleophile EF are also unique in each functional superfamily. Besides, hydrolase and transferase domains are frequently encountered as parts of multidomain structures with complex functions. Overall, specific environments of different folds predetermine their designability and, thus, potential for accommodating catalytic sites of different sizes and complexity (see figure S7 in supplementary file 1). Together with the above, diversity of sequence signatures and structural characteristics of the nucleophile EF shows that in future design it will be necessary to use the generalized description of the EF. The descriptor of EF could be expressed via a probabilistic model that unifies the features depending on the structure of the fold, its interactions with other domains (in case of multidomain structures), overall biochemical function. Descriptor is therefore a natural development of the representation of EF by adding the structural and interaction restraints to the conservation profile. Here we obtained a library of sequence profiles of EFs. We expect that the concept of EF and the library of descriptors of EFs derived from natural enzymes can be instrumental in future design and engineering efforts.

Discussion

The evolution of contemporary proteins is governed by the same laws of physics and chemistry, as it was in its very beginning. The difference is in the scale of events and their consequences: nowadays it is mostly mutational processes that modify existing functions and switch on/off the silent/native ones in promiscuous enzymes [71–73]. It can sometimes be supplemented by the recombination of functional domains into novel multidomain architectures, or by the formation of new specific folds [74]. There was a totally different scale of events that took place in the very beginning of protein evolution. It was presumably characterized by inefficient catalysis where the primitive prebiotic functions had to adjust to harsh environments and did not have the chance to evolve high substrate specificities. It was also the time of emergence of the first domains as structures that are able to withstand environments, to include diverse catalytic sites, and to support their functioning in form of multistep biochemical transformations involving more than one EF. These first 'events' in the protein structure's life should have resulted in a selection of the most important, and at the same time viable, elements, which paved the way for the protein evolution and left numerous traces in proteins existing today. The size and the shape of EFLs that we observe in contemporary proteins were predetermined by the polymer nature of polypeptide chains, and descendants of prebiotic ring-like peptides with simple functions that are still traceable in modern proteins. Stability in the hot environment of the origin of life and potential for the sequence diversification archetypal for the folds with high designability were presumably the decisive factors for their wide utilization in the emerging protein universe.

How did basic requirements together with three billion years of evolution shape up the structure and function of proteins? Pierre Gilles de Gennes wrote [75, 76] that Jacques Monod in his 1969s provisional lecture proposed that 'it is of some interest to estimate the minimum size required for comparatively long loops of the peptidic chain that linked together amino acid directly involved into the active site of a protein'. Inspired by the Monod's idea, de Gennes have estimated the minimal size of the loop and of the minimal functional domain—both are in a fair agreement with the numbers observed for natural proteins [22, 35, 36, 75, 76] (figure S1 in supplementary file 1). Considering contemporary proteins, we found that the number of EFLs per catalytic domain estimated for CDD database has a Poisson distribution with a mean about 3 (supplementary figure S4). Analysis of the active sites in the CSA shows that the number of active residues in catalytic sites in enzymes is also distributed according to Poisson distribution with a mean around 3.2 (supplementary figure S3 available at stacks.iop.org/PB/12/045002/mmedia). Therefore, the rough estimate suggests one catalytic residue per EFL as a typical building block of the catalytic site with three EFLs in a typical enzymatic domain. We, therefore, updated the schematic representation of an enzymatic receptor initially proposed by de Gennes and Monod (figure S2 in supplementary file 1), providing quantitative characteristics of catalytic site observed in contemporary natural proteins. Both, literature-based annotation of catalytic sites [50] and PSI-BLAST search performed for these sites reveal that most of the domains in modern proteins possess active sites with triads (144/393 occurrences, respectively) and tetrads (145/424) regardless of whether they host many superfamilies and families or not. There are also many active sites with two (dyads) or even one catalytic residue (monads) (see supplementary information tables S1–S2). Remarkably, compositions of the most abundant catalytic sites are also quite specific. According to literature annotation of catalytic sites [50], favorite residues for monads are either Glu, His (most frequent), or Asp. Three most frequent dyads are: Glu–Glu, Arg–Glu, and Asp–Glu; triads: Asp–Lys–Tyr, Arg–Asp–Glu, and Ala–Asp–His; and tetrads: Asp–Gly–His–Tyr, Arg–Asp–Glu, and Ala–Asp–His. Complete data on the statistics of catalytic sites and their compositions are present on web page with additional data (https://github.com/neksa/elementary-functions/tree/master/stats). All the above show that many catalytic functions do not require complex architectures of

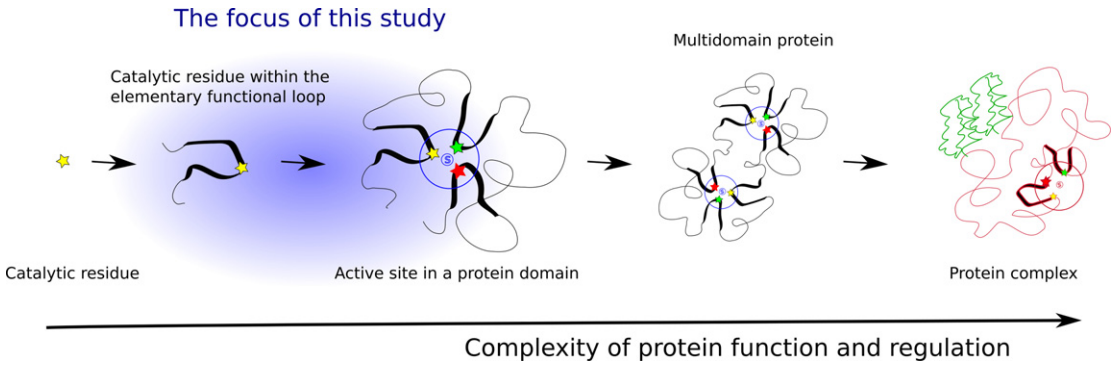

Figure 6. A sketch illustrating the increasing complexity in the structural, functional and regulatory organization of protein function. While domain recombination is characteristic to the later stages in protein evolution, we focus on the earliest evolutionary events and consider how the enzymatic domains themselves emerged.

the sites, apparently providing a strong support for simplified model of the catalytic sites envisioned by Jacques Monod (see for illustration figures S1 and S2 in the supplementary file 1). These pronounced features and restrictions of the catalytic sites of modern proteins show that decomposing natural enzymes into EFLs and learning from them how the function is organized is a feasible way to study the protein function and to develop ideas for how to design it.

The goal of this work was to link the realm of contemporary protein functions to its very origin in the beginning of the biological evolution. First, we wanted to look back and to find what were the EFs that came from the prebiotic world and served as building blocks of first enzymes (figure 6). Our assumption was that the very first EFs were indispensable for basic biochemical reactions, and, therefore, should be widely represented by their descendants in different protein domain superfamilies that may even have different folds. In this case, it would be possible to reconstruct signatures of ancient prototypes starting from the EFLs of contemporary proteins. Then, one can use these signatures along with other profiles of EFLs in the analysis of the protein folds, diversity their functional superfamilies, and evolutionary relations between them. Although our approach for reconstructing the prototypes of EFLs [6, 7] does not include a time line, an ancient character of the derived prototypes is corroborated by the detection of their representatives in many different protein domain superfamilies. That would not be the case if particular EF is invented *de novo* at some time-point of the protein evolution. Protein folds that form hubs in the graph of evolutionary relations (figure 4) are shown to be the ancient ones based on the observation that they are highly populated by different functional families and superfamilies, which is provided by their high designability [26]. The latter has also been shown as an inherent characteristic of the highly thermostable proteins that could survive in the hot conditions of the origin of life [24–26]. Another indication of the old evolutionary age of the 'hub-folds' is that they belong to the core of archaeal COGs, i.e. they provide scaffolds in most of the functional clades/groups-of-proteins [9]. Notably, both EFs and folds that we determined as the ancient ones are in a good agreement with those obtained with the help of phyletic approaches [31, 38, 39, 77] that invoke the time line into consideration. Considering ancient EFs, it should not be surprising that binding of nucleotides and heavy metals as well as cofactors containing them appeared to be the most abundant EFs in modern enzymes. Both nucleotides and metals were present in a prebiotic world, taking part in interactions between prebiotic peptides and nucleotides and working as catalysts. In the emerging DNA/RNA-protein world they became units of the first enzymes, and they have been involved into evolution of protein function ever since [77]. Metals are widely used as cofactors in the catalysis, activators, electron donor/acceptors in redox reactions [62, 63]. High energy ATP/GTP and their derivatives are indispensable sources of energy in numerous biochemical reactions; their phosphoryl groups are used for the phosphorylation. Additionally, nucleotide-containing cofactors work as donors/acceptors of electrons in redox reactions (FAD, FMN, NAD), donors of the methyl group (SAM), prosthetic groups Co-A, pyridoxal 5′-phosphate (PLP), SAM), and catalysts (thiamine diphosohate). The major hubs in the 'graph of the evolutionary relations' (figure 4) are formed by the highly designable folds that are present on the core of archaeal COGs [9], working thus in most of the protein functions. There many α/β folds among them, such as Rossman-like, hydrolase, α/β barrels, α/β sandwiches, SAM methyltransferases. Other representatives of folds with many families and superfamilies include all-α (four-helical bundle and α-α superhelix, and three-helical bundle), all-β (SH3 barrel and Ig-β sandwich), and $\alpha+\beta$ (ferredoxin-lke) folds. Regardless of exact type and content of secondary structure elements, these folds yield the best packing and balance between short- and long-range

stabilizing interactions, adopting numerous low-similarity sequences, and providing, thus, versatile scaffolds for the diversity of enzymatic functions.

To conclude, protein evolution has started from a limited number of basic EFs that arrived from the prebiotic world and formed first enzymatic folds with strong demands on their structures, sizes, stability, designability. Therefore, despite the complexity and diversity of contemporary enzymes, gaining an understanding of the very emergence and early evolution of protein function is very important not only for the basic studies but also for developing a theoretical foundation and computational approaches for future design efforts. The main tasks in protein design are: (i) to introduce a generalized representation of EF that would describe the diversity of its characteristics; (ii) to develop a theory and computational approach for combining EFs into desired enzymatic activities, depending on the predetermined requirements to designed proteins, such as their structure and stability. As a case study, we considered here the nucleophile—a very common EF in many biochemical transformations. It was presumably one of the first EFs that 'walked into' the functional folds from the prebiotic world. In principle, hydroxyl in water is perfectly capable of serving as a nucleophile in hydrolysis, particularly in metallo-hydrolases. However, the geometrical arrangement of the catalytic triad apparently provided a sufficient functional advantage, so that the conservatism of this EF in the catalytic triad in many enzymes corroborates its importance and the ancient nature. At the same time, nucleophile is a very frequent EF, which could have independently emerged in several protein families and at different times of the protein evolution. As a result, we observed that this EF has a quite different sequence signature and distinct structural implementation in PTPc, Lipase, and SGNH_hydrolase superfamilies. In Lipase superfamily, for example, the nucleophile serine yields the non-standard backbone torsion angles, whereas both PTPc and SGNH_hydrolase superfamilies exploit folds and overall mechanisms completely different to those of the typical serine proteases. Additionally, hydrolase/transferase folds containing the EFs of the nucleophile are frequently fused with other domains, forming complex biochemical functions that can, in turn, affect characteristics of the elementary ones. Therefore, in order to achieve a uniformed description of contemporary enzymes and their biochemical mechanisms, one has to develop a generic representation of their basic units—descriptor of the EF. Specifically, every EF should be represented by the set of characteristics, which contain exhaustive information on all possible sequences, structures, functional signatures, interactions etc present in different realizations of this EF. To this end, it should be a probabilistic model for the use in the protein design efforts, where realization of all parameters will depend on the requirements to the designed protein structure and function.

References

[1] Bairoch A 2000 The ENZYME database in 2000 *Nucleic. Acids Res.* **28** 304–5
[2] Holliday G L *et al* 2012 MACiE: exploring the diversity of biochemical reactions *Nucleic Acids Res* **40** D783–9
[3] Holliday G L *et al* 2005 MACiE: a database of enzyme reaction mechanisms *Bioinformatics* **21** 4315–6
[4] Akiva E *et al* 2014 The structure-function linkage database *Nucleic. Acids Res.* **42** D521–30
[5] Andreini C, Bertini I, Cavallaro G, Holliday G L and Thornton J M 2009 Metal-MACiE: a database of metals involved in biological catalysis *Bioinformatics* **25** 2088–9
[6] Goncearenco A and Berezovsky I N 2010 Prototypes of elementary functional loops unravel evolutionary connections between protein functions *Bioinformatics* **26** i497–503
[7] Goncearenco A and Berezovsky I N 2011 Computational reconstruction of primordial prototypes of elementary functional loops in modern proteins *Bioinformatics* **27** 2368–75
[8] Lupas A N, Ponting C P and Russell R B 2001 On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134** 191–203
[9] Goncearenco A and Berezovsky I N 2012 Exploring the evolution of protein function in archaea *BMC Evol. Biol.* **12** 75
[10] Keefe A D and Szostak J W 2001 Functional proteins from a random-sequence library *Nature* **410** 715–8
[11] Carny O and Gazit E 2011 Creating prebiotic sanctuary: self-assembling supramolecular peptide structures bind and stabilize RNA *Orig. Life Evol. Biosph.* **41** 121–32
[12] Gazit E 2007 Self-assembled peptide nanostructures: the design of molecular building blocks and their technological utilization *Chem. Soc. Rev.* **36** 1263–9
[13] Miller S L 1987 Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harb. Symp. Quant. Biol.* **52** 17–27
[14] Trifonov E N, Kirzhner A, Kirzhner V M and Berezovsky I N 2001 Distinct stages of protein evolution as suggested by protein sequence analysis *J. Mol. Evol.* **53** 394–401
[15] Bresler S E and Talmud D L 1944 The nature of globular proteins *C. R. Acad. Sci. URSS* **43** 310–4
[16] Bresler S E and Talmud D L 1944 A few consequences of the new hypothesis *C. R. Acad. Sci. URSS* **43** 349–50
[17] de Gennes P G 1979 *Scailing Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press)
[18] Flory P 1969 *Statistical Mechanics of Chain Molecules* (New York: Interscience)
[19] Svedberg T 1929 Mass and size of protein molecules *Nature* **123** 871
[20] Chothia C 1975 Structural invariants in protein folding *Nature* **254** 304–8
[21] Levitt M and Chothia C 1976 Structural patterns in globular proteins *Nature* **261** 552–8
[22] Berezovsky I N 2003 Discrete structure of van der Waals domains in globular proteins *Protein Eng.* **16** 161–7
[23] Koczyk G and Berezovsky I N 2008 Domain hierarchy and closed loops (DHcL): a server for exploring hierarchy of protein domain structure *Nucleic. Acids Res.* **36** W239–45
[24] Berezovsky I N and Shakhnovich E I 2005 Physics and evolution of thermophilic adaptation *Proc. Natl Acad. Sci. USA* **102** 12742–7
[25] England J L and Shakhnovich E I 2003 Structural determinant of protein designability *Phys. Rev. Lett.* **90** 218101
[26] Zeldovich K B, Berezovsky I N and Shakhnovich E I 2006 Physical origins of protein superfamilies *J. Mol. Biol.* **357** 1335–43
[27] Tawfik D S 2010 Messy biology and the origins of evolutionary innovations *Nat. Chem. Biol.* **6** 692–6

[28] Tokuriki N and Tawfik D S 2009 Protein dynamism and evolvability *Science* **324** 203–7
[29] Tokuriki N and Tawfik D S 2009 Stability effects of mutations and protein evolvability *Curr. Opin. Struct. Biol.* **19** 596–604
[30] Schimmel P R and Flory P J 1967 Conformational energy and configurational statistics of poly-L-proline *Proc. Natl Acad. Sci. USA* **58** 52–9
[31] Caetano-Anolles G *et al* 2009 The origin and evolution of modern metabolism *Int. J. Biochem. Cell Biol.* **41** 285–97
[32] McNaught A D and Wilkinson A I 2014 {UPAC}*Compendium of Chemical Terminology (The 'Gold Book')* 2nd edn (New York: Wiley)
[33] Yamakawa H and Stokmayer W H 1972 Statistical mechanics ofwormlike chains. 2. Excluded volume effects *J. Chem. Phys. Biol.* **57** 2843–54
[34] Shimada J and Yamakawa H 1984 Ring-closure probabilities for twisted wormlike chains. Application to DNA *Macromolecules* **17** 689–98
[35] Berezovsky I N, Grosberg A Y and Trifonov E N 2000 Closed loops of nearly standard size: common basic element of protein structure *FEBS Lett.* **466** 283–6
[36] Berezovsky I N and Trifonov E N 2001 Van der Waals locks: loop-n-lock structure of globular proteins *J. Mol. Biol.* **307** 1419–26
[37] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures *J. Mol. Biol.* **247** 536–40
[38] Caetano-Anolles G *et al* 2008 Origins and evolution of modern biochemistry: insights from genomes and molecular structure *Front Biosci.* **13** 5212–40
[39] Goldman A D, Baross J A and Samudrala R 2012 The enzymatic and metabolic capabilities of early life *PLoS One* **7** e39912
[40] Romero P A and Arnold F H 2009 Exploring protein fitness landscapes by directed evolution *Nat. Rev. Mol. Cell Biol.* **10** 866–76
[41] Glasner M E, Gerlt J A and Babbitt P C 2006 Evolution of enzyme superfamilies *Curr. Opin. Chem. Biol.* **10** 492–7
[42] Makarova K S *et al* 1999 Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell *Genome Res.* **9** 608–28
[43] Nasir A, Kim K M and Caetano-Anolles G 2014 Global patterns of protein domain gain and loss in superkingdoms *PLoS Comput. Biol.* **10** e1003452
[44] Nath N, Mitchell J B and Caetano-Anolles G 2014 The natural history of biocatalytic mechanisms *PLoS Comput. Biol.* **10** e1003642
[45] Furnham N *et al* 2012 Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies *PLoS Comput. Biol.* **8** e1002403
[46] Jencks W P 1987 *Catalysis in Chemistry and Enzymology* ed N Y Mineola (New York : Dover)
[47] Gutteridge A and Thornton J M 2005 Understanding nature's catalytic toolkit *Trends Biochem. Sci.* **30** 622–9
[48] Marchler-Bauer A *et al* 2013 CDD: conserved domains and protein three-dimensional structure *Nucleic. Acids Res.* **41** D348–52
[49] Derbyshire M K, Lanczycki C J, Bryant S H and Marchler-Bauer A 2012 Annotation of functional sites with the conserved domain database. Database (Oxford) 2012: bar058
[50] Furnham N *et al* 2014 The catalytic site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes *Nucleic. Acids Res.* **42** D485–9
[51] Hunter S *et al* 2009 InterPro: the integrative protein signature database *Nucleic. Acids Res.* **37** D211–5
[52] Crooks G E, Hon G, Chandonia J M and Brenner S E 2004 WebLogo: a sequence logo generator *Genome Res.* **14** 1188–90
[53] Tatusov R L, Koonin E V and Lipman D J 1997 A genomic perspective on protein families *Science* **278** 631–7
[54] Andreeva A, Prlic A, Hubbard T J and Murzin A G 2007 SISYPHUS—structural alignments for proteins with non-trivial relationships *Nucleic. Acids Res.* **35** D253–9
[55] Gough J, Karplus K, Hughey R and Chothia C 2001 Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure *J. Mol. Biol.* **313** 903–19
[56] Mistry J, Finn R D, Eddy S R, Bateman A and Punta M 2013 Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions *Nucleic. Acids Res.* **41** e121
[57] Gough J, Karplus K, Hughey R and Chothia C 2001 Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure *J. Mol. Biol.* **313** 903–19
[58] Xie L and Bourne P E 2008 Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments *Proc. Natl Acad. Sci. USA* **105** 5441–6
[59] Brakoulias A and Jackson R M 2004 Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching *Proteins* **56** 250–60
[60] Gold N D and Jackson R M 2006 Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships *J. Mol. Biol.* **355** 1112–24
[61] Kinjo A R and Nakamura H 2009 Comprehensive structural classification of ligand-binding motifs in proteins *Structure* **17** 234–46
[62] Rees D C and Howard J B 2003 The interface between the biological and inorganic worlds: iron–sulfur metalloclusters *Science* **300** 929–31
[63] Harel A, Bromberg Y, Falkowski P G and Bhattacharya D 2014 Evolutionary history of redox metal-binding domains across the tree of life *Proc. Natl Acad. Sci. USA* **111** 7042–7
[64] Fischer J D, Holliday G L, Rahman S A and Thornton J M 2010 The structures and physicochemical properties of organic cofactors in biocatalysis *J. Mol. Biol.* **403** 803–24
[65] Fischer J D, Holliday G L and Thornton J M 2010 The CoFactor database: organic cofactors in enzyme catalysis *Bioinformatics* **26** 2496–7
[66] Gerlt J A, Babbitt P C, Jacobson M P and Almo S C 2012 Divergent evolution in enolase superfamily: strategies for assigning functions *J. Biol. Chem.* **287** 29–34
[67] Gao M and Skolnick J 2012 The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation *Proc. Natl Acad. Sci.* **109** 3784–9
[68] Wong W C, Maurer-Stroh S and Eisenhaber F 2010 More than 1001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology *PLoS Comput. Biol.* **6** e1000867
[69] Henschel A, Winter C, Kim W K and Schroeder M 2007 Using structural motif descriptors for sequence-based binding site prediction *BMC Bioinformatics* **8** S5
[70] Carter P and Wells J A 1988 Dissecting the catalytic triad of a serine protease *Nature* **332** 564–8
[71] Khersonsky O, Malitsky S, Rogachev I and Tawfik D S 2011 Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions *Biochemistry* **50** 2683–90
[72] Khersonsky O, Roodveldt C and Tawfik D S 2006 Enzyme promiscuity: evolutionary and mechanistic aspects *Curr. Opin. Chem. Biol.* **10** 498–508
[73] Khersonsky O and Tawfik D S 2010 Enzyme promiscuity: a mechanistic and evolutionary perspective *Annu. Rev. Biochem.* **79** 471–505
[74] Geer L Y, Domrachev M, Lipman D J and Bryant S H 2002 CDART: protein homology by domain architecture *Genome Res* **12** 1619–23
[75] de Gennes P G 1990 *Introduction to Polymer Dynamics* ed L A R di Brozolo (Cambridge: Cambridge University Press)
[76] de Gennes P G 1998 *Simple Views on Condensed Matter* ed P G de Gennes (Singapore: World Scientific)
[77] Caetano-Anolles K and Caetano-Anolles G 2013 Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism *PLoS One* **8** e59300