Received: 4 January 2021 | Revised: 15 April 2021 | Accepted: 28 April 2021

DOI: 10.1002/prot.26089

RESEARCH ARTICLE

PROTEINS WILEY

Evidence for the preferential reuse of sub-domain motifs in primordial protein folds

Leonhard Heizinger | Rainer Merkl

Institute of Biophysics and Physical Biochemistry, University of Regensburg, Regensburg, Germany

Correspondence
Rainer Merkl, Institute of Biophysics and Physical Biochemistry, University of Regensburg, 93040 Regensburg, Germany.
Email: rainer.merkl@ur.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: ME 2259/4-1, SFB 960

Abstract

A comparison of protein backbones makes clear that not more than approximately 1400 different folds exist, each specifying the three-dimensional topology of a protein domain. Large proteins are composed of specific domain combinations and many domains can accommodate different functions. These findings confirm that the reuse of domains is key for the evolution of multi-domain proteins. If reuse was also the driving force for domain evolution, ancestral fragments of sub-domain size exist that are shared between domains possessing significantly different topologies. For the fully automated detection of putatively ancestral motifs, we developed the algorithm Fragstatt that compares proteins pairwise to identify fragments, that is, instantiations of the same motif. To reach maximal sensitivity, Fragstatt compares sequences by means of cascaded alignments of profile Hidden Markov Models. If the fragment sequences are sufficiently similar, the program determines and scores the structural concordance of the fragments. By analyzing a comprehensive set of proteins from the CATH database, Fragstatt identified 12 532 partially overlapping and structurally similar motifs that clustered to 134 unique motifs. The dissemination of these motifs is limited: We found only two domain topologies that contain two different motifs and generally, these motifs occur in not more than 18% of the CATH topologies. Interestingly, motifs are enriched in topologies that are considered ancestral. Thus, our findings suggest that the reuse of sub-domain sized fragments was relevant in early phases of protein evolution and became less important later on.

KEYWORDS
ancient modules, domain evolution, fold space, protein evolution

1 | INTRODUCTION

Proteins are indispensable to maintain the complex processes constituting life. The relevance and functional scope of proteins is exemplified by the fact that the human proteins consist of appr. 70 000 splice variants[1] and that enzymes catalyze about 4000 reactions.[2] Some complex tasks like protein biosynthesis via ribosomes or the assistance of the assembly of macromolecular structures via chaperones require large multiprotein complexes. However, most functions are fulfilled by proteins that exist as monomers or belong to small hetero-oligomers,[3,4] which requires the existence of a large number of highly specific proteins.

In contrast to the rich diversity of protein functions, the number of unique folds, that is, specific arrangements of major secondary elements with the same topological connections,[5] is drastically smaller. The claim from 1992 by Cyrus Chothia that nature is restricted to

approximately 1000 folds[6] has stood the test of time: The latest version (SCOP 2) of the SCOP database lists not more than 1388 different folds.[7]

Due to their evolutionary relationships, protein sequences and structures resemble each other, if they share a common ancestor. Thus, a certain level of sequence[8] or structure[9] similarity signals homology of two proteins. Additionally, the composition of most proteins complies with a common architectural scheme, comprising one or a sequential series of several building blocks named domains. Domains are independent evolutionary and functional entities, typically consist of 100 to 250 residues and fold into a compact 3D-structure independently of neighboring elements.[10] As evidenced by the Pfam database, more than 18 000 functionally different domains exist in nature[11] and the combination of these basic building blocks most plausibly explains the evolution of larger proteins.[12] Thus, the dominant processes that extend the repertoire of protein functions are duplication of sequences coding for one or several domains, their evolutionary modification, and for some of them their subsequent combination.[13]

Whereas this model for the evolution of multi-domain proteins is generally accepted and experimentally proven,[14,15] the evolution of domains is still a topic of recent studies. The probability that a functional protein can arise de novo has been considered unlikely,[16] which suggests in analogy to the evolution of multi-domain proteins that a reuse of basic building blocks also drives the genesis of individual domains.

In a pioneering analysis, Eck & Dayhoff have identified in ferredoxin a short repeat element binding an iron-sulfur cluster and suggested the doubling of a shorter protein for the evolution of ferredoxin.[17] By screening crystal structures for $C_\alpha - C_\alpha$ contacts less than 10 Å apart, structurally heterogeneous loops with a length of 20 to 50 residues have been detected and proposed as being remnants of prebiotic ring-like elements, whose combination resulted in modern folds.[18] Based on these findings, elementary functional loops (EFLs) have been characterized, these are structural-functional units that possess a closed loop structure and bear one or few residues that are involved in an elementary function.[19] Utilizing position-specific scoring matrices, 525 profiles of EFLs have been specified that include 249 ones involved in binding nucleotide-containing ligands.[20]

Alternatively, methods that compare structures and sequences irrespective of their functional role have been devised for the identification of protein segments of sub-domain size. The comparison of structures is often based on their superposition and the utilization of a similarity measure.[21] An even stronger indicator signaling a common ancestry of proteins is significant sequence similarity.[8] Due to their superior sensitivity, profile hidden Markov models (HMMs) are the method of choice for the detection of remote sequence homology.[22] Although several studies[23-27] used HMMs to account for sequence homology, the number of fragments that were found in similar sets of protein folds differ drastically: The database *Fuzzle* comprises more than 1000 fragments of various length,[26] whereas not more than 40 fragments to be found in 118 folds have been reported by A. Lupas and coworkers.[23]

We were interested in designing and utilizing an algorithm of highest sensitivity, whose application requires minimal user intervention to enable a comprehensive comparison of folds. This is why we implemented Fragstatt, which is based on cascaded HMMs. We concentrated on the analysis of clearly non-homologous proteins and assessed the detected motifs by means of statistical methods. After clustering of highly similar motifs, the large-scale application of Fragstatt identified only few fragments in a small number of putatively ancestral folds.

2 | METHODS

2.1 | Design of Fragstatt

The algorithm Fragstatt was implemented in Python and consists of the four modules *GraphCreator*, *PathFinder*, *PathAnalyzer*, and *MotifDetector*; compare Figure 1. The input are two HHblits cascades *c-tree*(P_i) and *c-tree*(P_j) for the proteins P_i and P_j as roots. The non-root nodes of the trees represent HMMs and the edges local HHblits alignments of two HMMs. The first module, *GraphCreator* combines the two cascades to one graph by identifying common nodes that represent the same protein family. The second module, *PathFinder* identifies candidate paths connecting P_i and P_j by performing a depth-first search with a maximum search depth of four. The graph may contain a large number of parallel edges, which are caused by several HHblits hits that align various subsequences of P_i and P_j. If a graph contains more than 10 000 edges, *PathFinder* randomly samples 10 000 of them. For each candidate path, *PathFinder* determines the overlap between the local alignments and discards all paths overlapping in less than five residues. As a first step, *PathAnalyzer* converts each of the paths identified by *PathFinder* to the multiple sequence alignment $MSA_{i,j}$ by joining the local pairwise alignments generated by HHblits. Subsequently, *PathAnalyzer* creates a matrix $\mathbf{H}_{ij}$, were each $H_{i,j}[k, l]$ indicates how often residues r_i^k and r_j^l from P_i and P_j were aligned in $MSA_{i,j}$. Based on these $H_{i,j}[k, l]$ values, the last module, *MotifDetector*, deduces an optimal traceback by using a modified Smith-Waterman approach.[28] Begin and end of the traceback specify two fragments $F_i^k = P_i[b_{i,k}, e_{i,k}]$ and $F_j^k = P_j[b_{j,k}, e_{j,k}]$ that represent the common motif M_k. The two fragments are then superimposed by means of TM-align to determine their *TM-score*.[9]

2.2 | Determination of prototypical *CA_motifs*

CD-HIT[29] with a cut-off of 90% sequence identity was applied to all entries of *CA_motifs* in order to reduce redundancy on the sequence level and the cluster centers and cluster members were stored. For all pairs of fragments, their *TM-score* was determined and a distance

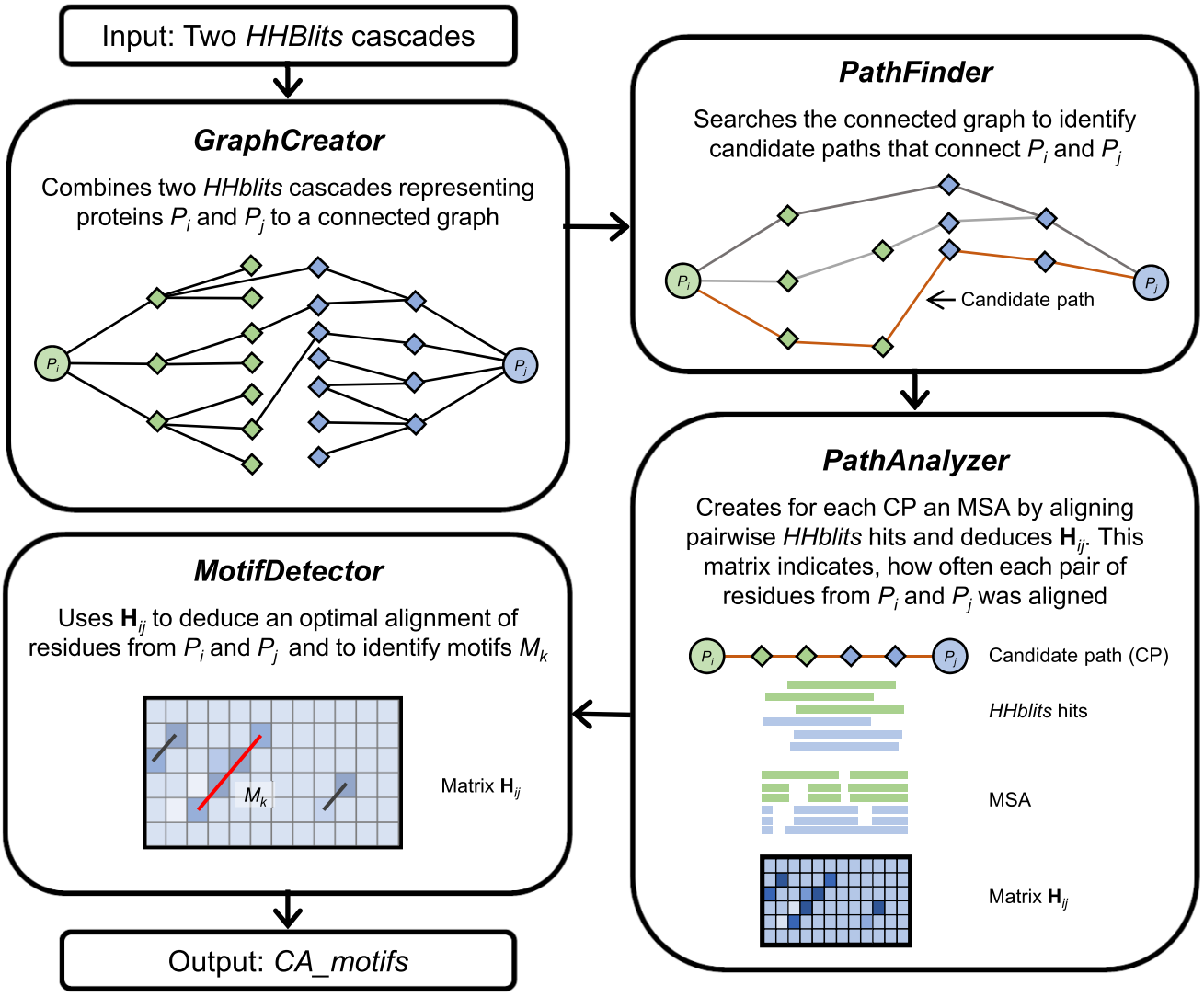

FIGURE 1 Principles of motif detection by means of Fragstatt. The first module, *GraphCreator* takes as input two HMM cascades that is, *c-tree*(P_i) and *c-tree*(P_j) and builds a combined graph by combining corresponding nodes. The module *PathFinder* searches for paths linking P_i and P_j. In this example, three paths were detected. *PathAnalyzer* combines the pairwise alignments of a path to a multiple sequence alignment (MSA) and determines a matrix $\mathbf{H}_{ij}$, where $H_{i,j}[k, l]$ indicates for each pair of residues r_i^k and r_j^l from P_i and P_j how often they were aligned. The last module, *MotifDetector* utilizes $\mathbf{H}_{ij}$ to determine motifs by a traceback similar to a Smith-Waterman algorithm. The corresponding fragments are finally superposed to compute the *TM-score* via TM-align [Color figure can be viewed at wileyonlinelibrary.com]

matrix $\mathbf{D}_{ij}$ was computed for all fragments specifying one of n cluster centers according to

$$D_{ij}[k,l] = \begin{cases} 1 & \text{if } 1.0 - \textit{TM-score}\left(F_i^k, F_j^l\right) > 0.45 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

which implements a minimal *TM-score* of 0.55.

This distance matrix was used for a *DBSCAN* clustering.[30] The parameter *minPts*, which defines the minimum number of neighbors a point must have to be considered as core point, was set to 2. The parameter *E* defines the neighborhood radius and was chosen by optimizing the number of clusters and the average *TM-score* of all pairs within each cluster. Fifty-one percent of the initial motifs could be assigned to a cluster.

2.3 | Searching multi-motif proteins

For each protein P_i, all fragments $F_i^k = P_i[b_{i,k}, e_{i,k}]$ were deduced from *CA_motifs*. To assess the number of positionally distinct sets of fragments, their Hausdorff distance[31] was calculated pairwise according to:

$$d_{Hausdorff}\left(F_i^k, F_i^l\right) = \max(|b_{i,k} - b_{i,l}|, |e_{i,k} - e_{i,l}|) \quad (2)$$

The corresponding distance matrix $\mathbf{D}_{ij}$ was used for a DBSCAN clustering. The parameter *minPts* was set to 2 and *E* was set to 10. For each protein, the result of the clustering was evaluated and those proteins with at least two positionally distinct motifs were saved in a list of multi-motif proteins.

3 | RESULTS

3.1 | Design principles of Fragstatt and benchmarking

For the subsequent usage, we defined the terms motif and fragment as follows: A super-secondary structure, which can be found in at least two proteins will be called a *motif*. The manifestation of a certain motif in a protein will be called a *fragment*. For a protein chain P_i with known sequence and structure, the manifestation F_i^k of motif M_k will be referred to as

$$F_i^k = M_k(P_i) = P_i[b_{i,k}, e_{i,k}] \quad (3)$$

where $b_{i,k}$ and $e_{i,k}$ are the begin and end position (respective residues) of fragment F_i^k in P_i.

Fragstatt (acronym for fragment instantiation) was customized to detect in non-homologous proteins relatively short, but similar subsequences that possess a matching 3D structure. In agreement with previous findings,[23,25] the detected fragments must comprise at least 15 residues, like small super-secondary structure elements. As fragments are considered of sub-domain size, their maximal size was limited to 60 residues. The 3D structures of fragments were compared by means of TM-align; generally, domains with a *TM-score* > 0.5 can be assumed to possess the same fold.[9,21] We regarded motifs as structurally equivalent, if the *TM-score* was ≥0.55, because this cut-off corresponds to a *P*-value of 0.01 which we deduced from a null-model of fragment comparison; see below.

The standard principle for the detection of common motifs relies on a pairwise comparison: Two non-homologous proteins P_i and P_j distinguished by different folds, are scanned for local sequence similarities. If both the sub-sequences and the sub-structures related to two fragments F_i^k and F_j^k in P_i and P_j possess a certain similarity, the fragments are considered as representatives of the motif M_k that corresponds to the mapping $M_k(P_i) \leftrightarrow M_k(P_j)$.

Fragstatt identifies local sequence similarities by means of HHblits,[32] which is a state-of-the-art method for the comparison of HMMs and the detection of remote homology.[33] In order to increase the sensitivity beyond previous studies,[25,34] Fragstatt cascades several HMM searches, because this approach has proven enhanced sensitivity in remote homology detection.[35]

Combining the outcome of cascading HMM alignments (compare Figure 1) related to a protein P_i results in a tree-like structure *c-tree* (P_i), where P_i is the root, and all internal nodes I_i^t are intermediates sharing local alignments with the corresponding parent and the child nodes. For the final step of motif detection, the 3D structures of root proteins P_i are required. Thus, these proteins were taken from PDB.[36] In order to achieve higher coverage, the HMMs specifying internal nodes were taken from Pfam,[11] which is more comprehensive than PDB, as it also contains proteins whose structure is unknown. In a preprocessing step, for each protein P_i, a *c-tree*(P_i) with a depth of three was computed and all hits with an E-value below 1.0 were accepted. A subsequent depth-first search determined for each pair P_i and P_j all paths $P_i \leftrightarrow I_i^{t_1} \leftrightarrow \ldots \leftrightarrow I_j^{t_2} \leftrightarrow P_j$ consisting of maximally four local HMM alignments. If these local alignments were consistently overlapping each other, they were merged to a multiple sequence alignment (MSA). A further processing of the MSA (see Methods [section 2]) led to the identification of fragments $F_i^k = P_i[b_{i,k}, e_{i,k}]$ and $F_j^k = P_j[b_{j,k}, e_{j,k}]$. Finally, the corresponding substructures were compared by means of TM-align[9] and if the *TM-score* was ≥0.55, F_i^k and F_j^k were considered as instantiations of motif M_k.

In order to confirm that the results generated with this fully automated Fragstatt protocol are comparable to those that were manually curated by experts in their fields, we recapitulated two recently reported experiments; for details see Supporting Information (Appendix S1). To begin with, we confirmed that Fragstatt is able to detect a motif shared between the $(\beta\alpha)_8$-barrel and the flavodoxin-like fold.[20] Moreover, we tried to recapitulate a set of 40 putatively ancestral peptide-sized motifs[23] (which we named *AncPept*) that was deduced from a comprehensive comparison of SCOP domains by means of HHsearch. Fragstatt recapitulated 85%, namely 34 of the 40 *AncPept* motifs and most plausibly, due to the manual curation and compilation of *AncPept*, these data were not fully reproducible with the automatic approach of Fragstatt. In summary, we concluded that the sensitivity of Fragstatt is comparable to alternative methods and that Fragstatt is capable of finding shared fragments in a fully automated manner, which was not feasible so far.

3.2 | A large-scale Fragstatt scan found motifs in 26 CATH architectures

As it was our aim to survey most of the "protein universe", that is, a large number of folds observed in nature, we had to carefully select the protein chains to be analyzed. In order to compile a representative but redundancy-free dataset, we opted for an analysis of single-domain proteins and the hierarchical classification of their structures by means of CATH.[37] From top down, the first three CATH levels are *class* (derived from secondary structure content), *architecture* (describing the gross orientation of secondary structures, independent of connectivity) and *topology* (clusters structures into fold groups according to their topological connections and numbers of secondary elements).[38]

The analysis of an initial dataset that was compiled on the lowest CATH level we considered, namely on topologies, gave rise to 77 million protein pairs to be analyzed and around 900 000 (1.2%) motifs that occurred in different CATH topologies were detected by *MotifDetector*. We named this set *CT_motifs*; however, the visual inspection of several motifs made clear that the CATH level

"topology" is inappropriate for the automatic identification of motifs to be found in evolutionary unrelated folds: A large number of these motifs occurred in proteins that are evolutionary closely related although they belong to different CATH topologies. A typical example are beta-propeller proteins that consist of 4 to 12 radially arranged beta-blades.[39] From an architectural point of view, the proteins share a common scheme of composition, namely the annular repetition of blades. It is assumed that beta-propeller proteins arose divergently and evolved by amplification and diversification after the formation of a prototypical blade motif.[39-41] As we were interested to identify motifs shared between evolutionary unrelated folds, we had to choose the next level of the CATH hierarchy, namely the level of different architectures to select protein pairs P_i and P_j. Unfortunately, this choice did not solve the problem completely, because beta-propellers are, depending on the number of blades, grouped into different CATH architectures. We could circumvent this issue by utilizing a "black list" containing combinations of architectures to be removed from the analysis. Note that this list does not exclude the comparison of, for example, propellers with other architectures.

Applying this criterion and accepting only fragments of length 15-60 residues, whose structural superposition resulted in a minimal *TM-score* of 0.55, Fragstatt found 12 532 motifs. This set, which we named *CA_motifs* because it arose by comparing different architectures, comprises 1.4% of the 900 000 *CT_motifs* Fragstatt found by comparing proteins with different CATH topologies. *CA_motifs* occur in 2870 different PDB chains that belong to not more than 26 CATH architectures and 245 CATH topologies. These low rates strongly suggest that motifs are rarely shared between proteins possessing different CATH architectures.

3.3 | *CA_motifs* are spread unevenly among CATH architectures

CATH architectures differ quite drastically in the number of assigned protein domains. Extreme examples are architectures 3.20 (alpha-beta barrel) which includes 16 668 domains and architecture 2.110 (4 propeller) which is comprised of not more than 55 domains. We wanted to confirm that our focusing on single-domain proteins did not introduce a bias and additionally assess the distribution of motifs identified by Fragstatt. This is why we determined for proteins chains their CATH classes and compared the frequency distributions of all CATH entries, single-domain proteins, and the proteins possessing *CT_motifs* or *CA_motifs*. As Figure 2 confirms, the histograms of CATH classes resulting from the full content of the database and the single-domain proteins are nearly identical. The *CT_motifs* are over-represented in the class of mainly alpha and under-represented in the class of mainly beta proteins. In contrast, the distribution of the *CA_motifs* agrees well with the assignment of the full CATH set. Thus, we concluded that by focusing on single-domain proteins, we selected a sample that adequately represents the full set of protein structures. Moreover, we did not observe a striking enrichment of motifs in one of the CATH classes.

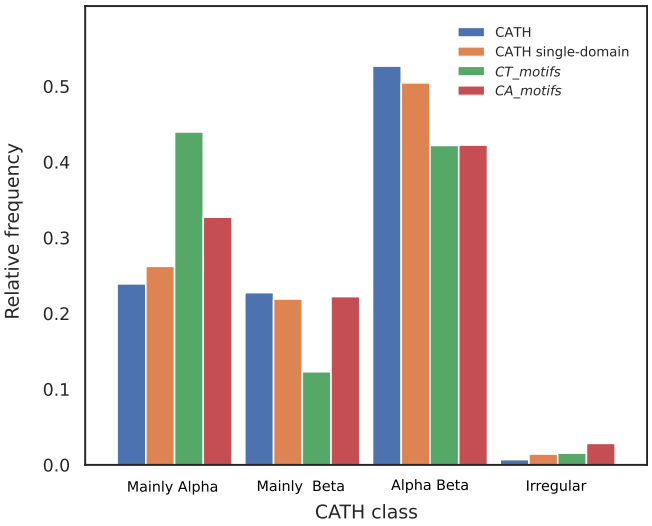

FIGURE 2 Frequency distributions of CATH classes. The bars represent the relative frequencies of the assignments to CATH classes for protein chains from four datasets. Blue: content of the complete CATH database; orange: distribution of CATH single-domain chains; green: distribution of *CT_motifs*; red: distribution of *CA_motifs* [Color figure can be viewed at wileyonlinelibrary.com]

In contrast, the analogous assessment of assigned proteins on the architecture level (compare Figure S3) revealed a severe bias for the full CATH content: The five most populated architectures are 3.40 (3-layer (aba) sandwich), 1.10 (orthogonal bundle), 3.30 (2-layer sandwich), 2.60 (sandwich), and 1.20 (up-down bundle). About 65% of all domains belong to one of these five architectures. Thus, in order to normalize the frequency $f(CA\_motifs_i)$ of a motif determined on the level of a CATH architecture, we calculated an enrichment factor E_i according to:

$$E_i = \log \frac{f(CA\_motifs_i)}{f(CA_i)} \quad (4)$$

where $f(CA_i)$ is the frequency of all single-domain proteins belonging to a certain CATH architecture CA_i.

In Table 1, the architectures are sorted according to the enrichment factor E_i. A positive number indicates that an architecture CA_i is overrepresented among the single-domain proteins contributing *CA_motifs*. In contrast, a negative E_i-value signals that fewer than expected proteins with architecture CA_i possess a *CA_motif*. The overrepresented architectures, ordered from high to low enrichment factor, are 2.140 (8 propeller), 3.20 (alpha-beta barrel), 1.25 (alpha horseshoe), 4.10 (irregular), 2.130 (7 propeller), 2.40 (beta barrel), 1.20 (up-down bundle), 1.10 (orthogonal bundle), and 2.70 (distorted sandwich). Except of architecture 2.140 (8 propeller, $E_i = 1.62$), sparsely populated architectures are underrepresented and the two highly populated "mainly alpha" architectures 1.10 (orthogonal bundle, $E_i = 0.29$) and 1.20 (up-down bundle, $E_i = 0.30$) are enriched. Moreover, most of these overrepresented architectures either show an internal symmetry like alpha-beta barrels or propellers or are repetitive like alpha horseshoes and up-down bundles.

TABLE 1 A comparison of the occurrence of CATH architectures in four datasets related to single-domain proteins

CATH architecture CA_i	*Enrichment* E_i	$f(CA\_motifs_i)$	$f(CA_i)$	$f(CA\_SD_i)$	$f(CT\_motifs_i)$
2.140 (8 propeller)	1.62	0.08	0.04	0.02	0.15
3.20 (alpha-beta barrel)	1.23	13.98	2.43	4.08	7.21
1.25 (alpha horseshoe)	1.13	7.63	1.19	2.45	3.71
4.10 (irregular)	0.80	3.19	0.70	0.43	1.56
2.130 (7 propeller)	0.78	1.26	0.32	0.58	1.63
2.40 (beta barrel)	0.33	5.62	4.99	4.03	2.79
1.20 (up-down bundle)	0.30	10.98	6.91	8.11	10.89
1.10 (orthogonal bundle)	0.29	20.35	15.55	15.25	29.34
2.70 (distorted sandwich)	0.07	0.65	0.60	0.60	0.18
2.120 (6 propeller)	−0.07	0.33	0.10	0.36	1.31
3.40 (3-layer(aba) sandwich)	−0.07	18.78	22.10	20.05	17.42
3.80 (alpha-beta horseshoe)	−0.10	0.47	0.25	0.52	0.17
3.10 (roll)	−0.52	3.54	3.28	5.93	3.91
3.90 (alpha-beta complex)	−0.58	2.19	5.13	3.93	1.87
3.30 (2-layer sandwich)	−0.69	6.87	17.51	13.66	10.08
2.102 (3-layer sandwich)	−0.75	0.06	0.04	0.12	0.61
2.20 (single sheet)	−0.77	0.28	1.16	0.60	1.58
2.30 (roll)	−0.81	1.90	3.10	4.26	1.66
2.10 (ribbon)	−1.08	0.37	1.39	1.08	0.70
2.80 (trefoil)	−1.12	0.14	0.24	0.43	0.03
3.50 (3-layer(bba) sandwich)	−1.76	0.06	1.31	0.37	1.33
2.60 (sandwich)	−1.91	1.24	9.86	8.35	1.25
3.70 (box)	−1.94	0.02	0.01	0.11	0.01
1.50 (alpha/alpha barrel)	−3.88	0.01	0.25	0.39	0.02
2.170 (beta complex)	−4.81	< 0.01	0.40	0.49	0.22
2.160 (3 solenoid)	−4.87	< 0.01	0.28	0.52	< 0.01

Note: The first column denotes the number of a CATH architecture CA_i and its name. E_i is the enrichment factor determined as log-odds ratio according to Equation (2). A positive value indicates for architecture CA_i an overrepresentation among the *CA_motifs*. The following two columns list the relative frequencies determining E_i. $f(CA\_motifs_i)$ is the frequency of single-domain proteins from CA_i contributing a motif at the architecture level and $f(CA_i)$ is the frequency of single-domain proteins belonging to CA_i. $f(CA\_SD_i)$ is the frequency of CA_i among all single-domain proteins and $f(CT\_motifs_i)$ is the frequency of CA_i among the single-domain proteins contributing a motif on the topology level. Rows are sorted according to the value of E_i.

As Fragstatt deduced motifs from protein pairs (P_i, P_j) possessing different CATH architectures, we wanted to know which pairs of architectures were linked, and if some combinations were particularly abundant. We counted how many motifs were detected for each combination of CATH architectures, listed the results in Table S3, and visualized the numbers as a network in Figure 3. Relatively few combinations of architectures constitute the greatest part of all hits: Most abundant are the following five combinations of two CATH architectures: 3.20 (alpha-beta barrel) ↔ 3.40 (3-layer(aba) sandwich), 1.10 (orthogonal bundle) ↔ 3.40 (3-layer(aba) sandwich), 1.20 (up-down bundle) ↔ 1.25 (alpha horseshoe), 1.20 (up-down bundle) ↔ 1.10 (orthogonal bundle), and 1.10 (orthogonal bundle) ↔ 3.30 (2-layer sandwich). This overrepresentation of few folds has a marked impact on the length distribution of the motifs. A strong bias originates from the comparison of CATH architectures 3.20 ↔ 3.40, because the outcome dominates the motifs longer than 40 residues (compare Figure S4).

In sum, five combinations constitute 57% of all motifs. With a frequency of 25%, motifs that occur both in architectures 3.20 (alpha-beta barrel) and 3.40 (3-layer(aba) sandwich) are the most abundant ones and one representant is the $\beta\alpha_2$ motif identified earlier.[25] Additionally, four of the five most frequent combinations belong to the class "mainly alpha"; moreover, symmetrical architectures like 3.40 (alpha-beta barrel) or repetitive architectures like 1.20 (up-down bundle) and 1.25 (alpha horseshoe) are abundant. Of the alpha helical proteins, architecture 1.10 (orthogonal bundle) seems most central, showing strong connections to architectures from all classes; compare Figure 3. The mean number of edges is 8.5, and we assume that the

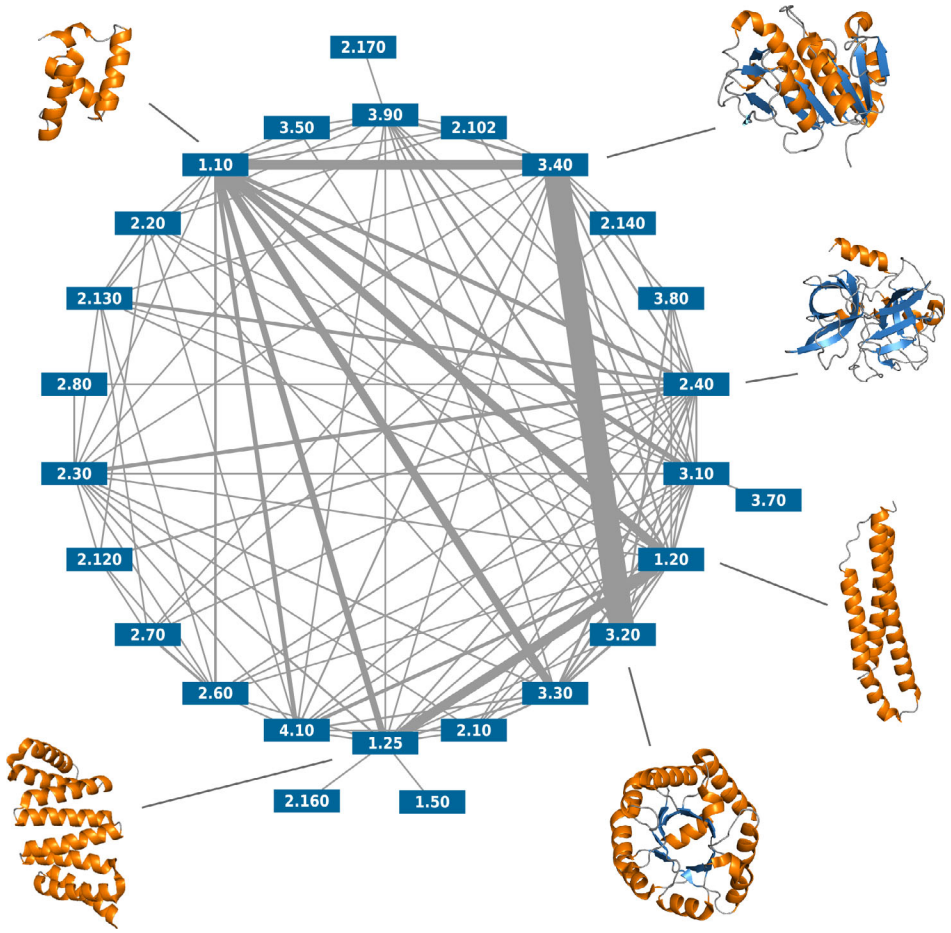

FIGURE 3 Abundance of Fragstatt motifs found in protein pairs of different CATH architectures. The nodes represent CATH architectures, which are connected by an edge, if Fragstatt detected a shared motif in members of these architectures. The width of the edges represents the number of detected motifs [Color figure can be viewed at wileyonlinelibrary.com]

relatively high number of edges of the architectures 2.40 (beta barrel, 18 edges), 3.10 (roll, 17 edges) and 1.20 (up-down bundle, 15 edges) is most likely due to their simple architecture and the existence of short repeating elements (beta strands and alpha helices) which can be found in many other architectures.

Next, we wanted to know whether there is a correlation of the age of folds and the abundance of *CA_motifs*. We used previously calculated values that are based on a maximum parsimony calculation[42] and the NCBI taxonomy tree. These ages of SCOP folds lie in the range [0, 1]; folds assigned an age 1.0 have been estimated to have evolved before the divergence of the superkingdoms.[43] We compiled the age distribution of *CA_motifs* by determining the CATH topology of the related fragments and the correspondence of CATH/SCOP folds by utilizing a Genome3D mapping.[44] We related the age distribution of *CA_motifs* to the age distribution of our full dataset and determined enrichment factors as log-odds ratios of corresponding frequency values (compare Figure 4A,B). Compared to the CATH single-domain dataset, the estimated age of folds bearing *CA_motifs* is significantly higher (*P*-value 1.3E-74, Mann-Whitney *U* test). As indicated by the log-odds ratios, *CA_motifs* are overrepresented in ancestral folds and underrepresented in more recent ones. Most interestingly, 61% of the *CA_motifs* occur in those 52% of all folds that have an age of 1.0, that is, evolved most likely prior to the divergence of the three superkingdoms.

A reason for the overrepresentation of *CA_motifs* in ancestral folds might be the higher flexibility of these folds in supporting demands imposed by different functions. More than 60% of all folds carry at most two enzymatic functions, but the $(\beta\alpha)_8$-barrel and the Rossman fold allow hundreds.[45] Thus, we estimated for the CATH single-domain dataset the designability[46] of each fold, which depends on its stability and compatibility with a large number of sequences by counting the number of related superfamilies $sfam_{fold}$.[47] Using the age distribution introduced above, we created a histogram of mean values $\overline{sfam_{fold}}$; compare Figure 4C. The histogram indicates a strong bias in the fold-specific number of superfamilies and proposes higher flexibility and designability of ancestral folds. The highest $\overline{sfam_{fold}}$ -value, namely 13.5, was observed for folds that have an age of 1.0.

In summary, these analyses show that the majority of *CA_motifs* was found in a small number of more ancient CATH architectures. Among these architectures, mainly alpha helical ones are overrepresented; moreover, the combination of CATH architectures 3.20 (alpha-beta barrel) and 3.40 (3-layer(aba) sandwich) contributes 25% to the *CA_motifs*. On the other hand, the many edges to be seen in Figure 3 indicate that less common motifs were detected in other CATH architectures.

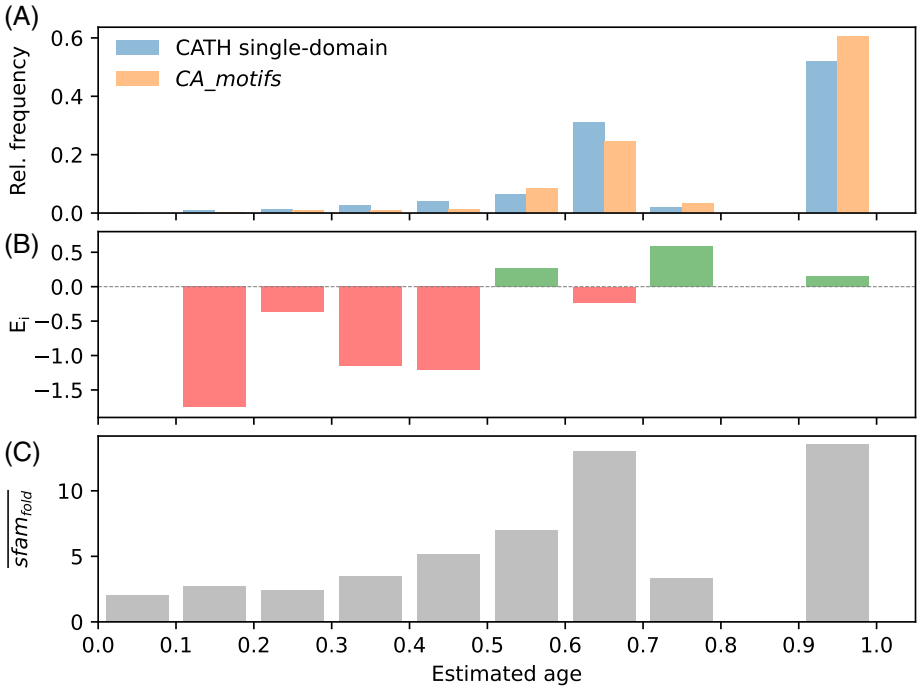

FIGURE 4 Age-specific distribution of folds and related superfamilies. The estimated ages of the folds are normalized to the range 0.0-1.0; folds having assigned an age of 1.0 are considered to have evolved before the divergence of the superkingdoms. (A) The relative frequencies of folds from the CATH single-domain dataset (blue) and from the subset of folds represented in *CA_motifs* (orange). (B) Enrichment E_i of *CA_motifs* indicated by log-odds ratios. A green bar indicates an enrichment and a red bar an underrepresentation of *CA_motifs* in folds of a certain age interval. (C) The age-related average number $\overline{sfam_{fold}}$ of superfamilies per fold determined for the CATH single-domain dataset [Color figure can be viewed at wileyonlinelibrary.com]

3.4 | Structural similarity of the *CA_motifs*

A critical parameter that affects the number of found motifs, is their structural similarity, which scholars consistently determine by computing the *TM-score*.[25,26,34] The *TM-score* is normalized to the range 0 to 1.0, where 1.0 indicates a perfect match between the two structures. Domains with a *TM-score* > 0.5 can be generally assumed to possess the same fold.[9,21] However, it was unclear to us whether this cut-off is adequate for the comparison of fragments of sub-domain size. More specifically, we wanted to ensure that the structural similarity of fragments that gave rise to *CA_motifs* is higher than that one to be expected by chance. Thus, we computed a "null model distribution" consisting of *TM-scores* resulting from a comparison of unrelated fragments. In order to create this fragment set, residue positions were randomly chosen in PDB entries contributing to *CA_motifs*. Subsequently, 12 532 fragments were excised with a length distribution following that of *CA_motifs* (compare Figure S4) and superposed to compute *TM-scores*.

Figure 5 shows the resulting distribution (named *CA_rand*) representing the *TM-scores* under the assumption that the null hypothesis (no structural similarity) is correct. We used *CA_rand* to fix the *TM-score* cut-off for a *P*-value of 0.01, which gave rise to the filter *TM-score* ≥ 0.55 for motif selection.

Figure 5 also shows the *TM-score* distribution *CA_all_motifs*, which results, if Fragstatt is applied to all of our chosen representants of CATH topologies without applying a *TM-score* cut-off for motif selection. The analysis of the two distributions makes clear that the comparison of randomly chosen fragments gives rise to a large number of *TM-scores* that are lower than those observed in *CA_motifs*, the mean values are 0.26 and 0.34, respectively. Both an unpaired *t*-test and a Wilcoxon signed-rank test gave highly significant results with a *P*-value <1E-99.

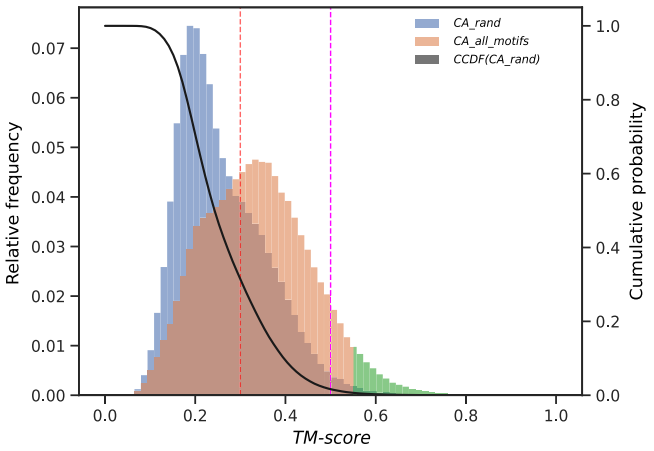

FIGURE 5 Histograms of *TM-scores*. The plots show the relative frequencies of *TM-scores* resulting from a comparison of randomly chosen fragments (set *CA_rand*, blue) and from *CA_all_motifs* (structurally unfiltered representants of CATH architectures, brown). The mean values are 0.26 and 0.34, respectively. The subset of *CA_all_motifs* surpassing the *TM-score* threshold of 0.55 is highlighted in green; these are the *CA_motifs*. *CCDF*(*CA_rand*) corresponds to the normalized right-tail of the *CA_all_motifs* histogram The thresholds 0.3 used to deduce *Fuzzle*,[26] and 0.5 used to deduce *AncPept*[23] are indicated by a red and purple dotted line, respectively

In order to estimate how the probability of Type I errors depends on the *TM-score* cutoff, we additionally determined the complementary cumulative distribution function *CCDF*(*CA_all_motifs*) and plotted it in Figure 5. As can be seen, the fraction of false positive hits increases drastically for *TM-score* cut-offs below 0.5; the empirically determined P-values are 0.01, 0.016, and 0.31, if *TM-score* cut-offs of 0.55 (here), 0.50,[23] and 0.30[26] are chosen. In summary, these results

confirm that most of the motifs identified by means of Fragstatt possess a larger *TM-score* than that expected by chance.

3.5 | Clustering results in 134 prototypical *CA_motifs*

CA_motifs are fragments of sub-domain size that were found by Fragstatt in two proteins with different CATH architectures. However, if the predecessors of the motifs emerged in the pre-LUCA era, it might be that the sequence similarity of modern manifestations of the same motif is too low to deduce homology, even with a highly sensitive HMM-based approach. Thus, it might be that the same or a highly similar motif occurs in several pairs of proteins. This is why we clustered all motifs M_i from *CA_motifs* to identify prototypical motifs M_i^*; for details see Supporting Information.

In summary, 134 prototypical motifs M_i^* were found, which are listed in Table S4. Three motifs (M_1^*, M_2^*, M_3^*) occur in more than four CATH architectures. Nine motifs occur in four ($M_4^* - M_{12}^*$) architectures, 36 motifs ($M_{13}^* - M_{48}^*$) in three architectures and the rest, summing up to 86 motifs ($M_{49}^* - M_{134}^*$), only occur in two architectures. This means that over 90% of the motifs are present in not more than

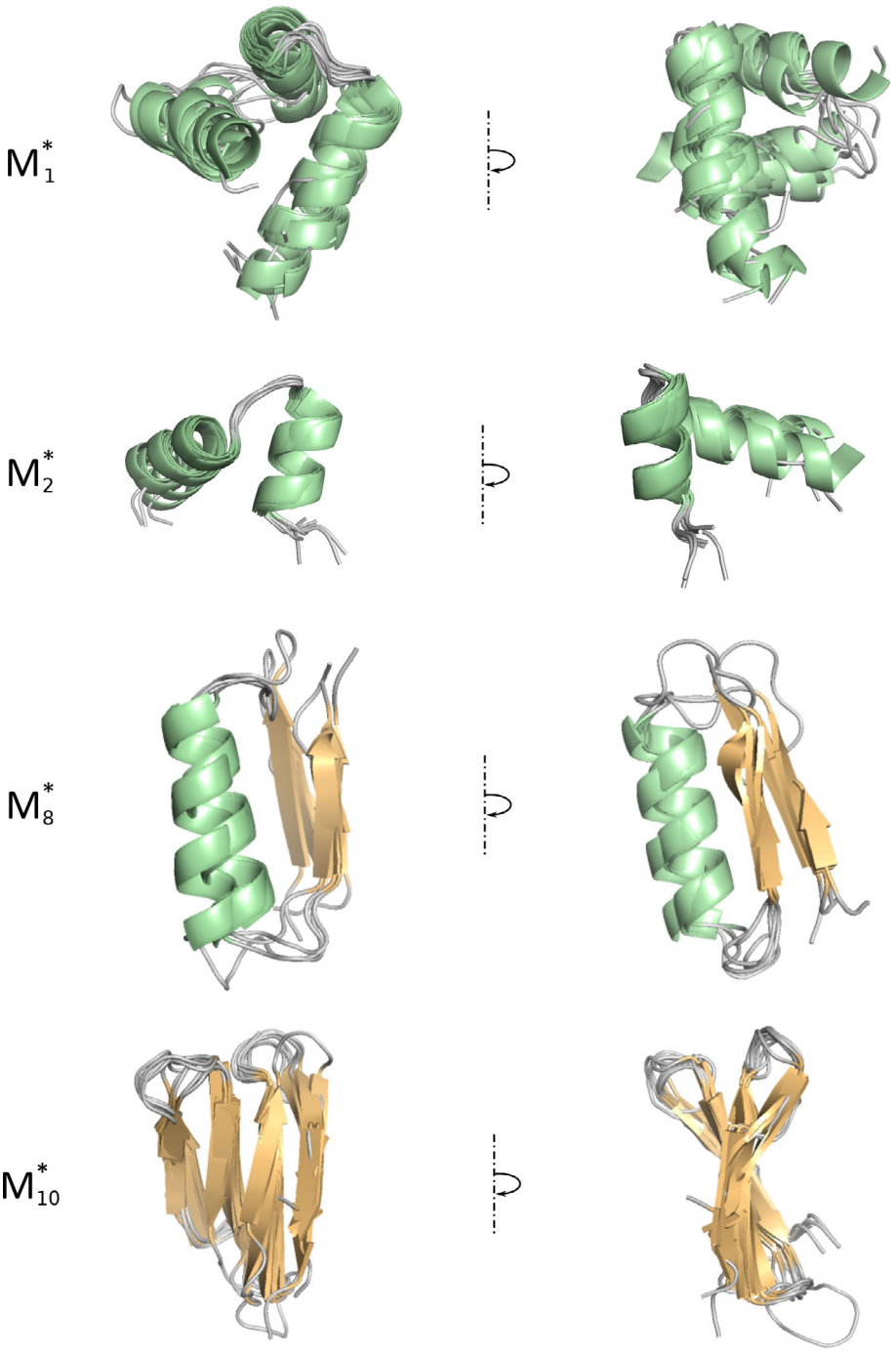

FIGURE 6 Examples of prototypical *CA_motifs*. The selected motifs are shown as a superposition of manifestations (fragments) in cartoon representation. M_1^* and M_2^* are typical α-helix-rich motifs, which are common among *CA_motifs*. The motif M_8^* has a more varied secondary structure, while M_{10}^* is a pure β-sheet [Color figure can be viewed at wileyonlinelibrary.com]

three architectures, at the most. The motifs, which were found in a large number of architectures are rich of alpha helices ($M_1^* - M_6^*$; compare Figure 6). In contrast, motifs with a more diverse secondary structure (like M_8^*) were found in fewer architectures. This was also observed for motifs rich of beta sheets (like M_{10}^*).

The five most abundant CATH architectures among the broadly represented motifs are 1.10 (orthogonal bundle), 3.40 (3-layer(aba) sandwich), 2.40 (beta barrel), 3.30 (2-layer sandwich), and 3.20 (alpha-beta barrel). This observation coincides with the frequencies of CATH architectures in *CA_motifs* and substantiates the finding that most of the motifs were found in a small set of CATH architectures.

3.6 | Functional characterization of prototypical M_i^*-motifs

In order to characterize the functional role of the 134 prototypical motifs, we opted for the annotations of BioLiP (version Feb 05, 2021, 529 047 entries). We selected this database, because some molecules like ethylene glycol or malonic acid are used as additives to solve protein structures and are annotated as ligands. Thus, not all ligands present in the PDB database are biologically relevant and the curators of BioLiP verify the biological relevance of the reported ligands.[48] The ligands bound by all M_i^*-motifs are listed in Table S5 and named according to the nomenclature of BioLiP.

BioLiP reports nucleotide binding for 11 (8%), and metal or ligand binding for 70 (52%) M_i^*-motifs. 30 (%) of the M_i^*-motifs bind at least two ligands and 64 (48%) motifs lack a reported binding function. The most abundant motif M_1^*, which occurs in 11 CATH architectures, is involved in binding nucleotides and HEM (heme). Motif M_7^* occurs in four CATH architectures and binds nucleotides, FES (iron-sulfur cluster), and ZN (zinc). Motif M_8^* binds NAD, NAP, and M7P (7-O-phosphono-D-glycero-alpha-D-manno-heptopyranose). The binding of five ligands is reported for motif M_{69}^*, namely GOL (glycerol), 3HC (3-hydroxybutyryl-coenzyme A), XE (xenon), FE (iron), and MN3 (manganese). Motifs M_{107}^* binds four ligands, namely nucleotides, III (peptides), TFP (trifluoperazine), and CA (calcium). M_{114}^* binds 4NB (4-nitrobenzoic acid), FEO (ferriooxyiron), FE (iron), and ZN (zinc). To a great extent, these findings suggest for M_i^*-*motifs* a role in the coordination of cofactors, which is in agreement with the function deduced for *AncPept* motifs.[23]

3.7 | Identifying proteins possessing more than one motif

If ancestral motifs have been the building blocks for the evolution of proteins domains, it should be possible to find among the proteins contributing to *CA_motifs* some that possess two or more motifs. After the elimination of motifs that correspond to overlapping fragments, we found 135 proteins that contained at least two non-overlapping fragments and all cases were inspected manually. Twenty-two proteins were discarded due to the low quality of the superpositions of the motifs (helix only or too short) leaving 113 cases, which belonged to 15 different CATH topologies; compare Figure 7. 90% of the cases belong to the three CATH topologies 3.20.20 (TIM barrel, 53%), 3.40.50 (Rossmann fold, 21%), and 1.25.40 (serine threonine protein phosphatase 5, 16%). We manually classified the 113 cases according to the three criteria *repetitive*, *overlapping* and *fragmented*; compare Figure S6. *Repetitive* indicates that the same motif was found multiple times at different positions in the same protein chain. *Overlapping* indicates that the positions of the motifs overlap and *fragmented* indicates that inside a larger motif smaller subfragments occurred that were interpreted as individual motif regions.

One hundred six of the 113 cases were repetitive and mainly based on motifs detected between TIM, that is, $(\beta\alpha)_8$-barrels and the Rossmann fold. Among the four cases that were non-repetitive and non-fragmented was the in silico designed protein OR258 (PDB-ID 4J29_A), which was discarded. The remaining three cases will be detailed in the following.

The transcriptional regulator PlcR (PDB-ID 2QFC_A) contains three distinct and separated motifs; compare Figure 8A. This protein belongs to the CATH superfamily 1.25.40.10 (tetratricopeptide repeat domain, TPR) and all superfamily members share the repetition of TPR motifs. PlcR contains two TPR motifs, which were also found by Fragstatt in (a) two DNA-binding proteins (PDB-ID 5K98_B and 3EUS_B) from the CATH superfamily 1.10.260.40 (lambda repressor-like DNA-binding domains), (b) a transcription related protein (PDB-ID 3D3B_A), and (c) a lipid transport related protein (PDB-ID 2RKL_A) from the CATH superfamily 1.10.940.10 (NusB-like) and the CATH superfamily 1.20.5.420 (immunoglobulin FC, subunit C), respectively. The third motif links PlcR to a mitochondrial protein (PDB-ID 1OM2_A), which belongs to the superfamily 1.20.960.10 (mitochondrial outer membrane translocase complex, subunit Tom20 domain). In PlcR, a long α-helix interconnects motifs 1 and 2 with motif 3.

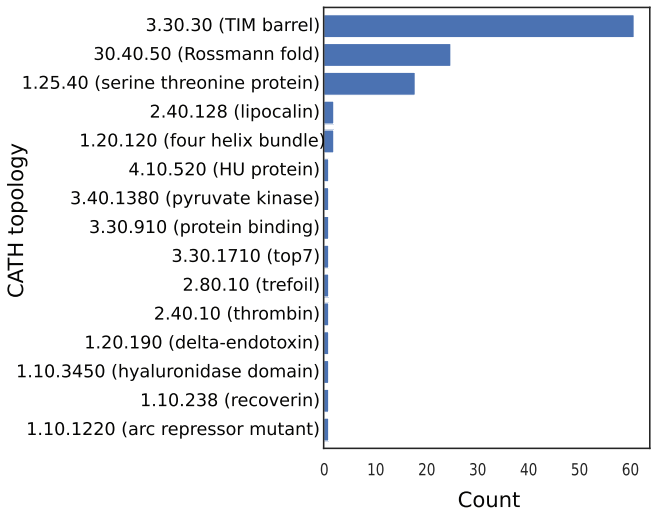

FIGURE 7 CATH topologies of multi-motif proteins. 90% of the cases belong to the three CATH topologies 3.20.20 (TIM barrel, 53%), 3.40.50 (Rossmann fold, 21%) and 1.25.40 (serine threonine protein phosphatase 5, 16%). The histogram lists absolute numbers of multi-motif proteins [Color figure can be viewed at wileyonlinelibrary.com]

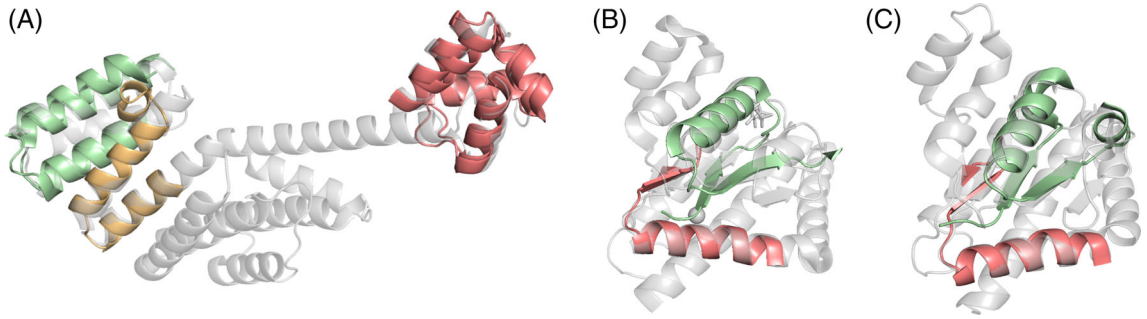

FIGURE 8 Proteins possessing multiple motifs. A, The two repetitive TPR-like motifs of PlcR (PDB-ID 2QFC_A) are shown in green and orange. The manifestations of a third, distinct motif is shown in red. Multiple motifs found in the two homologs B, GmhA (PDB-ID 2XBL_B) and C, diaA (PDB-ID 2YVA_A). For both proteins the same two motifs were detected: A beta-alpha-beta motif (green) and an alpha-beta motif (red) [Color figure can be viewed at wileyonlinelibrary.com]

The isomerase GmhA (PDB-ID 2XBL_B) and the initiator-associating protein diaA (PDB-ID 2YVA_A) both possess a Rossmann fold and belong to the CATH superfamily 3.40.50.10490 (glucose-6-phosphate isomerase like protein; domain 1). In both cases (compare Figure 8B,C) a beta-alpha-beta motif is detected, linking GmhA and diaA to two TIM barrel proteins (PDB-ID 1JCN_A and PDB-ID 1I4N_A). Both TIM barrel proteins belong to the CATH superfamily 3.20.20.70 (aldolase class I). The second motif in the GmhA and diaA, an alpha-beta motif, links them to a DNA-binding protein (PDB-ID 2NDP_A), which belongs to the CATH superfamily 4.10.520.10 (IHF-like DNA-binding proteins).

In summary, by analyzing 309 proteins that contributed to *CA_motifs*, we found only two cases of folds, namely PlcR plus GmhA and diaA (the latter two are members of the same superfamily) that possesses two clearly distinct motifs. In combination with our other results, this finding strongly argues against the frequent recombination of sub-domain sized motifs in the evolution of protein domains.

4 | DISCUSSION

4.1 | Several approaches provide no convincing evidence for the combination of motifs in non-repetitive folds

Since Dayhoff and Eck postulated the hypothesis that modern proteins emerged from the recombination of smaller recurring fragments,[17] it has been shown that repetition plays a fundamental role in evolution.[49] Gene duplication followed by fusion and diversification is a dominant evolutionary mechanism.[50,51] A prototypical example is the $(\beta\alpha)_8$-barrel fold that most likely evolved via a two-step evolutionary pathway from a $(\beta\alpha)_2$-motif constituting a quarter-barrel.[52] Similar cases of a conserved internal symmetry can be observed in various protein folds, for example, beta-propellers,[41] β-trefoils,[53] or TPR repeat folds,[54] which further substantiates the idea that protein domains can emerge from relatively small protein fragments. Based on these observations, the idea of proteins built from smaller modules has been developed; for a review see Söding & Lupas[55].

AncPept,[23] the *bridging themes*,[27] *Fuzzle* [26] and our *CA_motifs* specify four sets of putatively ancestral protein motifs. *AncPept*, the *bridging themes*, and *Fuzzle* recruit motifs from pairwise alignments of HMMs representing SCOP domains. *CA_motifs* are based on HMM cascades and the CATH classification. We opted for this approach to improve detection sensitivity and to explore the effect of alternative structural classification schemes on the detection of putative ancestral protein motifs.

The *AncPept* motifs have been found in 118 (10%) of the 1194 considered SCOP folds.[23] We concluded that this low abundance of primordial motifs is no convincing evidence for the existence of universally used building blocks. This insight has driven the sensitivity-aware design of Fragstatt and prompted us to utilize HMM cascades and low probability HHblits alignments to detect more motifs. We ensured that false positive alignments have no effect by deducing fragments via a traceback; compare the content of the matrix **H** shown in Figure S1. Despite these sensitivity-increasing efforts, the *CA_motifs* occur in not more than 245 (18%) of the 1391 considered CATH topologies, which is less than twice the coverage of the *AncPept* motifs and the utilization of all motifs is far from universal.

A critical parameter with a crucial influence on the number of proposed motifs is the cut-off chosen for the *TM-score*. For the determination of *AncPept* and *CA_motifs*, the cut-off values were 0.5 and 0.55, respectively. In the set of the *bridging themes*, 30% have a TM-score < 0.3 and only 25% a score > 0.5.[27] For the identification of *Fuzzle* motifs, a *TM-score* threshold of 0.3 in combination with an RMSD threshold of 3 Å and no upper or lower bound for the length of the motifs has been used.[26] As a consequence, the average length of *Fuzzle* motifs is 64 residues and the motifs were found in 519 (43%) of the 1221 considered SCOP folds. As Figure 5 demonstrates, a threshold of 0.3 is close to the maximum of randomly sampled *TM-scores*, which is 0.2. Moreover, this cut-off corresponds to a significantly increased false positive rate indicated by a *P*-value of 0.31. Thus, these structural correspondences are from a twilight zone. Nevertheless, it is remarkable that motifs from the largest set, namely *Fuzzle*, have been detected in not more than half of the SCOP folds.

An indicator for the universality of a motif is the number of folds or topologies in which it was detected. The most universal *AncPept* motif has been found in 14[23] and the most universal prototypical *CA_motif* in 13 folds or topologies. However, the majority of *CA_motifs* occur in only two, three or four folds or topologies. For motif

representation, *Fuzzle* utilizes a network[26]: nodes represent proteins containing fragments and links represent motifs. Interestingly, 2% of the proteins contribute 80% of the motifs and the number of links per node follows a power-law distribution, a reuse pattern also observed for motifs deduced from the ECOD database.[24]

In conclusion, these findings strongly suggest that these motifs are no universal building blocks used during the evolution of many different folds. This notion is further supported by the fact that we found only two clear cases of multi-motif folds; compare Figure 8.

4.2 | The usage of motifs seems restricted to early phases of fold evolution

F. Jacob's phrases "Nature is a tinkerer, not an inventor" and "the probability that a functional protein would appear *de novo* by a random association of amino acids is practically zero"[16] emphasize the role of gene duplication and disregard the possibility that new genes might arise de novo. However, the limited spread of motifs in a small number of folds does not provide convincing evidence for the modular structure of protein domains and the frequent reuse of motifs during the evolution of the full fold space.

On the other hand, former results and those presented here are compatible with F. Jacob's idea and suggest that the reuse of short motifs played a role in the origin of protein folds that date back to the formation of RNA-motif complexes in early phases of protein genesis.[23] A mapping of the five most ancient SCOP folds[56] onto the CATH database showed that they belong to the CATH topologies *Rossmann fold*, *TIM barrel*, *Trp operon repressor* and *alpha-beta plaits*. These topologies are all observed among the *CA_motifs* and similarly among the *AncPept* and *Fuzzle* motifs.

The earliest phases of protein evolution must have involved the ab initio invention of new folds, which were flexible enough to support a primordial form of life.[13] The age-specific distribution of $\overline{sfam_{fold}}$-values, which reflects designability, supports this notion; compare Figure 4C. Generally, more than 60% of folds carry out one or two enzymatic functions, but some folds, like the $(\beta\alpha)_8$-barrel or the Rossman fold, catalyze hundreds of reactions.[45] Although these folds are considered ancestral,[56] they contain the most reliably identified motifs. Thus, it is likely that the combination of motifs—often in a repetitive manner—was relevant for the creation of primordial proteins, perhaps by a combination of elementary functional loops.[57,58] It is likely that structure and sequence similarity faded away for those copies of motifs that were less relevant for protein function or stability. Additionally, it might be that due to the evolution of a more sophisticated DNA replication and protein synthesis machinery, the combination of motifs became less likely for the genesis of more recent folds. Thus, in contrast to F. Jacob's intuition, the de novo gene birth seems to continuously contribute to the recruitment of protein-coding sequences as evidenced by recent laboratory experiments.[59-62]

ACKNOWLEDGEMENTS

We thank Reinhard Sterner for continued and generous support. Rainer Merkl received support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via SFB 960 (project Z2b) and the project ME 2259/4-1. The funding sources played no role in the design of the study, collection, analysis, and interpretation of data or in writing the manuscript.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTERESTS

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets supporting the conclusions of this article are available in the https://github.com/merkllab/motif_reuse repository. Program code is available from the corresponding author on reasonable request.

ORCID

Rainer Merkl https://orcid.org/0000-0002-3521-2957

REFERENCES

1. Cunningham F, Achuthan P, Akanni W, et al. Ensembl 2019. *Nucleic Acids Res*. 2019;47(D1):D745-D751.
2. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res*. 2000;28(1):304-305.
3. Goodsell DS, Olson AJ. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*. 2000;29:105-153.
4. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA. Assembly reflects evolution of protein complexes. *Nature*. 2008;453(7199):1262-1265.
5. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247:536-540.
6. Chothia C. Proteins. One thousand families for the molecular biologist. *Nature*. 1992;357(6379):543-544.
7. Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res*. 2020;48(D1):D376-D382.
8. Pearson WR. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics*. 2013;42. https://doi.org/10.1002/0471250953.bi0301s42.
9. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33(7):2302-2309.
10. Doolittle RF. The multiplicity of domains in proteins. *Annu Rev Biochem*. 1995;64:287-314.
11. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019;47(D1):D427-D432.
12. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. *Cell*. 2000;101(6):573-576.
13. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. *Science*. 2003;300(5626):1701-1703.
14. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J 3rd. The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*. 2005;62(4):435-445.
15. Apic G, Russell RB. Domain recombination: a workhorse for evolutionary innovation. *Sci Signal*. 2010;3(139):pe30.
16. Jacob F. Evolution and tinkering. *Science*. 1977;196(4295):1161-1166.
17. Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science*. 1966;152(3720):363-366.
18. Berezovsky IN, Grosberg AY, Trifonov EN. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett*. 2000;466(2–3):283-286.

19. Berezovsky IN, Guarnera E, Zheng Z. Basic units of protein structure, folding, and function. *Prog Biophys Mol Biol*. 2017;128:85-99.
20. Zheng Z, Goncearenco A, Berezovsky IN. Nucleotide binding database NBDB—a collection of sequence motifs with specific protein-ligand interactions. *Nucleic Acids Res*. 2016;44(D1):D301-D307.
21. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*. 2010;26(7):889-895.
22. Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998;14(9):755-763.
23. Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. *Elife*. 2015;4:e09410.
24. Nepomnyachiy S, Ben-Tal N, Kolodny R. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proc Natl Acad Sci U S A*. 2017;114(44):11703-11708.
25. Farías-Rico JA, Schmidt S, Höcker B. Evolutionary relationship of two ancient protein superfolds. *Nat Chem Biol*. 2014;10(9):710-715.
26. Ferruz N, Lobos F, Lemm D, et al. Identification and analysis of natural building blocks for evolution-guided fragment-based protein design. *J Mol Biol*. 2020;432(13):3898-3914.
27. Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N. Bridging themes: short protein segments found in different architectures. *Mol Biol Evol*. 2021. https://doi.org/10.1093/molbev/msab017.
28. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147(1):195-197.
29. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658-1659.
30. Hahsler M, Piekenbrock M, Doran D. Dbscan: fast density-based clustering with R. *J Stat Softw*. 2019;91(1):1-30.
31. Chavent M. A Hausdorff distance between hyper-rectangles for clustering interval data. In: Banks D, McMorris R, Arabie G, eds. *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation*. Berlin: Springer; 2004.
32. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*. 2011;9(2):173-175.
33. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML. Firestar-advances in the prediction of functionally important residues. *Nucleic Acids Res*. 2011;39. https://doi.org/10.1093/nar/gkr437.
34. Alva V, Remmert M, Biegert A, Lupas AN, Söding J. A galaxy of folds. *Protein Sci*. 2010;19(1):124-130.
35. Kaushik S, Nair AG, Mutt E, Subramanian HP, Sowdhamini R. Rapid and enhanced remote homology detection by cascading hidden Markov model searches in sequence space. *Bioinformatics*. 2016;32(3):338-344.
36. Berman HM, Westbrook J, Feng Z, et al. The protein data Bank. *Nucleic Acids Res*. 2000;28(1):235-242.
37. Sillitoe I, Dawson N, Lewis TE, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res*. 2019;47(D1):D280-D284.
38. Greene LH, Lewis TE, Addou S, et al. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res*. 2007;35:D291-D297.
39. Afanasieva E, Chaudhuri I, Martin J, et al. Structural diversity of oligomeric beta-propellers with different numbers of identical blades. *Elife*. 2019;8. https://doi.org/10.7554/eLife.49853.
40. Kopec KO, Lupas AN. Beta-propeller blades as ancestral peptides in protein evolution. *PLoS One*. 2013;8(10):e77074.
41. Chaudhuri I, Söding J, Lupas AN. Evolution of the β-propeller fold. *Proteins*. 2008;71(2):795-803.
42. Edwards H, Abeln S, Deane CM. Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS Comp Biol*. 2013;9(11):e1003325.
43. Winstanley HF, Abeln S, Deane CM. How old is your fold? *Bioinformatics*. 2005;21(Suppl 1):i449-i458.
44. Lewis TE, Sillitoe I, Andreeva A, et al. Genome3D: exploiting structure to help users understand their sequences. *Nucleic Acids Res*. 2015;43:D382-D386.
45. Tóth-Petróczy A, Tawfik DS. The robustness and innovability of protein folds. *Curr Opin Struct Biol*. 2014;26:131-138.
46. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. *Science*. 1996;273(5275):666-669.
47. Goncearenco A, Berezovsky IN. Protein function from its emergence to diversity in contemporary proteins. *Phys Biol*. 2015;12(4):045002.
48. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res*. 2013;41:D1096-D1103.
49. Lee J, Blaber M. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proc Natl Acad Sci U S A*. 2011;108(1):126-130.
50. He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*. 2005;169(2):1157-1164.
51. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. *J Genet*. 2013;92(1):155-161.
52. Richter M, Bosnali M, Carstensen L, et al. Computational and experimental evidence for the evolution of a $(\beta\alpha)_8$-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *J Mol Biol*. 2010;398(5):763-773.
53. Broom A, Doxey AC, Lobsanov YD, et al. Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. *Structure*. 2012;20(1):161-171.
54. Main ER, Lowe AR, Mochrie SG, Jackson SE, Regan L. A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr Opin Struct Biol*. 2005;15(4):464-471.
55. Söding J, Lupas AN. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*. 2003;25(9):837-846.
56. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. The origin, evolution and structure of the protein world. *Biochem J*. 2009;417(3):621-637.
57. Aharonovsky E, Trifonov EN. Sequence structure of van der Waals locks in proteins. *J Biomol Struct Dyn*. 2005;22(5):545-553.
58. Goncearenco A, Berezovsky IN. Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics*. 2010;26(18):i497-i503.
59. Weisman CM, Eddy SR. Gene evolution: getting something from nothing. *Curr Biol*. 2017;27(13):R661-R663.
60. Carvunis AR, Rolland T, Wapinski I, et al. Proto-genes and de novo gene birth. *Nature*. 2012;487(7407):370-374.
61. Schmitz JF, Bornberg-Bauer E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Res*. 2017;6:57.
62. Tretyachenko V, Vymětal J, Bednárová L, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci Rep*. 2017;7(1):15449.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Heizinger L, Merkl R. Evidence for the preferential reuse of sub-domain motifs in primordial protein folds. *Proteins*. 2021;1–13. https://doi.org/10.1002/prot.26089