Multiple Profile Models Extract Features from Protein Sequence Data and Resolve Functional Diversity of Very Different Protein Families

R. Vicedomini,[†,1,2] J.P. Bouly,[†,1,3] E. Laine,[1] A. Falciatore,[1,3] and A. Carbone[*,1,4]

[1]CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative - UMR 7238, Sorbonne Université, 4 place Jussieu, 75005 Paris, France

[2]Institut des Sciences du Calcul et des Données, Sorbonne Université, Paris, France

[3]CNRS, Institut de Biologie Physico-Chimique, Laboratory of Chloroplast Biology and Light Sensing in Microalgae - UMR7141, Sorbonne Université, Paris, France

[4]Institut Universitaire de France, Paris 75005, France

***Corresponding author:** E-mail: alessandra.carbone@lip6.fr.

[†]These authors contributed equally to this work.

Associate editor: Aida Ouangraoua

Abstract

Functional classification of proteins from sequences alone has become a critical bottleneck in understanding the myriad of protein sequences that accumulate in our databases. The great diversity of homologous sequences hides, in many cases, a variety of functional activities that cannot be anticipated. Their identification appears critical for a fundamental understanding of the evolution of living organisms and for biotechnological applications. ProfileView is a sequence-based computational method, designed to functionally classify sets of homologous sequences. It relies on two main ideas: the use of multiple profile models whose construction explores evolutionary information in available databases, and a novel definition of a representation space in which to analyze sequences with multiple profile models combined together. ProfileView classifies protein families by enriching known functional groups with new sequences and discovering new groups and subgroups. We validate ProfileView on seven classes of widespread proteins involved in the interaction with nucleic acids, amino acids and small molecules, and in a large variety of functions and enzymatic reactions. ProfileView agrees with the large set of functional data collected for these proteins from the literature regarding the organization into functional subgroups and residues that characterize the functions. In addition, ProfileView resolves undefined functional classifications and extracts the molecular determinants underlying protein functional diversity, showing its potential to select sequences towards accurate experimental design and discovery of novel biological functions. On protein families with complex domain architecture, ProfileView functional classification reconciles domain combinations, unlike phylogenetic reconstruction. ProfileView proves to outperform the functional classification approach PANTHER, the two k-mer-based methods CUPP and eCAMI and a neural network approach based on Restricted Boltzmann Machines. It overcomes time complexity limitations of the latter.

Key words: **genome, metagenome, evolution, functional classification, protein classification, profile model, profile, cryptochrome, photolyase, photoreceptor, WW domain, glycoside hydrolase, Radical SAM, Haloacid Dehalogenase, B12-binding domain containing, methylthiotransferase, SPASM/twitch domain containing.**

Introduction

Functional classification of biological sequences is fundamental to understanding the ever-increasing genomic and metagenomic sequence data accumulating in our databases. This quest depends on the correct domain annotation of coding genes (Ponting and Dickens 2001; De Filippo et al. 2012; Prakash and Taylor 2012) which, in the past, was handled by sequence homology and feature-based approaches.

The first and most intuitive approach searches for sequences homologous to already known protein or domain sequences (Hawkins et al. 2006; Wass and Sternberg 2008; Loewenstein et al. 2009; Clark and Radivojac 2011; Törönen et al. 2018) and does so either by direct pairwise sequence alignment or by passing through protein signatures, which are descriptions of protein or domain families derived from multiple sequence alignments. It is based on the "orthology-function conjecture" for which orthologs carry out biologically equivalent functions in different organisms, in contrast to paralogs whose functions typically diverge after duplication (Gabaldón and Koonin 2013). Due to complex processes of evolution, many homologs have diversified their functions and the sequence homology approach should be applied with great awareness: different levels of similarity in homology should induce different levels in functional annotation transfer. This represents a serious pitfall for this approach. A second pitfall is linked to the production of profile models, describing features conserved across sequences. Indeed, these families may consist of a few members highly divergent from each other (rare) or a continuum of thousands of sequences due to the absence of strong functional/evolutionary pressure, which challenges the definition of the family and produces totally degenerated super-family/clan models (most frequent) of restrained use.

The second class of methods is based on the selection of an appropriate set of features (such as short sequence segments or wavelet decompositions) (Karchin et al. 2005; Wen et al. 2005; Bonetta and Valentino 2020; Wan and Jones 2020). Other computational schemes use protein structure (Pazos and Sternberg 2004; Pal and Eisenberg 2005; Lee et al. 2007; Dawson et al. 2017), phylogenetics and evolutionary relationships (Eisen 1998; Engelhardt et al. 2005, 2011; Gaudet et al. 2011; Sahraeian et al. 2015; Gumerov and Zhulin 2020), interaction and association data (Deng et al. 2004; Letovsky and Kasif 2003; Vazquez et al. 2003; Nabieva et al. 2005; Sharan et al. 2007; Cao et al. 2014; Pham and Lichtarge 2020), and a combination of these (Shin et al. 2007; Furnham et al. 2012; Boari de Lima et al. 2016; Cao and Cheng 2016; Zhang et al. 2017; Kulmanov and Hoehndorf 2020), with the evident dependence on the availability of different data-types and a large and highly diversified dataset of sequences.

Novel computational approaches that classify sequences by function and overcome the intrinsic limitations of existing methods would help screen sequences to design accurate experiments directed to functional testing and to discover new functions. ProfileView is a computational method conceived for this purpose, capable of classifying hundreds/thousands of homologous sequences into functional groups. It is strongly based on the understanding of the structure of sequence data imposed by the evolutionary history of the sequences. The first main step of ProfileView is to encode functional and structural information belonging to the protein family into multiple profile models that capture the diversity of the homologous sequences in the family. Based on the set of different models for the family, the second main step of ProfileView is to define an original sequence space which organizes sequences by function. Biologically interpretable information and functional motifs are extracted from the classification process. In other words, family members are organized in a tree structure, where subfamily delineation is possible thanks to the hierarchical organization. The presence of multiple functions in a family or subfamily makes it desirable to subdivide its members into smaller groups in order to capture differences in function-related features at a lower level than the subfamily. ProfileView representative models and their specific conserved motifs have proven to be good indicators of this functional delineation. ProfileView can be applied on a large scale to a wide variety of datasets.

In the past, the usage of multiple profile models demonstrated to be powerful in the context of domain annotation (Bernardes et al. 2016; Ugarte et al. 2018), where they have proven be highly accurate on whole genomes and metagenomic/metatranscriptomic datasets, allowing the discovery of new sequences enriching protein families (Fortunato et al. 2016; Amato et al. 2017). Here, we take on a new challenge and use these models to capture the variety of functional motifs characterizing a protein family. Their construction requires a relatively small number of sequences (a minimum of 20), and therefore, they can encode even functional motifs that are poorly represented within large sets of natural sequences, generating a possibly very large motifs diversification.

To highlight its power and generality, we applied ProfileView to seven protein families whose members are characterized by a large functional diversity, multiple members are functionally well-characterized proteins and subfamilies delineations have been validated experimentally together with their functional motifs: the Cryptochrome/Photolyase Family (CPF), the glycoside hydrolase enzymes GH30 family, the enzyme superfamily Haloacid Dehydrogenase (HAD/β-PGM/Phosphatase-like subgroup) and four others (the WW domains and three protein subgroups belonging to the enzyme superfamily Radical SAM, the B12-binding domain containing, the Methylthiotransferase and the SPASM/twitch domain containing). These families and subgroups allowed us to demonstrate the power in feature extraction, the simplicity in the interpretability of the results and the methodological approach, and the computational efficiency of ProfileView compared to a recent artificial neural networks approach to sequence classification (Tubiana et al. 2019). Comparisons are also made with the PANTHER classification system (Mi et al. 2012, 2013), the CUPP (Barrett and Lange 2019) and the eCAMI (Xu et al. 2020) platforms. For each protein family, ProfileView agrees with all available experimental data. For those sequences that were not experimentally validated before, ProfileView provided a functional classification supported by functional motifs.

New Approaches

Proteins carry out their functions primarily through their constituent motifs and domains. Motifs and domains are evolutionarily more conserved than other regions of a protein and tend to evolve as units, which are gained, lost, or combined together as one module (Basu et al. 2009). A domain is a conserved sequence pattern, defined as an independent structural unit. It consists of more than 40 and up to 700 residues, with an average length of 100 residues. A motif is a short conserved sequence pattern usually smaller than a domain. It is often associated with a distinct structural site performing a particular function. Our methodological approach to sequence classification, ProfileView, explores domain functions in proteins, possibly with complex domain architectures, and makes the hypothesis that the set of positions in a protein structure that are essential for its functional activity might have evolved within domains in alternative ways leading to functional differentiation within a protein family. Identifying these differences at the residue level then becomes crucial for functional classification, and ProfileView meets the challenge by identifying the diversity of functional motifs from sequences.

Converting Sequences in Multidimensional Vectors with Profile Models

The ProfileView method is outlined hereafter and illustrated in figure 1. ProfileView takes as input a set of homologous sequences and a protein domain, and returns a classification of the sequences in functional subgroups together with functional motifs contained in the domain and characterizing the subgroups.

The first main idea of ProfileView is to extract conserved patterns from the space of available sequences through the construction of many profile models for a protein domain family that should sample the diversity of the available homologous sequences and reflect shared structural and functional characteristics. These models, called Clade-Centered Models or CCM (Bernardes et al. 2016; Ugarte et al. 2018), are built as conservation profiles from close sequences. Compared to consensus models (e.g., a pHMM Eddy 1998) which are constructed from large sets of homologous sequences including distant ones, CCMs avoid the loss of functional signals due to distant sequences. To build CCMs, we consider the *full* set of Pfam sequences S^i associated with a domain D^i (Finn et al. 2014) and, for each sequence $s_j \in S^i$, we construct a CCM "seeded" from that sequence; if S^i is too large (e.g., comprising tens of thousand sequences), we sample its sequences by first clustering S^i as explained in Materials and Methods. A CCM seeded from s_j is built as a pHMM from a set of UniProt sequences close to s_j (see Materials and Methods). Such a CCM displays features characteristic of s_j and that might differ for $s_k \in S^i$. The more s_j and s_k are divergent, the more the CCMs seeded from them are expected to highlight different features. CCM high specificity, obtained by considering UniProt domain sequences that display a high sequence identity to the seed sequence s_j, captures feature characteristics of protein interaction sites and/or determinants of functional specificity for the protein family. Note that in the past, we constructed CCMs to improve domain annotation (Bernardes et al. 2016; Ugarte et al. 2018) and, for those models, we employed less restrictive conditions for sequence selection in UniProt. In practical terms, this first main idea of ProfileView is implemented into a precompiled library of models associated with the protein domain given as input (fig. 1*A*).

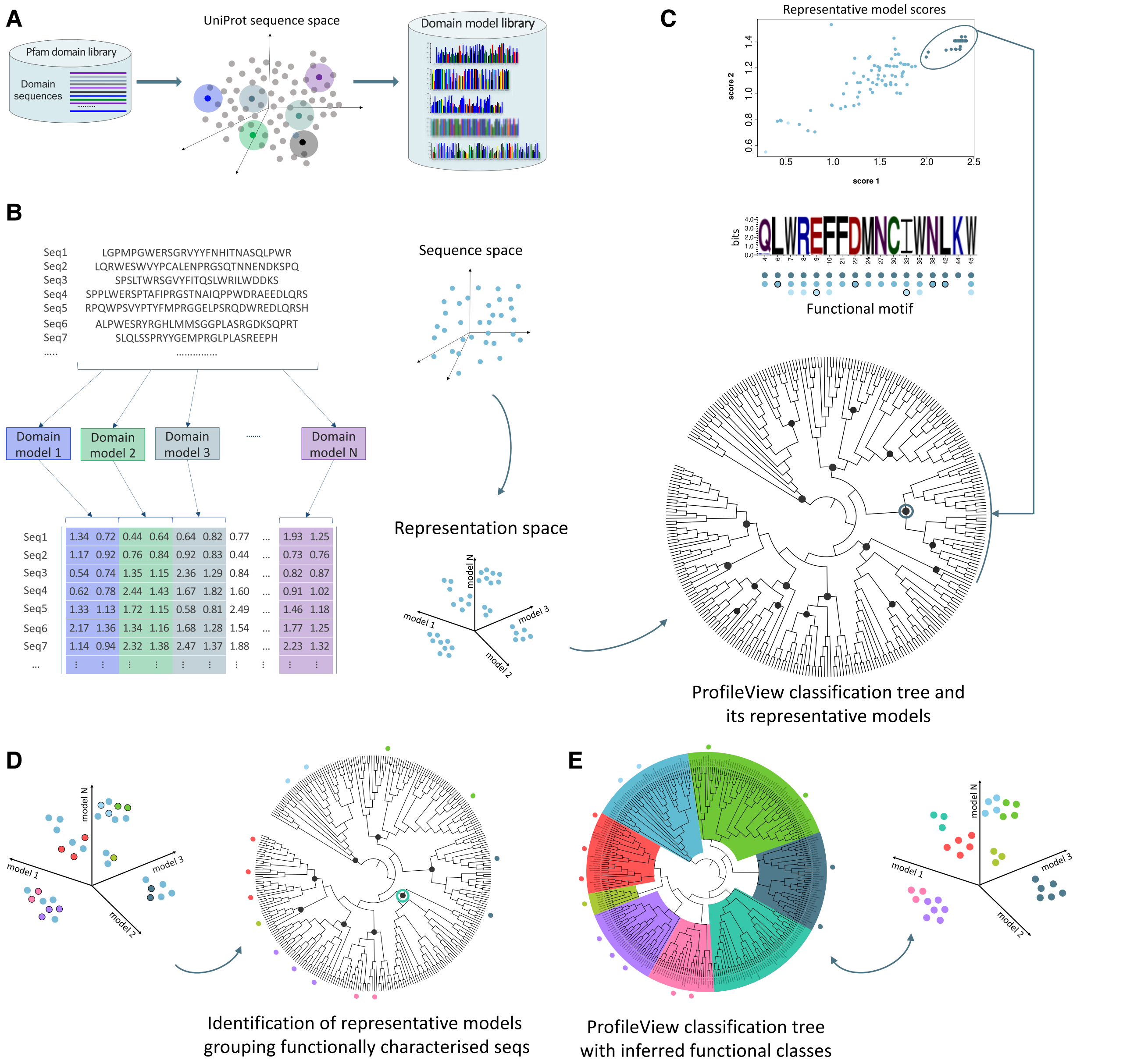
The second main idea of ProfileView is to use CCMs to embed input sequences into a multidimensional representation space, where each dimension is associated with a CCM (fig. 1*B*). Namely, for each input sequence to be classified, each model is matched against the sequence, and the value of the match, expressing how close a model is to the sequence, is recorded as a vector entry (see colored rows in the matrix of fig. 1*B*, left). This space is thought of as a "functional space" because nearby sequences, matching similar profile motifs, are supposed to share the same functional motifs. ProfileView clusters sequences (converted into vectors) within this space by hierarchical clustering and provides a classification tree whose internal nodes are, whenever possible, annotated by representative models and by functional motifs (fig. 1*B*, right, and 1*C*). Subtrees endowed with representative models are indicators of functional specificities and we shall argue that they can be used to subdivide family or subfamily members into smaller groups, in order to capture differences in function-related features of the family, that is, creating groups that preferably include only one function (fig. 1*D*). Such subtrees will be described by functional motifs, that is groups of positions, not necessarily consecutive in the protein sequence, that are conserved in most sequences of the subtree. Positions in a functional motif can be specific to the subtree or shared among subtrees allowing for overlapping positions between representative motifs.

The steps of ProfileView described in figure 1*A*–*C*, identifying subtrees and their representative motifs, do not use any known functional information from experimentally characterized sequences. This latter will be used for validating the method (fig. 1*D*) and for functional inference (fig. 1*E*) while searching for new sequences sharing a known function. All details of the ProfileView pipeline are explained in Method.

Results

Protein Families Analyzed with ProfileView, Validation and Inference

ProfileView was run on seven protein families. Three of the analyzes are detailed below and four in supplementary text 1, Supplementary Material online. The protein sequences to be classified for these families present different characteristics, listed in table 1 (supplementary tables S1, S2, Supplementary Material online) and supplementary text 1, Supplementary Material online. Their sequence length spans from 30aa to 750aa and sequence similarity varies from 30% to more than 50% (supplementary table S1 and text 1, Supplementary Material online). For each family, ProfileView classification is based on one or two similar Pfam domains occurring in their architecture. Length, sequence similarity, and sequence identity for the domain regions contained in the sequences to be classified are reported in supplementary tables S2 and text 1, Supplementary Material online. Number, sequence identity, and sequence similarity of the seed sequences used to construct the ProfileView model libraries are described in supplementary table S3 and text 1, Supplementary Material online. For a comparative view, see supplementary figure S1, Supplementary Material online.


A
Pfam domain library
Domain sequences
UniProt sequence space
Domain model library
B
Seq1 LGPMPGWERSGRVYYFNHITNASQLPWR
Seq2 LQRWESWVYPCALENPRGSQTNNENDKSPQ
Seq3 SPSLTWRSGVYFITQSLWRILWDDKS
Seq4 SPPLWERSPTAFIPRGSTNAIQPPWDRAEEDLQRS
Seq5 RPQWPSVYPTYFMPRGGELPSRQDWREDLQRSH
Seq6 ALPWESRYRGHLMMSGGPLASRGDKSQPRT
Seq7 SLQLSSPRYYGEMPRGLPLASREEPH
....
Domain model 1
Domain model 2
Domain model 3
....
Domain model N
Sequence space
Representation space
model N
model 1
model 2
model 3
C
Representative model scores
score 2
score 1
0.6
0.8
1.0
1.2
1.4
0.5
1.0
1.5
2.0
2.5
bits
Functional motif
ProfileView classification tree and its representative models
D
model N
model 1
model 2
model 3
Identification of representative models grouping functionally characterised seqs
E
model N
model 1
model 2
model 3
ProfileView classification tree with inferred functional classes

Seq1	1.34	0.72	0.44	0.64	0.64	0.82	0.77	...	1.93	1.25
Seq2	1.17	0.92	0.76	0.84	0.92	0.83	0.44	...	0.73	0.76
Seq3	0.54	0.74	1.35	1.15	2.36	1.29	0.84	...	0.82	0.87
Seq4	0.62	0.78	2.44	1.43	1.67	1.82	1.60	...	0.91	1.02
Seq5	1.33	1.13	1.72	1.15	0.58	0.81	2.49	...	1.46	1.18
Seq6	2.17	1.36	1.34	1.16	1.68	1.28	1.54	...	1.77	1.25
Seq7	1.14	0.94	2.32	1.38	2.47	1.37	1.88	...	2.23	1.32
...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Fig. 1. Schema of the ProfileView approach, validation and functional inference. (*A*) Model library construction in ProfileView: representative sequences of the domain under consideration are selected from the Pfam domain library. For each sequence, ProfileView searches for its close homologs in UniProt (colored disks around the seed Pfam sequence, colored dot) and constructs a profile model seeded from that sequence. (*B*) Given a set of unaligned sequences to classify, construction of the representation space: homologous protein sequences (light blue dots in sequence space, center) are encoded into multidimensional vectors by the profile models constructed in A. For each sequence, each profile model contributes two numerical scores to the vector, the normalized bit-score and the normalized weighted bit-score (see V and IX in Materials and Methods). By clustering points in the multidimensional representation space, ProfileView outputs a classification tree of the set of sequences where internal nodes are annotated by representative models (black dots), whenever possible. (*C*) Analysis of the set of sequences in a subtree (see circled root in B) which is endowed with a representative model. The 2-dimensional plot illustrates all sequences in the tree as points defined by their model's two scores. Note that these scores are described in the two columns of the score matrix in B associated with the model. Sequences in the subtree (pastel blue points) are scored the highest. A functional motif, characterizing sequences in the subtree, is associated with the representative model. (*D*) ProfileView validation on a set of sequences with a functionally characterized function. Their position is identified in the tree (colored dots) and the subtrees grouping together sequences of the same functional class are used to evaluate ProfileView. The existence of a representative model is indicated by a black dot. (*E*) Subtrees endowed with representative models in *D* are used to infer a functional classification of sequences locating near functionally characterized ones. The emerald green subtree, endowed with a representative model (*D*), groups no characterized sequence, indicating a potentially new functional class.

ProfileView was validated on seven independent test sets constituted by human curated functionally characterized sequences belonging to these seven protein families. Within a family, sequences are characterized in several functional subgroups, going from a minimum of four to a maximum of nine (table 2 and supplementary text 1, Supplementary Material online). In particular, the difficulty in classification is expected to be nonuniform over

Table 1. Characteristics of the Seven ProfileView Analyzes.

Superfamily/Family	Characteristics of seqs to be Classified			Information on the Model Library Construction			
	#seqs	#filt seqs	#func seqs	Pfam Domain (accession code)	#Pfam seqs	Clust cond	#profile models
Cryptochrome/Photolyase (CPF)*	397	307	72	FAD (PF03441)	4,615	—	4,615
Glycoside hydrolase family 30 (GH30)*	1,803	1,675	695	Glyco-hydro-30 (PF02055) Glyco-hydro-30-2 (PF14587)	1,894	—	1,894
Haloacid Dehalogenase* HAD/β-PGM/ Phosphatase-like	391	259	259	HAD (PF12710) HAD_2 (PF13419)	35,416	≥40%	4,075
WW domain	349	349	54	WW (PF00397)	5,634	—	5,634
B12-binding domain containing	273	258	258	B12-binding (PF02310) B12-binding_2 (PF02607)	12,241	≥60%	3,504
Radical SAM Methylthiotransferase	400	393	393	Radical_SAM (PF04055)	83,232	≥40%	4,501
SPASM/twitch domain containing	128	29	29	SPASM (PF13186)	6,469	≥60%	2,663
	128	115	115	Radical_SAM (PF04055)	83,232	≥40%	4,501

Note.—List of protein families discussed in the main text (starred) and four more discussed in supplementary text 1, Supplementary Material online. For each family, we report some characteristics of the sequences to be classified (number of sequences, number of sequences after filtering (steps II and III of the pipeline), number of sequences with known function) and some information on the model library construction (Pfam domain used in classification, number of Pfam domain sequences, MMseq2 clustering condition (when clustering is applied), number of constructed models). Further features are described in supplementary tables S1, S2, Supplementary Material online. The SPASM/twitch domain containing family is considered twice because classified both with the SPAM domain and the Radical_SAM domain.

Table 2. Summary of ProfileView Performance in Classifying Functionally Characterized Sequences.

Protein Family	# func subgrs	Validation on Subtrees						Table	Figure
		With Models			w/o Models				
		U	M	W	U	M	W		
Cryptochrome/Photolyase (CPF)	5	1	0	71	1	0	71	S4	S2
Glycoside hydrolase family 30 - CAZy families	9	106	5	584	0	1	694	S6	5
HAD/β-PGM/Phosphatase-like	6	0	0	259	0	0	259	S8	6

Note.—For each protein family, the number of functional subgroups used in the evaluation is reported. The total number of unclassified (U), misplaced (M), and well-classified (W) sequences is identified with respect to subtrees endowed with a representative model or not. Names of supplementary tables and figures where ProfileView performance is described in detail, for each functional subgroup, are given.

different functional subgroups. Figure 1*D* describes the validation pipeline (see Materials and Methods).

To validate ProfileView classification, we determined whether functionally characterized sequences of the same functional group localize together in the ProfileView tree. Ideally, one would like ProfileView to split characterized sequences belonging to *n* functional subgroups into *n* distinguished subtrees endowed with representative models (fig. 1*D*). Hence, if sequences belonging to the same functional class are grouped together in a single subtree endowed with a representative model (see Materials and Methods), we consider them well-classified (W in table 2). Some sequences might remain unclassified (U), some others misplaced (M) in subtrees of the wrong functional subgroup, and several sequences of the same functional subgroup might group together in some subtree which is not represented by a model. These different possibilities are indicators of the difficulty in classifying sequences within subgroups. Dropping the condition on the existence of representative models on subtrees, allows to show that, very often, the topology of classification trees groups together unclassified sequences belonging to known functional groups, as for the Glycoside hydrolase family 30 (GH30) for instance (table 2). In table 2 and supplementary text 1, Supplementary Material online, for each protein family, we provide a summary of ProfileView performance by reporting the total number of unclassified, misplaced and well-classified sequences in ProfileView trees. A detailed description of ProfileView performance, for each functional subgroup, is found in supplementary tables and supplementary figures cited in table 2 (supplementary text 1, Supplementary Material online).

ProfileView identified a large number of functionally known positions and specific protein residues in interaction with either nucleic acids, amino acids or small molecules. For two families, the CPF and the GH30, we shall show in detail how ProfileView can provide a functional classification for a large number of functionally uncharacterized sequences (fig. 1*E* and table 2), and novel information on conserved amino acids that could be useful to design testing experiments (see also the detailed analysis of the WW domain family in supplementary text 1, Supplementary Material online). Subtrees endowed with representative models and grouping sequences of a specific function are used to infer the function for all sequences in the subtree, and subtrees endowed with representative models and grouping sequences with no functional characterization (e.g., emerald green tree in fig. 1*E*) are used as indicators of potentially new functional classes.

ProfileView on the CPF Family

The CPF, involved in the interaction with nucleic acids, amino acids, and small molecules, is widely distributed in all kingdoms of life (Sancar 2003; Brettel and Byrdin 2010; Chaves et al. 2011; Jaubert et al. 2017). CPF members share the same fold, yet can perform very different functions and have completely different partners: cryptochromes (CRY) are mainly photoreceptors (PR) using light to activate specific signaling pathways; some CRY also acts as light-independent transcriptional regulators of the circadian clock; photolyases (PL) are light-activated enzymes repairing UV-damaged DNA (cyclobutane pyrimidine dimer (CPD) lesions or (6-4) lesions). There are five main CPF functional groups: circadian, (6-4) photolyase, CPD photolyase, ssDNA photolyase, and photoreceptor. (See supplementary text 2, Supplementary Material online, section 1, for more description.) On the other hand, sequence similarity highlighted a finer classification of CPF proteins splitting the five groups in several subgroups (see, for instance, Emmerich et al. 2020). Based on the current literature, we could identify 11 distinct subgroups: (6-4) photolyase, Animal photoreceptor cryptochrome (PR CRY), transcriptional regulator, CRY DASH, CRY Pro, Classes I, II, III CPD photolyase, Plant-like photoreceptor CRY, Plant photoreceptor CRY, and a new NCRY subgroup. Some of these subgroups of sequences have been experimentally characterized to share the same function, as (6-4) photolyase and CRY Pro (fig. 2*B*, supplementary fig. S2, Supplementary Material online), and some others are associated with very similar sequences, as (6-4) photolyase, Animal photoreceptor CRY, and transcriptional regulator (supplementary figs. S3, S4, Supplementary Material online).

In our analysis, we make the hypothesis that the FAD (flavin adenine dinucleotide) binding domain, occurring in all CPF sequences, contains all functional information leading to a functional diversification of the family. Indeed, all CPFs noncovalently bind FAD and share a mechanism of FAD photoreduction by intra-protein electron transfer (Björn 2015). FAD can be in different oxidation and protonation states (Sancar 2003), specifically associated with different functions. The FAD domain is known to interact specifically either with the damaged DNA, with other domains present in CPF proteins (e.g., C-ter extensions in some photoreceptor cryptochromes) or with other protein partners (Czarna et al. 2013).

ProfileView is validated on two different types of data: functionally characterized CPF sequences and functionally characterized positions within CPF sequences. These latter are compiled in a manually curated list of positions (supplementary file, Supplementary Material online

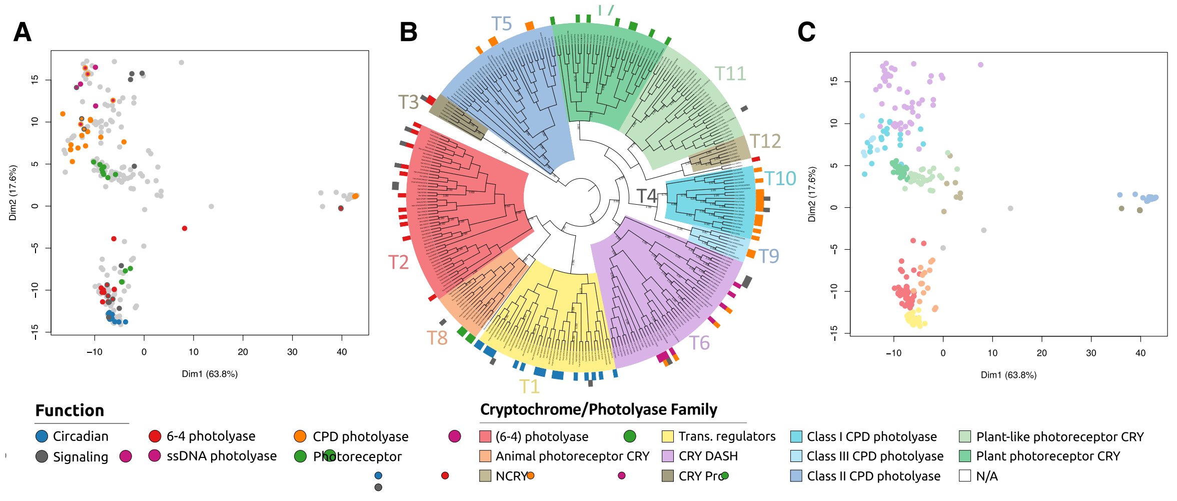


Fig. 2. ProfileView representation space for the CPF family, classification tree, validation on experimental data, and inference. (*A*) Two-dimensional projection of the ProfileView representation space for 307 FAD-binding domain CPF sequences obtained by Principle Component Analysis (PCA). The axes correspond to the first and second PCA components explaining the 63.8% and 17.6% of the dispersion, respectively. Seventy-two experimentally functionally classified sequences are colored (legend "Function"). Unclassified sequences are left light gray. When a sequence is known to have a double function, both colors are indicated and the inside color refers to the known primary function. For instance, five of the eight ssDNA photolyase sequences (red purple) located on the top of the plot are double function (compare to the first ring in B). (*B*) ProfileView classification tree. External colored squares define known functions for the sequences (legend "Function"). Some functionally characterized sequences are known to hold multiple functions and are labeled by two colors. The function "signaling" (dark gray) refers to signaling processes of different nature (photoreceptor, transcription, unknown). Numbers on the internal nodes correspond to the percentage of sequences in the corresponding subtree that are separated from the remaining sequences in the tree by the a representative model occurring in the model library (see supplementary fig. S2, Supplementary Material online for details). Colored subtrees are identified by representative models and they correspond to known CPF classes (legend "Cryptochrome/Photolyase Family"), with the exception of the NCRY subtree. (*C*) Inferred function for unclassified sequences (gray dots in A), where colors (legend "CPF") correspond to the identified subtrees endowed with representative models on the ProfileView tree in B.

"CPF_mutants_used_for_validation.xlsx") from the literature. Furthermore, we combined them with structural modeling to analyze CPF subgroups in detail.

Validation of ProfileView on the Functional Diversity of CPF Members

The ProfileView representation space shows a coherent organization, where sequences with the same functional characterization (see "Function" in fig. 2*A* and *B* and supplementary fig. S2, Supplementary Material online) tend to occur together in space (fig. 2*A*, table 2). Their localization is analyzed further in the ProfileView classification tree, comprised of 11 subtrees ($T_1 \ldots T_3$, $T_5 \ldots T_{12}$ in fig. 2*B* and supplementary fig. S2, Supplementary Material online). We observe an almost perfect split of 71 out of 72 functionally characterized CPF sequences (see labels in the outer circles of fig. 2*B*, colors as in "Function") across the eight distinguished subtrees $T_1 \ldots T_8$, all endowed with a representative model. Their classification in the five CPF functional classes is described in supplementary table S4, Supplementary Material online. Note that the (6-4) photolyase class divides in two subtrees, distinguishing (6-4) photolyase sequences (T_2) from the known CRY Pro sequences (T_3). This split, also recognized in the phylogenetic trees of the CPF family (supplementary figs. S3 and S4, Supplementary Material online; see also figs. 3*A*, 3*C* and 3*D*), is due to sequence divergence highlighting specific traits of the CRY Pro sequences such as the four FeS-binding cysteines which are missing in the (6-4) photolyase subgroup (Ma et al. 2019). In contrast, the CPD photolyase class divides in two distinguished subtrees corresponding to the subgroups Classes I and III CPD photolyase (T_4) and Class II CPD photolyase (T_5), which are not identified in the phylogenetic trees of the CPF family. In particular, ProfileView further divides Classes I and III CPD photolyase into two subtrees, one for Class I CPD photolyase (T_{10}) and the other for Class III CPD photolyase (T_9). Finally, the photoreceptor sequences are divided in two subtrees, one grouping Animal photoreceptor CRY (T_8) and the other Plant photoreceptor CRY (T_7).

Most importantly, at the root, the ProfileView tree topology organizes large subtrees consistently with known functional subgroups (fig. 3*A*). Namely, the ProfileView tree separates light-independent circadian transcriptional regulator CRY (T_1) from the light-dependent (6-4) photolyase (T_2) and Animal photoreceptor CRY (T_8; fig. 3*E*). It also clearly separates the DNA repair (6-4) photolyase from the Animal photoreceptor CRY. It reconciles classes I and III CPD photolyase into a single subtree (T_4), while keeping them distinct (T_9, T_{10}), and it clearly separates them from Plant (T_7) and Plant-like photoreceptor CRYs (T_{11}; fig. 3*F*). For the characterized sequences displaying double function (supplementary fig. S2, Supplementary Material online), their DNA repair/photolyase activity (either CPD or (6-4)) is consistently determined by ProfileView that groups these sequences in the photolyase subtrees. At the best of our knowledge, these sharp separations, in agreement with known functional characterizations, have never been obtained by sequence analysis before.

Interestingly, the ProfileView tree allowed for the identification of a yet functionally uncharacterized subtree (T_{12}, named NCRY; see the light beige subtree in figs. 2*B*

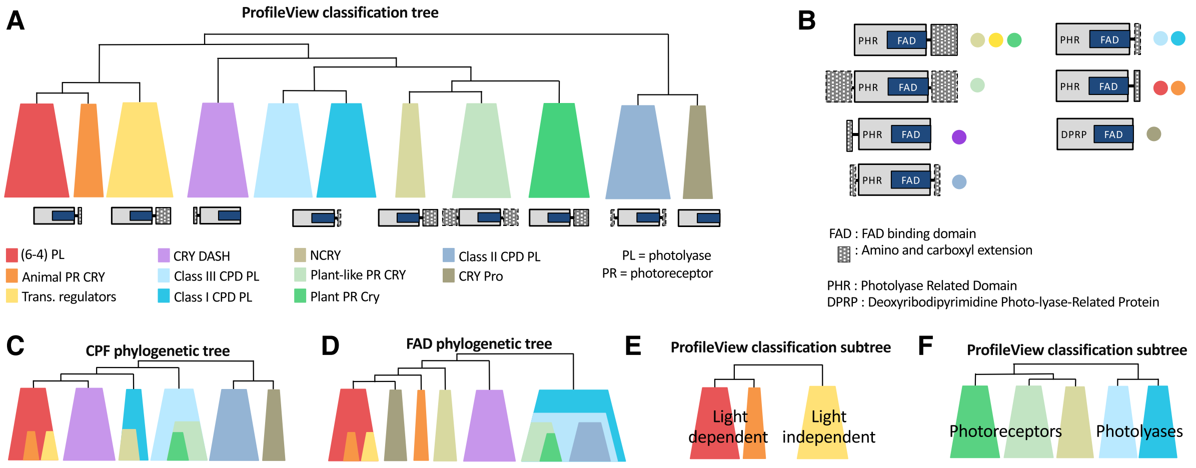


Fig. 3. Topological comparison between the ProfileView classification tree and the phylogenetic trees for the CPF family and the FAD-binding domain. (*A*) Schema illustrating the topological structure of the ProfileView tree in figure 2*B* and supplementary figure S2, Supplementary Material online. Colors correspond to subtrees where the characterized sequences of the same functional group are over-represented (bottom). The domain architectures known to be characteristic of each subtree are reported (see *B* for more details). (*B*) Domain architectures for proteins belonging to different subtrees of *A* are reported (colors as in *A*). C- and N-terminal regions are indicated with gray boxes. Dashed border lines indicate terminal regions present only occasionally in an architecture. (*C*) Scheme of the main topological structure of the CPF phylogenetic tree constructed from the 307 CPF sequences containing the FAD binding domain. Colors as in *A*. See the CPF phylogenetic tree in supplementary figure S3, Supplementary Material online. (*D*) Scheme of the main topological structure of the FAD phylogenetic tree constructed from the 307 FAD-binding domain sequences. Colors as in *A*. See the FAD phylogenetic tree in supplementary figure S4, Supplementary Material online. (*E*,*F*) Two zooms on subtrees of the ProfileView classification tree involving classes of CPF sequences described in *A*. Colors as in *A*.

and 3A) of proteins showing strong sequence divergence. The same subtree was also identified by sequence similarity network analysis in (Emmerich et al. 2020) without inferring any functional classification for it, and by the phylogenetic tree based on the FAD-binding domain in CPF sequences (FAD tree, for short; fig. 3D) which places it close to the Animal photoreceptor CRY (T_8) and CRY DASH (T_6). ProfileView positions NCRY close to the Plant photoreceptor CRY and Plant-like photoreceptor CRY (green subtrees in fig. 3A). In contrast, the phylogenetic tree of CPF sequences (CPF tree, for short) includes NCRY within Class I CPD photolyase (cyan subtree in fig. 3B). To our knowledge only one protein from this family has been characterized and it was shown to bind FAD but to lack DNA repair/photolyase activity (Worthington et al. 2003) which is in accordance with the position of this family in our functional tree. This finding highlights the potential of ProfileView to reveal novel functional classes within a protein family. (See also supplementary fig. S5, Supplementary Material online.)

Overall, the 11 subtrees in figure 3A ($T_1 \ldots T_3$, $T_5 \ldots T_{12}$ in figure 2B and supplementary fig. S2, Supplementary Material online) are uniquely associated with known functional classes (see supplementary table S4, Supplementary Material online). This provides the first proof of the method's classification power for inferring known functions and suggesting potentially new ones.

Comparison of the ProfileView Tree with the FAD and CPF Phylogenetic Trees

The comparison of ProfileView classification tree (fig. 2B) with the CPF tree (supplementary fig. S3, Supplementary Material online) and the FAD tree (supplementary fig. S4, Supplementary Material online) highlights important differences in the topological organization of major functional classes. A drawing in figures 3A, 3C and 3D compares the three trees for easy visualization. We notice that the CPF phylogenetic tree (fig. 3C): (1) incorrectly groups sequences exhibiting disparate functions, for instance Plant photoreceptor CRY and Plant-like photoreceptor CRY are clustered within Class III CPD photolyase; (2) hides the NCRY subtree within class I CPD photolyase; (3) mixes light-dependent and light-independent proteins in a subtree where Animal photoreceptor CRY and circadian transcriptional regulator are clustered within (6-4) photolyase sequences.

It is interesting to notice that the compatibility of domain architectures associated with different functional classes of CPF sequences (fig. 3B) is coherent with the ProfileView tree topology (fig. 3A, bottom) and much less so with the CPF phylogenetic tree. Compare, for instance, the architectures for the classes Plant-like photoreceptor CRY, Plant photoreceptor CRY and NCRY, or those for classes I and III CPD photolyase. All members of these classes have a PHR domain in which a specific CPF FAD-binding domain is found, but C- and N-terminal extensions of variable sequence or length. The architectures for Plant-like photoreceptor CRY, Plant photoreceptor CRY and NCRY possess N- or C-terminal extensions whereas classes I and III CPD photolyase only possess the PHR domain. Classes which are topologically close in the ProfileView tree preserve sequence/length characteristics of C- and N-terminal regions and agree with what is expected in contrast to the subtrees of the CPF phylogenetic tree. Similar observations can be highlighted by comparing the ProfileView tree with the FAD phylogenetic tree (fig. 3D).

Representative Models, Motifs and Validation of ProfileView on Functionally Characterized Positions

ProfileView associates representative models and functional motifs to the subtrees of its classification tree. They are used to highlight subfamily delineations and molecular determinants underlying functions and interactions, respectively.

A representative model for a subtree of the ProfileView tree is a profile model that ideally "separates" the sequences in a subtree from all other sequences in the tree (step IX of ProfileView pipeline in Materials and Methods). Representative models can be used to subdivide members of a family or subfamily into smaller groups to capture function-related differences at a lower level of the ProfileView tree, that is, creating groups that preferably include only one function. CPF subtrees corresponding to known functional classes in figure 2B and supplementary figure S2, Supplementary Material online are characterized by representative models that separate at least 50% of the sequences in a subtree from all other sequences in the ProfileView tree (see labels reporting the proportion of sequences supported by a model on the nodes in supplementary fig. S2, Supplementary Material online). An automatic procedure in ProfileView identifies representative models.

Given a representative model for a subtree, the set of conserved positions in the model uniquely defines a motif for the subtree (step X of ProfileView pipeline in Materials and Methods). The motifs associated with the 11 CPF functional subtrees are reported in supplementary figures S6, S7, Supplementary Material online with the exception of Classes I and III CPD photolyase, known to share the same function, which we grouped together by considering the representative model of the minimal subtree including both classes. The only subtree associated with two distinct representative models, covering two different regions of the FAD-binding domain sequence, is the light-independent transcriptional regulator tree (supplementary fig. S8, Supplementary Material online, fig. 4A and B). Figure 4C shows the alignment of the two transcriptional regulator motifs with the (6-4) photolyase motif and the Animal photoreceptor CRY motif, where positions 50–126 are not covered by the two transcriptional regulator motifs. These positions comprise the FAD-binding domain region directly involved in proton or electron transfer to the FAD chromophore and provide evidence that proton/electron transfer is not involved in the function of light-independent transcriptional

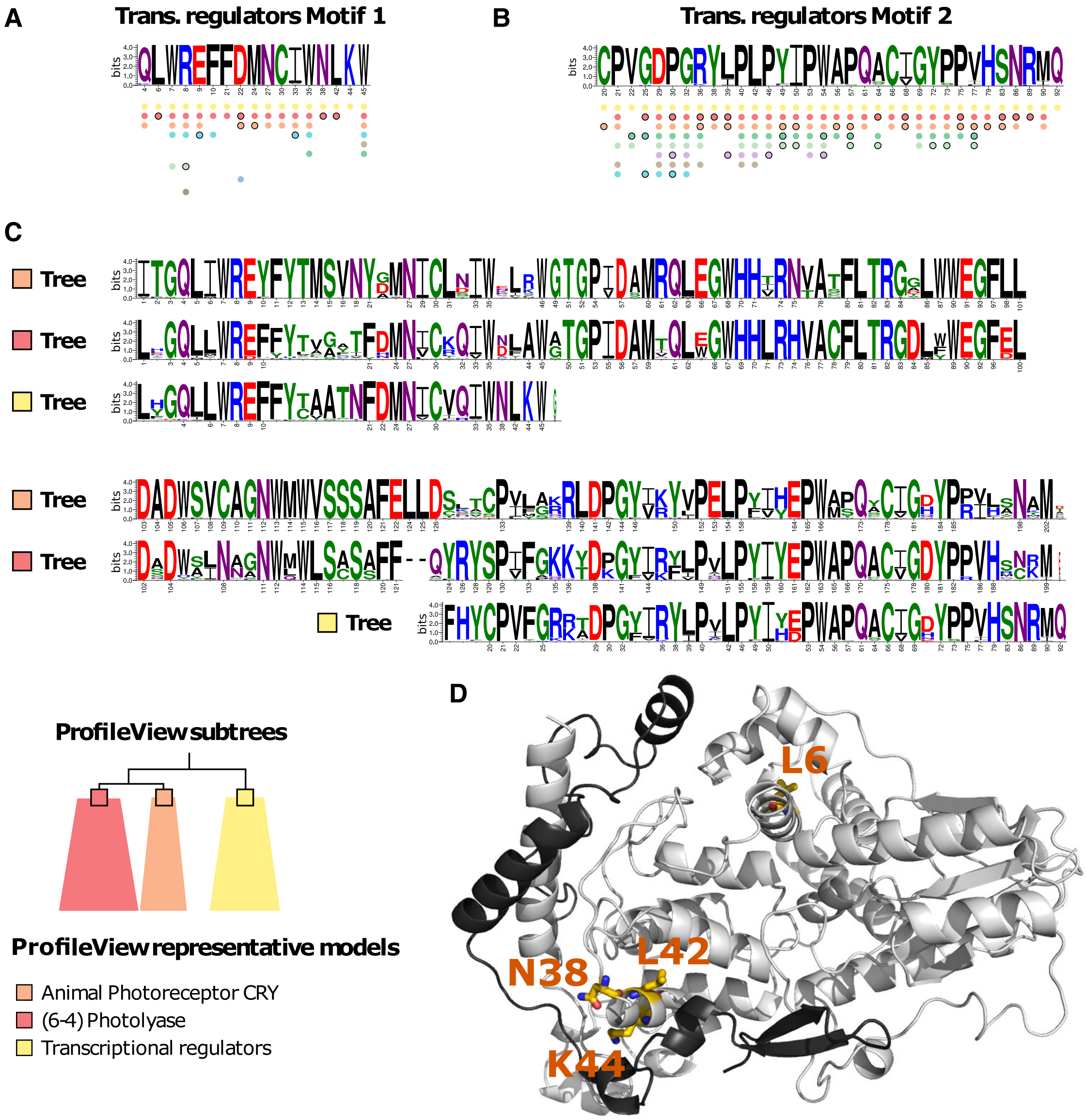

A
Trans. regulators Motif 1
QLWREFFDMNCIWNLKW
B
Trans. regulators Motif 2
CPVGDPGRYLPLPYIPWAPQACIGYPPVHSNRMQ
C
Tree
Tree
Tree
Tree
Tree
Tree
bits
ProfileView subtrees
ProfileView representative models
Animal Photoreceptor CRY
(6-4) Photolyase
Transcriptional regulators
D
L6
N38
L42
K44

FIG. 4. Transcriptional regulator motifs and their comparison with (6-4) photolyase and Animal photoreceptor CRY motifs. (*A*,*B*) two motifs of conserved residues present in light-independent transcriptional regulator sequences. They are extracted from two representative models (supplementary fig. S8, Supplementary Material online) of the "yellow" subtree of figure 3*A* and *E* (see bottom). Numbers (under the letters) correspond to positions in a model, and they are not comparable between motifs. A colored dot, piled below a motif, indicates that the corresponding position is well conserved (see Materials and Methods) in the representative model of the subtree of that color in figure 3*A*. Circled dots indicate positions that are less conserved (see Materials and Methods). For each motif, colored dots are ordered, from top to bottom, depending on best E-values given by hhblits to the pairwise model alignments. (*C*) Representative motifs associated with the transcriptional regulator (yellow), (6-4) photolyase (red) and Animal photoreceptor CRY (orange) subtrees of ProfileView tree (bottom) are aligned. Numbered positions correspond to conserved positions belonging to the associated representative motif. The absence of the number indicates less conserved positions. The alignment has been constructed using transcriptional regulator motifs as template models and all others as query models. The length of a motif depends on the length of the associated model, selected as best representing the sequences in a subtree. (*D*) PDB structure (4CT0) of the interacting mouse cryptochrome mCRY1 (grey) and Period2 mPER2 (black) involved in the circadian clock. Residues L6, N38, L42, and K44 are specific to light-independent transcriptional regulators (supplementary text 2, Supplementary Material online).

repressors (fig. 4C) despite the importance of the FAD chromophore in their regulation (Hirano et al. 2017).

To validate ProfileView motifs, we exploited the functional information derived by characterized mutations and looked whether their conserved amino acid positions would identify known functional natural variations, single amino acid residue replacements by site-directed mutagenesis or random mutagenesis, and structural specificity when structures were available. For this purpose, we manually curated a list of experimentally characterized positions in the CPF sequences (see supplementary file, Supplementary Material online "CPF_mutants_used_for_validation.xlsx"). Most of these positions display mutations causing loss of function or phenotypic changes. They are often involved in binding with other proteins, DNA substrates or with the cofactor FAD; active amino acids involved in catalytic or allosteric sites, such as DNA repair for photolyases or post-translational modifications in CRY, are also identified. Supplementary table S5, Supplementary Material online summarizes how many ProfileView positions are validated by current experimental evidence. Interestingly, ProfileView finds a number of highly specific positions for CPF functional classes that have not been reported in the literature before. We discussed these positions together with other observations in supplementary text 2, Supplementary Material online. They illustrate the great deal of functional information that can be extracted from representative motifs and be used to design tailored experiments for discovering new functional activities or novel biological mechanisms involving the FAD-binding domain.

How ProfileView Representative Motifs Distinguish Evolutionary Close Sequences?

ProfileView can distinguish very similar sequences associated with different functions. We illustrate this crucial feature with a concrete example, based on representation models and motifs. CPF sequences U5NDX3 and R7UL99 are grouped together by phylogenetic analysis because they are very similar (sequence identity is 61.8% and sequence similarity is 74.7%) and are classified in different functional groups by ProfileView, as a photolyase and a transcriptional regulator respectively. The conserved positions belonging to the photolyase functional motif (motif called “(6-4) photolyase” in supplementary fig. S6, Supplementary Material online) are shown in the alignment reported in supplementary figure S9, Supplementary Material online. For almost all positions in the motif the corresponding amino acid is conserved in both sequences (green) as expected by the high sequence identity of the alignment. For positions 1 (L), 33 (I), and 135 (K) in the motif, the amino acid is conserved only in the U5NDX3 sequence (the corresponding amino acids in R7UL99 are colored blue). Viceversa, positions 136 (K) and 160 (Y) in the motif are conserved in R7UL99 but not in U5NDX3. Even in the presence of this high sequence conservation, the (6-4) photolyase motif distinguishes the sequences by providing higher matching value for U5NDX3 than for R7UL99. Among the five positions, two of them, 1 and 135, are highly conserved in the photolyase family and variable in the transcriptional regulator family (see missing yellow dots below the (6-4) photolyase motif in supplementary fig. S6, Supplementary Material online) making the U5NDX3 sequence closer in classification space to the photolyase subgroup than R7UL99. Note that these observations concern the dimension of ProfileView classification space which is associated with the “(6-4) photolyase” model. Ultimately, it is the contribution of all profile models, one for each dimension of the space, that will define the position of the sequences bringing them closer either to the photolyase subgroup or the transcriptional regulator subgroup.

A second example is reported in supplementary figure S10, Supplementary Material online for sequences Q6MDF3, D8UF46, and Q485Z2. The position of the sequences in the CPF phylogenetic trees (supplementary figs. S3 and S4, Supplementary Material online) could wrongly suggest an ancestral function, conserved in paraphyletic groups separated by clades where neofunctionalization would occur. In contrast, a sequence alignment analysis (see legend in supplementary fig. S10, Supplementary Material online) driven by representative motifs highlights specific positions that explain the functional classification of the three sequences.

ProfileView on the GH30 Family of the CAZy Database

The glycoside hydrosylases (EC 3.2.1.-), in short GH, are a widespread group of enzymes which hydrolyse the glycosidic bond between two or more carbohydrates or between a carbohydrate and a noncarbohydrate moiety. Their classification, based on substrate specificity and occasionally on molecular mechanisms, has proven to be particularly difficult. For this purpose, a vast knowledge on these enzymes has been meticulously curated in the CAZy database (Lombard et al. 2014). The GH30 is one of the GH families that has been organized in subfamilies in CAZy (http://www.cazy.org/GH30.html). It counts nine different subfamilies (GH30-1,..., GH30-9) corresponding to 11 different enzymatic chemical reactions. Some of these subfamilies are functionally classified in CAZy and some others are left unclassified.

Validation of ProfileView on the Functional Diversity of GH30 Sequences

We considered a set of 1675 GH30 sequences and their 695 functionally classified sequences in CAZy (table 2). ProfileView representation space and ProfileView tree for these sequences have been constructed using models coming from two similar PFAM domains: PF02055 (Glyco_hydro_30) and PF14587 (Glyco_hydr_30_2).

ProfileView classifies 584 out of 695 sequences well, within eight subtrees (T_1, T_2, T_3, T_6, T_7, T_{10}, T_{11}, T_{12} in fig. 5) and leaves 106 sequences unclassified and five misplaced (supplementary tables S6 and S7, Supplementary Material online, and fig. 5). Noticeably, all unclassified sequences in CAZy subfamilies GH30-3, GH30-4, and GH30-5 are grouped by ProfileView into three subtrees missing a representative model (T_4, T_5, T_8, respectively). All misplaced sequences in CAZy subfamily GH30-6 are grouped into the same subtree missing a representative model (T_9). Furthermore, the same subtrees separate well the EC numbers in CAZy functional annotation (supplementary table S7, Supplementary Material online).

In figure 5, the existence of multiple representative models for the internal nodes of the ProfileView classification tree highlights a possible functional subcharacterization for several CAZy subfamilies. For instance, note that the two CAZy reactions 3.2.1.45 and 3.2.1.21 +3.2.1.37 for GH30-1 are identified in distinguished subtrees (green and violet labels are associated with reactions

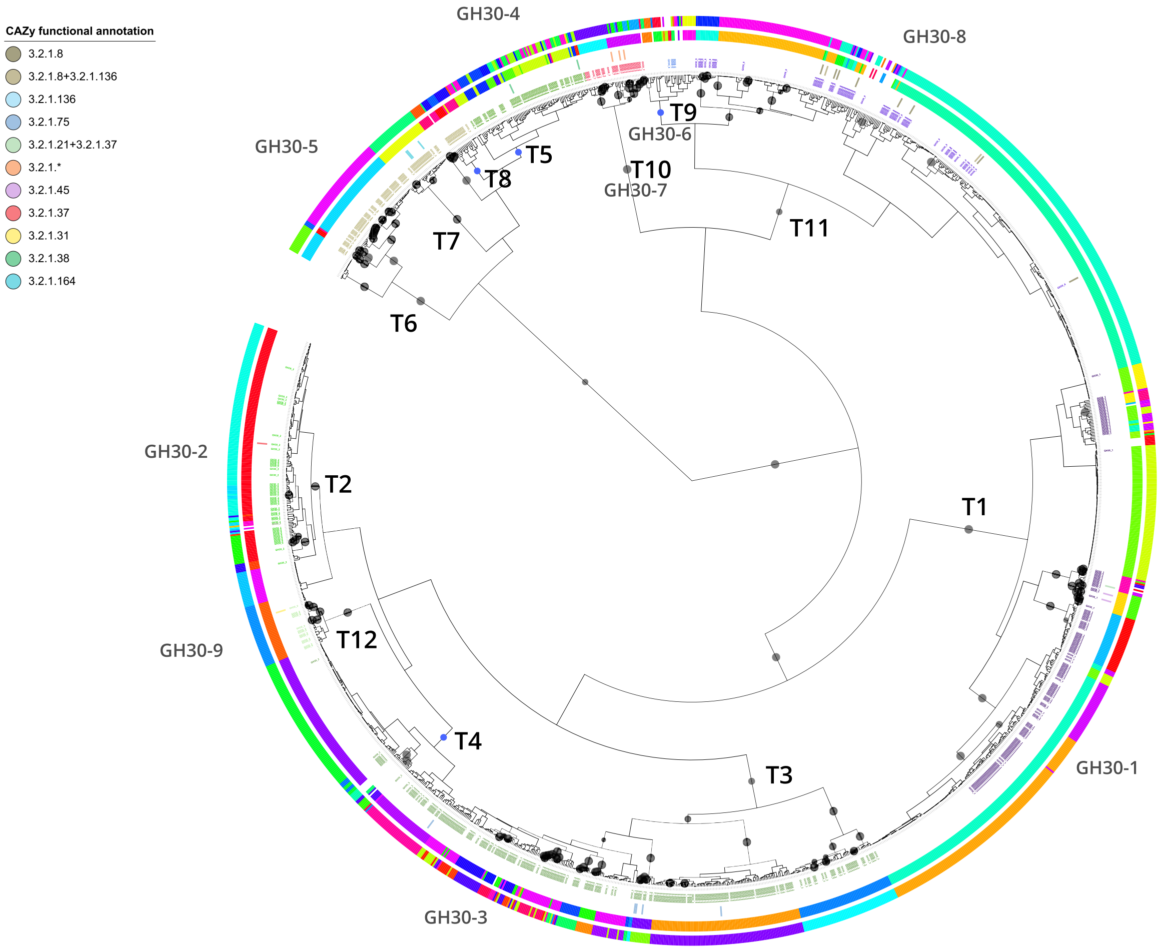


Fig. 5. ProfileView tree of GH30 sequences. The tree is based on the construction of models for the two pfam domains PF02055 (Glyco_hydro_30) and PF14587 (Glyco_hydr_30_2). Black dots in the tree indicate the existence of representative models separating at least 75% of the sequences in the subtree (note that lowering the threshold to 50% provides comparable results). The first external ring contains the labels of CAZy subfamilies (GH30-1, …, GH30-9), also indicated in larger characters on the annotated tree for an easier reading. Sequences and their classification correspond to those used in figure 3 of Barrett and Lange (2019). The second ring reports the existence of an "EC number" providing the functional annotation in CAZy. The EC numbers and their associated colors are indicated on the top left (GH30-1: 3.2.1.45 and 3.2.1.21+3.2.1.37; GH30-2: 3.2.1.37; GH30-3: 3.2.1.75; GH30-4: 3.2.1.38; GH30-5: 3.2.1.164; GH30-6: –; GH30-7: 3.2.1.*; GH30-8: 3.2.1.8, 3.2.1.136, 3.2.1.8+3.2.1.136; GH30-9: 3.2.1.31). The third ring reports CUPP clustering (Barrett and Lange 2019). Different colors are used to indicate different CUPP clusters. See supplementary table S6, Supplementary Material online. The fourth and most external ring reports eCAMI clustering (Xu et al. 2020). Different colors are used to indicate different eCAMI clusters.

3.2.1.45 and 3.2.1.21+3.2.1.37 in fig. 5) separated by a representative model. Furthermore, for the GH30-3 subfamily, several sequences labeled by CAZy reaction 3.2.1.75 occur in different subtrees endowed with representative models, highlighting potential functional differences within this subfamily.

ProfileView on the Enzyme Superfamilies of the Structure-Function Linkage Database

The Structure–Function Linkage Database (SFLD) is a manually curated classification resource describing structure–function relationships for functionally diverse enzyme superfamilies (Schnoes et al. 2009; Akiva et al. 2014). Despite their different functions, the members of these superfamilies "look-alike," which facilitates annotation errors. We challenge ProfileView against these sets of sequences and show that its classification meets the functional information in SFLD.

SFLD is organized in superfamilies whose members are subdivided into subgroups using sequence information, and finally into families, that is, sets of enzymes known to catalyze the same reaction using the same mechanistic strategy. Subgroups are not organized by function, and the functional specificity of the sequences is detailed at the family level. We consider two different superfamilies, Haloacid Dehydrogenase and Radical SAM, because of their wide variety of functions. Indeed, the Haloacid Dehydrogenase family is characterized by 25 subgroups organized in 22 families and 20 different reactions, and the Radical SAM family by 58 subgroups organized in 98 families and 85 reactions (see sfld.rbvi.ucsf.edu/archive/django/superfamily/index.html for a detailed description). We analyzed the HAD/β-PGM/Phosphatase-like subgroup of Haloacid Dehydrogenase and three subgroups of Radical SAM: B12-binding domain containing, Methylthiotransferase and SPASM/twitch domain containing (see supplementary text 1, Supplementary Material online). ProfileView functional classification has been validated on the SFLD families associated with the four subgroups.

ProfileView on the HAD/β-PGM/Phosphatase-like Subgroup

The 259 characterized functions included in this subgroup comprise 2-haloacid dehalogenase, beta-phosphoglucomutase, phosphonoacetaldehyde hydrolase, and phosphatases of various specificities (see sfld.rbvi.ucsf.edu/archive/django/subgroup/1129/index.html). We run ProfileView on a model library constructed from the two similar Pfam domains HAD and HAD_2 (see tables 1 and 2). ProfileView classifies well all-known sequences belonging to known characterized functions in distinguished subtrees endowed with representative models. Neither unclassified nor misplaced sequences were identified, as illustrated in figure 6 and supplementary table S8, Supplementary Material online.

Comparison of ProfileView with Other Computational Approaches

ProfileView is compared with the PANTHER classification system (Mi et al. 2012, 2013) and the two k-mer-based platforms CUPP (Barrett and Lange 2019) and eCAMI (Xu et al. 2020). One more comparison with CUPP and the state-of-the-art neural network approach based on Restricted Boltzman Machines (RBM) described in (Tubiana et al. 2019) are reported in supplementary text 1, Supplementary Material online (on the WW domain family). In all comparisons, ProfileView outperforms or is on par with the functional classification considered.

ProfileView and PANTHER

PANTHER (Mi et al. 2012, 2013) is a large curated biological database of gene/protein families and their functionally related subfamilies which has been designed to classify and identify the function of gene products. PANTHER provides data and tools to group sequences in functional clusters. Unlike ProfileView, it does not organize them in a distance tree, missing the possibility to identify large-scale functional properties for groups of sequences that cluster together, such as light-dependent/independent CPF sequences.

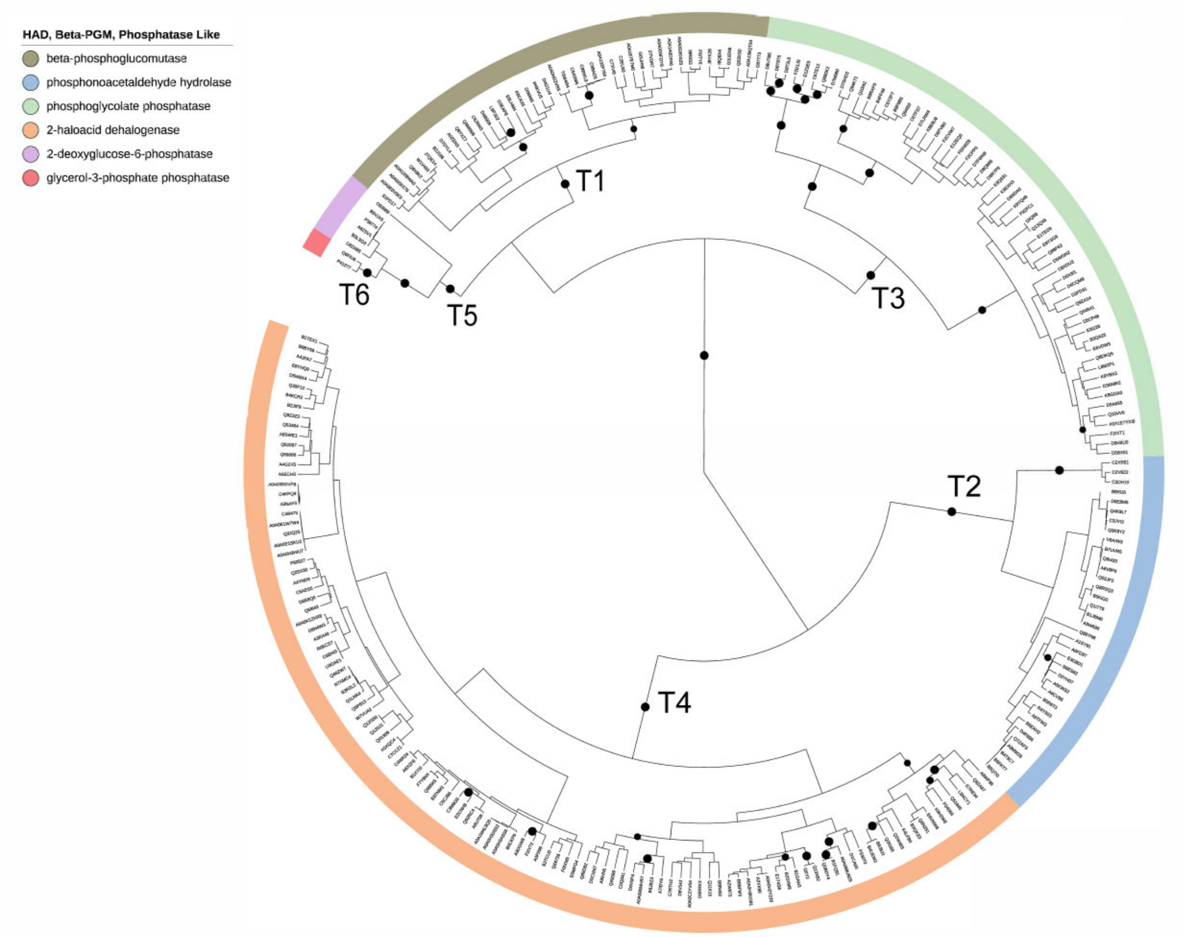

HAD, Beta-PGM, Phosphatase Like
● beta-phosphoglucomutase
● phosphonoacetaldehyde hydrolase
● phosphoglycolate phosphatase
● 2-haloacid dehalogenase
● 2-deoxyglucose-6-phosphatase
● glycerol-3-phosphate phosphatase
T1
T2
T3
T4
T5
T6

FIG. 6. ProfileView classification tree of the HAD/β-PGM/Phosphatase-like subgroup of Haloacid Dehydrogenase in SFLD. Validation test of ProfileView performance. See supplementary table S8, Supplementary Material online.

We annotated the 307 sequences belonging to the CPF family with PANTHER and compared PANTHER to ProfileView on the 72 functionally characterized CPF sequences. For easier visualization, we reported PANTHER classification of the full set of CPF sequences on both the ProfileView classification tree and the CPF distance tree in supplementary figures S11 and S12, Supplementary Material online. Supplementary table S9, Supplementary Material online reports PANTHER classification of the 72 functionally characterized sequences. As ProfileView, PANTHER associates sequences in Class II CPD photolyase and CRY Pro to specific functional classes. In contrast, it associates many functional classes with the Plant Photoreceptor CRY. This implies, on the one hand, that PANTHER does not sharply identify the subset of Plant Photoreceptor CRY sequences and, on the other hand, that it suggests a finer functional delineation within this subset. In this regard, we notice that PANTHER associations in the Plant Photoreceptor CRY subtree of supplementary figure S11, Supplementary Material online correspond to the topology of the ProfileView Plant Photoreceptor CRY subtree. Moreover, several PANTHER classes (e.g., cryptochrome-1, (6-4) photolyase isoform A, and SI:CH1073-390K14.1) are associated with distinct CPF classes. Some associations are clearly faulty as it is the case for the (6-4) PL sequences annotated as circadian regulators, and for the Class I CPD photolyase sequences classified as (6-4) PL. Note that PANTHER class SI: CH1073-390K14.1 recognizes both Plant photoreceptor Cry and Class III CPD photolyase, and that sequences in the NCRY subtree are annotated as (6-4) photolyase while they are PRs according to us and to (Emmerich et al. 2020).

ProfileView and the Two k-mer-based Platforms CUPP and eCAMI

CUPP (Barrett and Lange 2019) and eCAMI (Xu et al. 2020) are two computational approaches designed to classify carbohydrate-active enzymes by using short peptide/ k-mer sequences expected to be enzyme specific. In CUPP and eCAMI, proteins sharing the same peptide profile are claimed to share the same function.

The set of GH30 sequences used to validate ProfileView (fig. 5) was also used for the evaluation of CUPP (Barrett and Lange 2019). CUPP splits these sequences in 33 groups and organizes them in a dendogram (Barrett and Lange 2019) whose topology is reported in supplementary figure S13A, Supplementary Material, online. The dendogram is composed of nine subtrees corresponding to the nine CAZy subfamilies. A schematic comparison of CUPP dendrogram (Barrett and Lange 2019) and ProfileView tree is given in supplementary figure S13B, Supplementary Material, online. Both their topologies highlight the separation of the CAZy subfamilies GH30-1, GH30-2, GH30-3, and GH30-9 from the other subfamilies. ProfileView tree separates further subfamilies GH30-4 and GH30-5 from the remaining ones.

A detailed analysis of the CAZy subfamilies indicates similar sequence organization for both ProfileView and CUPP. For instance, CUPP organizes GH30-1 sequences by splitting them in five clusters (Barrett and Lange 2019) that are easily identified in ProfileView tree, where three representative models are associated with three of CUPP clusters (purple, fuchsia, and dark blue in third circle of annotation in fig. 5). In contrast, the classification of CAZy subfamilies GH30-4 and GH30-5 (fig. 5) highlights a large number of CUPP clusters while ProfileView groups GH30-5 into three main subtrees and GH30-4 into one. Two of the three ProfileView subtrees grouping GH30-5 are characterized by representative models. Interestingly, the remaining sequences are clusterized by CUPP into several clusters and no representative model is found by ProfileView, indicating the difficulty of both methods in classifying this group of sequences.

On the GH30 family, eCAMI performs very similarly to CUPP (compare the two most external layers of fig. 5). eCAMI tends to group sequences in a larger number of clusters than CUPP. This cluster fragmentation corresponds to small ProfileView subtrees and affects the same sets of sequences that are of difficult classification for CUPP.

To test the general applicability of ProfileView versus CUPP and eCAMI, both designed to classify carbohydrate-active enzyme sequences, we also compared the three approaches on the CPF sequences. CUPP and eCAMI were run using both FAD and PHR sequences. CUPP tree and its associated clusters are represented in supplementary figure S14, Supplementary Material online for FAD sequences (see also supplementary fig. S15, Supplementary Material online). CUPP: (1) groups all together the CPF classes “Transcriptional regulators,” (6-4) photolyase and Animal photoreceptor CRY. Hence, distinguished functions are shared in the same subtree. In particular, it does not distinguish light dependent from light independent protein sequences; (2) does not distinguish Classes I and III CPD PL; (3) places the CRYPro subtree far from the remaining subtrees while, in ProfileView, CRYPro is located closer to Class II CPD photolyase; (4) splits the CRY DASH tree into two distinguished subtrees, one of which contains no sequence with a known functional annotation.

In addition, CUPP successfully classifies a larger number of sequences (corresponding to the leaves left uncolored in supplementary fig. S14, Supplementary Material online) in the CPF family compared to ProfileView, which did not find sufficient confidence among its models to include some input sequences in its tree (Steps II and III of ProfileView pipeline in Materials and Methods). Viceversa, there are sequences that have been classified by ProfileView and that do not belong to CUPP classification (see uncolored sequences within CUPP clusters in supplementary fig. S14, Supplementary Material online). We also notice that, like ProfileView, CUPP: (1) groups Class II CPD photolyase in a single subtree, and (2) distinguishes NCRY sequences.

When CUPP considers the whole PHR sequence, the topology of the CUPP tree (supplementary fig. S16B, Supplementary Material online) gets closer to ProfileView topology even though CUPP mixes up Classes I and III CPD photolyase as well as light-dependent (6-4) photolyase and Animal photoreceptor CRY sequences; the NCRY subtree locates close to photolyases (supplementary fig. S15, Supplementary Material online); the higher number of CUPP clusters fragments the functional organization, as for instance for Class II CPD PL.

CUPP and eCAMI clustering can be visualized on the ProfileView tree in supplementary figure S17, Supplementary Material online. As observed for the GH30 family, eCAMI tends to group sequences in a larger number of clusters than CUPP. For CPF sequences, it separates some Animal photoreceptor CRY sequences from transcriptional regulators and (6-4) photolyases, clustered together by CUPP. It does not separate transcriptional regulators and (6-4) photolyases though. Furthermore, in agreement with ProfileView, it distinguishes between Classes I and II CPD photolyase sequences based on the FAD domain. In contrast, eCAMI divides the set of characterized sequences of Class I CPD photolyase on FAD and PHR sequences into several clusters (supplementary fig. S17, Supplementary Material online).

Overall, this analysis highlights CUPP's and eCAMI limitations in handling arbitrary protein families.

Discussion

The availability of large amounts of (meta)genomic data allows for a deeper exploration of living organisms and the processes underpinning their genetic, phylogenetic, and functional diversification. Computational approaches, capable of highlighting these diversities and identifying what is functionally novel in sequence information, will make the first fundamental step in the discovery of new candidates whose functional activity will be tested experimentally. Moreover, due to the huge amount of sequences that will be acquired in coming years (1 zetta-bases/year are expected in 2025 Stephens et al. 2015), there will no longer be a way to examine this mass of data with an "expert eye" and computational approaches will play a key role in extracting new information and in functional classification.

Today, we can characterize homologs on the basis of their similarity using distance measures that model the evolution of the entire sequences. However, as shown here and elsewhere (Schnoes et al. 2009; Mi et al. 2012; Akiva et al. 2014; Barrett and Lange 2019), this computational approach is insufficient to provide insights on the functional activities of proteins, and a large number of sequences are still not functionally annotated. Some of these protein families, like the seven families discussed in this study, are extremely important in medicine, biology, environmental science and biotechnology due to their key roles in cancer biology, DNA repair, drug delivery strategies, chronobiology, and photobiology, specific enzymatic reactions, the formation of protein–protein interaction networks, optogenetics. Thanks to their key role, over the decades, experiments have accumulated an enormous amount of functional information that we have used to validate the ProfileView approach. ProfileView functional organization of the seven families considered here agrees with experimental evidence.

ProfileView highlights that the functional classification of proteins depends on the nonlinear contribution of many profile models and that conserved patterns in sequences are not sufficient alone to discriminate diversified functions of complex protein families. This change of perspective in functional classification underlies the complexity of the question and explains why this problem remains wide open today despite the clear interest in classifying protein families that have been extensively studied in molecular biology, such as transporters, signaling, and transcription factors. Not least, the recent focus on *de novo* genes points out that the notion of "function" is more complicate than one might expect (Keeling et al. 2019).

By constructing multiple profile models characterizing different conserved motifs in homologous domain sequences, ProfileView captures functional signals and, by combining them, is able to successfully classify large datasets.

Its main advantages compared to approaches developed before are as follows: (i) ProfileView is alignment-free and avoids errors due to the difficulty of comparing distant homologs; (ii) several profile models represent more precisely than a single consensus model the functional variability of protein families; (iii) large amounts of data are not required to learn features and perform classification because of a relatively small number of profile models, which is reduced to a few thousand, and a construction of profile models from a few dozen sequences; (iv) functional annotation of many sequences does not need to be known to explore with precision the space of sequences and classify them; (v) ProfileView is a general approach applicable to proteins of arbitrary length and function. Moreover, once a domain library is constructed, ProfileView is computationally efficient in screening very large sets of homologous sequences in a reasonable time.

ProfileView demonstrated to discover potentially interesting CPF proteins whose function could be tested experimentally to identify new light-responsive proteins with novel features. Photoactive proteins are of interest for biotechnology and any computational approach to finding them is desired. More broadly, ProfileView has the potential to greatly expand our understanding of the mechanisms developed by nature to exploit light for functional purposes. ProfileView also organized the WW domain family in subtrees of sequences, corresponding to a large spectrum of differences in binding affinity to various ligands, which have been experimentally observed. It demonstrated that a large variety of sequence motifs covers this spectrum and it identified these motifs. It could classify protein superfamilies in the manually curated CAZy and SFLD databases by accurately identifying differences in their multiple enzymatic reactions. Compared to Tubiana et al. (2019), a computational approach also based on sequence analysis, ProfileView described differences among binding motifs in much greater detail, opening new avenues in the discovery of alternative binding patterns in protein–protein interaction networks. It has been compared favorably to other classification tools like PANTHER, CUPP, and eCAMI, on the CPF, the WW domains and the GH30 family classified in the Carbohydrate-Active Enzymes database CAZy.

The ProfileView method makes no assumptions about the complexity of the domain architecture of a protein family. For the applications discussed here, ProfileView operates under the assumption that the analysis of a single domain is sufficient to functionally classify very different protein families, possibly with complex domain architecture. For CPF sequences, we observed that ProfileView classification tree is in agreement with the known domain architectures for CPF subfamilies and potentially on their evolution. This result highlights an "extra" structure on the CPF family and points to a much more general question about the precise relationships between the evolution of domain architectures and the evolution of functions for a protein family. In principle, we cannot exclude that domain multiplicity and/or domain order might play a role for functional classification of some protein families (Basu et al. 2009; Lees et al. 2016). To test this hypothesis, ProfileView could be used to systematically classify known protein families based on single domains. Such classification will clarify, on the one hand, the large-scale applicability of ProfileView and, on the other hand, will contribute to our understanding of the (combined) evolution of functions and domain architectures.

On the methodological side, ProfileView addresses the problem of extracting biological information about protein families from the huge space of natural sequences. Sampling of distant sequences could be realized using different distance measures. This is an important direction of investigation that could lead to finer biological information extracted from sequences.

From the algorithmic point of view, ProfileView is surprisingly simple compared to the Restricted Boltzmann Machines (RBM) model used in Tubiana et al. (2019) to classify WW domain homologs. RBM are generative stochastic (single layer) artificial neural networks that can learn a probability distribution over a set of inputs. Once the machine is trained on protein sequences, it can be used to either generate new protein sequences that look "alike" the ones that have been used in the training or to estimate the probability for a sequence to be generated by the model. RBM learn collective modes by extracting short sequence motifs from sets of sequences based on correlation patterns among alignment positions. These motifs might reveal structural, functional and phylogenetic features and they are used to define a representation space where to classify sequences. RBM generative nature makes training challenging by an algorithmic point of view since intensive sampling from large training sets is required. In contrast, ProfileView constructs profile models seeded from distant homologous sequences. To construct a model, ProfileView requires a very small number of natural sequences, 20 or more, that are similar to the seed sequence. Also, ProfileView makes no use of positional correlations nor generates artificial sequences. Its profile models encode conserved patterns ignoring those parts of the homologous sequences appearing variable (see discussion on the two CPF sequences U5NDX3 and R7UL99 above). The number of models is not a restriction for the construction of the classification space.

The intrinsic simplicity of ProfileView makes it possible to envision new directions of investigation such as the design of a ProfileView extension that could consider motifs across multiple domains, for proteins of complex domain architectures. Indeed, the fine understanding of functional mechanisms might need more sophisticated computational approaches than ProfileView. For instance, for the CPF classification based on the FAD-binding domain, ProfileView highlights functional differences between large classes of CPF sequences, helping to model the proximity between these classes with an appropriate identification of a functional tree topology. To find functional differences within classes and to anticipate the existence of a double function (see supplementary fig. S2, Supplementary Material online), the interplay between domains in the CPF sequence might have to be considered as highlighted in Rosensweig et al. (2018).

ProfileView is deeply rooted in the evolutionary information encoded in genomic sequences. For this reason, it is expected to contribute to fundamental questions of genome evolution, such as accurate reconstruction of gene duplication history. A fundamental question in this regard concerns the functional distinction of paralogous genes within a phylogenetic tree and within a single species. The power of multiple profile models in identifying different functional determinants between homologs should be able to do the same between paralogs.

Last, even though ProfileView has been applied here to the classification of entire protein sequences, it can handle metagenomic sequences as well. In this regard, it is important to highlight that the majority of metagenomic and metatranscriptomic data come from organisms that cannot be cultured and may never be isolated. Therefore, new conceptual approaches to explore their biology in complex ecosystems are desperately needed. ProfileView increases knowledge on the biology of organisms whose ecological role has been recognized (e.g., marine microbes) but which are still not accessible to functional investigations, thus opening a new avenue for functional exploration.

Materials and Methods

Datasets Used to Validate the Method

Datasets of Sequences to be Classified

The seven protein families used to evaluate ProfileView performance are listed in table 1 (first column). Their sets of homologous sequences to be classified (see table 1, second column, for their number) were retrieved from publicly available databases (see below). For each family, a subset of sequences was functionally characterized (see table 1, fourth column) and we used it for evaluation. All families show multiple functions (table 2, second column). The different characteristics of the protein families are reported in table 1 and supplementary table S1, Supplementary Material online.

CPF sequences were retrieved from UniProt, JGI projects (genome.jgi.doe.gov), and OIST projects (marinegenomics.oist.jp). The set was constructed according to two main criteria and contains: (1) CPF sequences known to have a specific function according to experimental evidence reported in the literature (see supplementary file, Supplementary Material online for bibliographical references); (2) CPF sequences that span the entire tree of life; they belong to 146 species, 74 classes, and 40 phyla (see supplementary file, Supplementary Material online for the detailed list). In the text, a "CPF sequence" refers to the full-length CPF sequence comprising the PHR domain, including the FAD-binding domain, and possibly the C- and N-terminal extensions, whereas a "FAD sequence" refers to the FAD-binding domain sequence exclusively.

The set of WW domain sequences was constructed by combining the datasets of natural sequences analyzed in (Otte et al. 2003; Ingham et al. 2005; Russ et al. 2005; Tubiana et al. 2019). Sixty sequences were experimentally characterized (Otte et al. 2003; Ingham et al. 2005; Russ et al. 2005), and the remaining ones were randomly selected in comparable proportion from the three sets classified in (Tubiana et al. 2019) as types I, II/III, and IV.

The set of GH30 sequences is the same as that used in (Barrett and Lange 2019) (file GH30.faa provided with the CUPP program v1.0.14 and containing 1803 sequences) and described in the Carbohydrate-Active Enzymes database CAZy (http://www.cazy.org/GH30.html). It is organized into several subfamilies of the CAZy classification. Some of these subfamilies are functionally classified by CAZy and others are left unclassified. We used the annotation files in Barrett and Lange (2019), where 721 of the 1803 sequences have mapping/labeling to subfamilies GH30-1 through GH30-9. Note that, of the 1675 sequences retained for ProfileView analysis after filtering, 695 have a label in the GH30 ProfileView tree (table 1).

The set of sequences of the HAD/β-PGM/Phosphatase-like subgroup of the Haloacid Dehalogenase (HAD) superfamily and the three subgroups of the Radical-SAM superfamily (B12-binding domain containing, Methylthiotransferase and SPASM/twitch domain containing) were retrieved from the Structure-Function Linkage Database (SFLD) (Schnoes et al. 2009; Akiva et al. 2014). More precisely, each subgroup is defined by the union of the sets of annotated sequences associated with its families in SFLD. Given a subgroup, we considered all of its families, even if they were represented by very few sequences, possibly only one.

Datasets of Sequences Used for Model Library Construction

For all protein families, the domain(s) considered for model construction, their Pfam accession code, and the number of models constructed are shown in table 1. The seed sequences seeding the models were retrieved from the Pfam *full* set associated with the Pfam domain used for classification. The seed sequences for FAD were retrieved from Pfam v31 while for all other domains we used Pfam v32. For the three families, GH30, HAD/β-PGM/Phosphatase-like and B12-binding domain containing, Pfam contains two similar domains (see table 1, fifth column) and we used the union of the Pfam sequences from both these domains.

For each seed sequence, the homologous sequences used to build the profile model were retrieved from Uniclust30, which is UniProtKB clustered to 30% identity and for which a HHblits database is provided (Mirdita et al. 2017).

Clade-Centered Models and a Multi-source Functional Annotation

Widely used methods searching for homologous domain sequences (Altschul et al. 1997; Eddy 2011; Remmert et al. 2011) rely on a single-source annotation strategy, where a single profile model (e.g., a pHMM Eddy 1998), generated by the consensus of a set of homologous sequences, is used to represent a protein domain. The single-source strategy usually works well for rather conserved homologous sequences, but when the sequences are highly divergent, the consensus signals become too weak to generate a useful probabilistic representation and the global consensus models do not properly characterize domain features. A *multi-source* domain annotation strategy (Bernardes et al. 2016), in which protein domains are represented by several profile models, called *Clade-Centered Models* (CCM), was implemented in CLADE (Bernardes et al. 2016) and MetaCLADE (Ugarte et al. 2018) for genomes and metagenomes/metatranscriptomes, respectively. There, we showed that CCMs significantly improve domain annotation for both complete genomes and metagenomic/metatranscriptomic sequences. Because of their proximity to protein sequences, CCMs are more specific and functionally predictive than canonical global consensus models.

Here, we construct and use CCMs differently, with the goal of better resolving the functional organization of sequences within protein families. In order to capture conserved motifs that might be of functional relevance to the family, we build CCMs that are highly specific. The motifs, which consist of conserved positions on subsets of close homologs, will likely belong to protein interaction sites and will be determinants of functional specificity. To construct CCMs (see below, step I of the ProfileView pipeline), we consider the *FULL* set of sequences S^i associated with a Pfam domain D^i (Finn et al. 2014) and, for each sequence $s_j \in S^i$, we construct a profile HMM by retrieving a set of homologous sequences close to s_j from

UniProt. Such a model displays features that are characteristic of s_j and that might differ from other sequences $s_k \in S^i$. The rationale is that the more divergent s_j and s_k are, the more we expect CCMs to highlight different features within a protein family.

The ProfileView Method

The ten main steps of the ProfileView pipeline are explained in detail below. A flowchart is provided in figure 7. A hands-on description of the ten steps for the CPF family is given in supplementary text 2, Supplementary Material online.

ProfileView takes as input a Pfam domain *D* and a set of homologous sequences $\mathcal{S}$ to be classified. If there are similar Pfam domains (Pfam usually names them with a numerical extension, as for instance HAD and HAD_2), the user may decide to provide several alternative domains as input and build the model library $\mathcal{M}_\mathrm{D}$ from multiple domains accordingly. The output of ProfileView is a classification tree with representative models associated with internal nodes of the tree, when they exist, and functional motifs characterizing the sequences in the corresponding subtree.

I. Model Library Construction (fig. 1A). To construct a library of models $\mathcal{M}_\mathrm{D}$ for the domain *D*, we considered

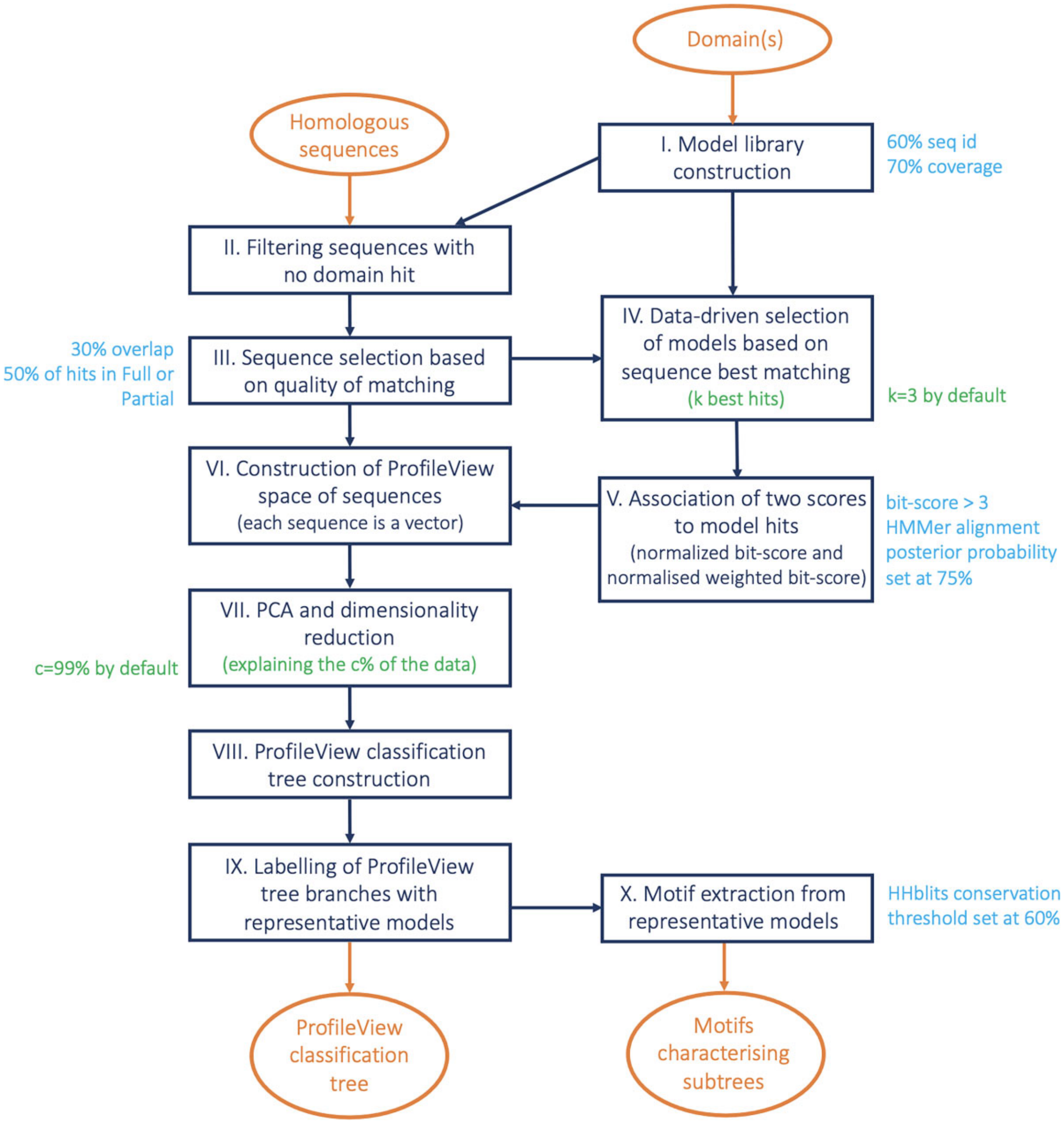

Domain(s)
Homologous sequences
I. Model library construction
60% seq id
70% coverage
II. Filtering sequences with no domain hit
30% overlap
50% of hits in Full or Partial
III. Sequence selection based on quality of matching
IV. Data-driven selection of models based on sequence best matching
(k best hits)
k=3 by default
VI. Construction of ProfileView space of sequences
(each sequence is a vector)
V. Association of two scores to model hits
(normalized bit-score and normalised weighted bit-score)
bit-score > 3
HMMer alignment posterior probability set at 75%
VII. PCA and dimensionality reduction
(explaining the c% of the data)
c=99% by default
VIII. ProfileView classification tree construction
IX. Labelling of ProfileView tree branches with representative models
X. Motif extraction from representative models
HHblits conservation threshold set at 60%
ProfileView classification tree
Motifs characterising subtrees

Fig. 7. ProfileView flowchart. The ProfileView pipeline is organized in ten main steps: (I) construction of the model library for a domain or a few similar domains chosen by the user, (II) sequence filtering based on matching/unmatching of the models on a sequence, (III) sequence selection based on the quality of a match, (IV) filtering of models to reduce model redundancy, (V) association of two scores to each model hit, (VI) construction of ProfileView space of sequences, (VII) dimensionality reduction of ProfileView sequence space, (VIII) construction of ProfileView classification tree, (IX) identification of the best representative models for subtrees, (X) extraction of functional motifs from representative models. User-editable parameters are highlighted in green and those that remain fixed are highlighted in cyan.

sequences from the FULL dataset in Pfam database (Finn et al. 2014) as "seeds" for the models. For each sequence, we built a CCM (Bernardes et al. 2016) by searching in Uniclust30 for highly significant matches of homologous sequences having at least 60% identity with the seed sequence and covering at least 70% of it. More precisely, a multiple sequence alignment is built using the command `hhblits` of the HH-suite (Remmert et al. 2011) (with parameters `-qid 60 -cov 70 -id 98 -e 1e-10` and database `uniclust30_2017_10`) and subsequently converted into a pHMM with HMMER (Eddy 1998) in order to perform a sequence-profile comparison. Typically, profile models are built from 50 to 100 sequences; a minimum of 20 sequences (heuristically determined threshold) is required.

Note that the FULL set of Pfam sequences associated with a domain might be very large and contain tens of thousands of sequences. If so, we sample a few thousand sequences, by first clustering the set with MMseq2 (https://github.com/soedinglab/MMseq2). The `easy-cluster` command of mmseqs is used with two parameters: `--min-seq-id` sets the minimum sequence identity for clustering, and `-c 0.8` considers matches above this fraction of aligned/covered query/target residues. MMseq2 is used to cluster close sequences together and for this, we required that sequences in a cluster had more than either 40 or 60% sequence identity (default set at 50%) depending on the protein family, so that several thousands (up to 6000) representative sequences could be identified from different clusters. The selected representative sequences were used as seeds to build profile models, as indicated above.

If several similar Pfam domains are considered, as for the GH30 and HAD families, the above procedure is applied to the union of Pfam sequences associated with all domains. (Ultimately, we observed that several similar domains slightly improve classification.)

II. Sequence Filtering. After constructing the set of models $\mathcal{M}_D$ for the domain *D*, we discarded from the input set of sequences $\mathcal{S}$ all sequences against which HMMER (version 3.1b2) found no domain hit, regardless of the hit score. For all protein families, table 1 (third column) reports the number of sequences after filtering. Note that this filtering step, based on multiple profile models, can identify domains in divergent sequences where the consensus Pfam model cannot provide a hit. See supplementary figure S18A, Supplementary Material online for an illustration of sequence filtering.

III. Sequence Selection. Each CCM in $\mathcal{M}_D$ is mapped against the set $\mathcal{S}$ of all input sequences using HMMER. Let $\mathcal{H} = \{h_{s,m} \mid s \in \mathcal{S},\ m \in \mathcal{M}_D,\ \text{score}(h_{s,m}) > 0\}$ be the set of hits $h_{s,m}$ provided by `hmmsearch`, where *s* is a sequence of $\mathcal{S}$, *m* is a model of $\mathcal{M}_D$ and score($h_{s,m}$) is the bit-score assigned to $h_{s,m}$. The bit-score is a log-odds ratio score (in base two) comparing the likelihood of the pHMM to the likelihood of a null hypothesis (i.e., an i.i.d. random sequence model). More formally,

$$\text{score}(h_{s,m}) = \log_2 \frac{\Pr(s \mid m)}{\Pr(s \mid \text{null})}$$

where $\Pr(s \mid m)$ is the probability of the pHMM *m* generating the sequence *s* and $\Pr(s \mid \text{null})$ is the probability of *s* being generated by the null model (Barrett et al. 1997).

We partitioned the hit set $\mathcal{H}$ in three subsets Full($\mathcal{H}$), Overlap($\mathcal{H}$), Partial($\mathcal{H}$), where Full($\mathcal{H}$) contains all hits that fully cover the associated model, Overlap($\mathcal{H}$) contains all hits involving the extremes of a sequence covered only partially by the associated model (this situation corresponds to an "incomplete" sequence), and Partial($\mathcal{H}$) contains all remaining hits. See supplementary figure S18B, Supplementary Material online for an illustration of the three matching types. More formally, a hit $h_{s,m} \in \mathcal{H}$ belongs to Full($\mathcal{H}$) if the aligned region of *m* to *s* (excluding gaps) covers at least 90% of the length of *m*. If $h_{s,m}$ represents an overlap between *s* and *m* (allowing an overhang length of at most the 10% of the sequence length) then $h_{s,m} \in$ Overlap($\mathcal{H}$). Otherwise, $h_{s,m} \in$ Partial($\mathcal{H}$).

To eliminate potentially incomplete sequences, a sequence *s* is retained only if:

1) either at most the 30% of its hits belong to Overlap($\mathcal{H}$),
2) or, at least the 50% of its hits belong to either Full($\mathcal{H}$) or Partial($\mathcal{H}$).

These two conditions were introduced to take into account that Pfam could also contain partial sequences which could lead to the construction of very short models (which could be fully aligned in potentially incomplete sequences). We refer to the reduced set of sequences as $\mathcal{S}^*$.

IV. Data-Driven Selection of Models based on Sequence Best Matching. In order to restrict the analysis to a reduced set of models that remains representative of $\mathcal{M}_D$, we kept only those models that achieve one of the *k* best scores for at least a sequence in $\mathcal{S}^*$, for $k = 3$ (default). The rationale of this model filtering is to get rid of "noisy" models and, at the same time, significantly reduce the size of $\mathcal{M}_D$, from some thousands down to a few hundreds. We refer to the reduced set of models as $\mathcal{M}_D^*$. The parameter *k* can be set by the user.

V. Association of Two ProfileView Scores to Model Hits: The Normalized Bit-score and the Normalized Weighted Bit-score. Let L_s be the number of positions in a sequence *s* that match to a model *m* in a sequence-profile alignment (that is, no gap is considered in the counting). Given a hit $h_{s,m}$, we define the following two scores for it:

- a normalized bit-score $ns(h_{s,m}) = \text{score}(h_{s,m})/L_s$;
- a normalized weighted bit-score $nws(h_{s,m}) = \text{Wscore}(h_{s,m})/L_s$, where Wscore($h_{s,m}$) is the sum of bit-scores over those positions in the sequence-profile alignment having a bit-score ≥ 3 (that is, the positions

where m and s strongly agree). More formally, let $\sigma(s_i, m_j) = \log_2 (e(s_i, m_j)/bg(s_i))$ be the log-odds ratio of a residue s_i being emitted from a match state m_j with emission probability $e(s_i, m_j)$ and with null model background frequency $bg(s_i)$, defined by HMMER during the model construction and differing between amino acids (Eddy 1998). Given the list $\langle(s_{i_1}, m_{j_1}), \ldots, (s_{i_K}, m_{j_K})\rangle$ of the aligned residues of s against the model states of m and such that the posterior probability, computed by HMMER, of each aligned pair is greater than 75%, we define $\mathrm{Wscore}(h_{s,m}) = \sum_{z=1|\sigma(s_{i_z}, m_{j_z})\geq 3}^{K} \sigma(s_{i_z}, m_{j_z})$.

Both scores are computed for all hits $h_{s,m}$ and used to construct the ProfileView space of sequences.

VI. The Construction of a ProfileView Space of Sequences (fig. 1*B*). For each sequence $s \in \mathcal{S}^*$, we construct a vector v_s, where the dimension of v_s is $2|\mathcal{M}^*_{\mathrm{D}}|$ and $|\mathcal{M}^*_{\mathrm{D}}|$ is the number of models in $\mathcal{M}^*_{\mathrm{D}}$. The vector v_s contains the pairs of values $ns(h_{s,m})$ and $nws(h_{s,m})$, for each $m \in \mathcal{M}^*_{\mathrm{D}}$. If a model m does not have a hit on the sequence $s \in \mathcal{S}^*$, then we assume that $h_{s,m} \notin \mathcal{H}$ and let $ns(h_{s,m}) = 0$ and $nws(h_{s,m}) = 0$. Hence, we say that the ProfileView space $\mathcal{PV}$ is a $2|\mathcal{M}^*_{\mathrm{D}}|$-dimensional space, where each dimension is associated with either the normalized bit-score or the normalized weighted bit-score for some model $m \in \mathcal{M}^*_{\mathrm{D}}$. Each sequence is a point in $\mathcal{PV}$ and its position reflects the proximity of the sequence to CCMs in $\mathcal{M}^*_{\mathrm{D}}$.

VII. PCA and Dimensionality Reduction for ProfileView Space of Sequences. After constructing the ProfileView space $\mathcal{PV}$ for the sequences $s \in \mathcal{S}^*$, Principal Component Analysis (PCA) is performed to reduce its number of dimensions. More precisely, $\mathcal{PV}$ is reduced to a p-dimensional space $\mathcal{PV}^*$, where p is the minimum number of principal components that explain the c% of variance for the set $\mathcal{S}^*$. By default, $c = 99\%$. This value should decrease the number of dimensions to a few dozens. If a protein family is characterized by a large diversity of representative sequences, the user may have to loosen the constraints on variance by setting c to smaller values. c is a parameter that can be set by the user.

VIII. The ProfileView Tree Construction (fig. 1*B*). Sequences are clustered in $\mathcal{PV}^*$ using a hierarchical agglomerative strategy. Namely, we considered the Euclidean distance between vectors and Ward's minimum variance method for merging clusters. The logic of this criterion is to select, at each step, the pair of clusters that minimize the total variance within the cluster after the merging. Starting from all clusters being singletons, this bottom-up algorithm completes in $|\mathcal{S}^*| - 1$ agglomerative steps and allows to represent clusters in a hierarchical way and to define a rooted tree. More precisely, it produces a binary tree where each internal node defines a cluster of two or more elements (according to the chosen merge criterion). Moreover, in such a tree, the distances/dissimilarities between the merged clusters are encoded as edge weights.

IX. Association of Representative Models to ProfileView Subtrees (fig. 1C). To better explore subtrees in the ProfileView tree, potentially associated with known functions, we associated a *representative model* to the sets of sequences that label their leaves. Intuitively, a representative model separates a subset of sequences $\mathcal{C}$ from the rest of the sequences of the tree (this set is designated $\mathcal{S}^* \setminus \mathcal{C}$) in the ProfileView space $\mathcal{PV}^*$ (see fig. 1C). Given a model m in the library, let us call $\mathcal{C}^*_m$ the maximal subset of $\mathcal{C}$ where the model assigns higher scores to sequences in $\mathcal{C}^*_m$ than to sequences in $\mathcal{S}^* \setminus \mathcal{C}$. This must apply to at least one of the metrics—*ns* and *nws**—which define $\mathcal{PV}^*$ (see step III). For each model m in the library, we compute $\mathcal{C}^*_m$ and choose the model with a $\mathcal{C}^*_m$ of largest cardinality as the *representative model* of $\mathcal{C}$. If two models have the same maximum cardinality, we choose the model m that provides the best separation, that is, the model that maximizes the distance between the centroids of the sets $\mathcal{C}^*_m$ and $\mathcal{S}^* \setminus \mathcal{C}$ (again, computed according to the *ns* and *nws* metrics). If $\mathcal{C}$ is the set of sequences of a subtree T of the ProfileView tree (which is not the entire tree), then a *representative model m for $\mathcal{C}$ is associated with the root of T* when the following two conditions are met: (1) $\mathcal{C}^*_m$ includes at least half of the sequences in $\mathcal{C}$ and (2) $\mathcal{C}^*_m$ contains at least one sequence from each of the child subtrees of T. Note that a node in the ProfileView tree might be left without a representative model. When ProfileView returns a representative model for a node of the tree, it also returns a list of suboptimal models covering either the same amount of sequences $|\mathcal{C}^*_M|$ or 90% of $|\mathcal{C}|$.

X. Motif Extraction from Representative Models. A motif extracted from a representative model is the set of all amino acids characterizing well-conserved columns (i.e., match states) in the sequence alignment associated with the model, according to the hhblits' definition. That is, given a column of the multiple sequence alignment related to the model, an amino acid is *well conserved* if it occurs with a probability ≥ 0.6 before adding pseudo-counts and including gaps in the fraction count. See figure 1C.

On the Size and Diversity of the Model Library

ProfileView model library construction (step I) is designed in such a way that sequences coming (possibly extracted) from the Pfam FULL set would best represent the protein family. To show the relevance of ProfileView's automatic construction, we have tested ProfileView on smaller sets of Pfam "seed" sequences and have shown that not having a large and diverse set of sequences hinders classification performance. More precisely, we built two model libraries for the CPF family by considering 10 and 100 "seed" sequences, respectively. These sequences have been randomly chosen in the FULL set of CPF sequences. The resulting ProfileView trees are reported in supplementary figures S19 and S20, Supplementary Material online. A straightforward comparison with figure 3 and supplementary figure S2, Supplementary Material online shows the limitations of a functional classification based on a restricted number of models.

Parameters Used in ProfileView Analysis of the Seven Protein Families

ProfileView was run with the same default parameters $k = 3$ and $c = 99\%$ on all protein families in table 1, with the exception of the WW domain, which is characterized by a wide variability of sequences. For the WW domain family, we set $k = 5$ and $c = 80\%$ (see steps IV and VII). The parameter $c = 80\%$ allowed to obtain a space of 11 dimensions starting from a total of 2,488, versus 206 dimensions obtained with a threshold of $c = 99\%$. The parameter $k = 5$ allowed to increase the number of best matching models to 1,244 versus 845 obtained with $k = 3$. Intuitively, to classify datasets of sequences with high variability, such as the WW domain family, the number of models representing the dataset should be large (>1,000).

Motif Graphical Representation

The model logos were built using the Weblogo python package (Crooks et al. 2004) (version 3.7) which allowed us to easily export sequence logos (Schneider and Stephens 1990). Amino acids are colored according to chemical properties: neutral polar amino acids (G, S, T, Y, C) show in green, acidic polar (Q, N) violet, positively charged (K, R, H) blue, negatively charged (D, E) red, and hydrophobic (A, V, L, I, P, W, F, M) black.

The graphical representation of a motif associated with some representative model has been augmented by additional information that helps to easily compare the motif across representative models. Namely, we have highlighted, by a colored "dot," those positions that are well conserved in other representative models. Given a reference model M_r and a query model M_q, a dot is put under a well-conserved column of M_r, if there exists a column in the query model M_q: (1) aligning in hhblits with a score greater than +1.5 (i.e., fairly similar amino acid profiles) and posterior probability greater than 0.8; (2) containing a most conserved amino acid which is the same as in M_r and is also well conserved. A circled dot indicates an aligned column in M_q satisfying 1 but not 2. This means that the most conserved amino acid in M_r shows <60% frequency in M_q. Note that, in this case, M_r and M_q might display different most conserved amino acids.

It is important to note that given two models and one position, the score assigned to that position in the hhblits pairwise alignment of the models depends on the reliability of the query-template alignment (https://github.com/soedinglab/hh-suite/wiki). Depending on which of the models is considered as template, the scores assigned to the same position may vary (confidence values are obtained from posterior probabilities calculated in the Forward-Backward algorithm of hhblits). In particular, hhblits warns that the confidence score for an aligned position depends on the alignment confidence of the neighboring regions. As a result, the alignment score of some conserved positions may decrease due to the presence of a highly variable region in their vicinity, possibly containing gaps. This explains why, for the aligned positions of two motifs, we may miss to indicate related positions or we may display different color dots. An example of missing related positions is illustrated by position 102 in the NCRY motif and position 103 in the Plant photoreceptor CRY motif of CPF. The two motifs clearly diverge within the region adjacent to positions 102/103, justifying a difficult model alignment and a low confidence score at 102/103. A second example, illustrating the asymmetry of colored dots, is position 102 in the NCRY motif aligned with position 95 in CRY Pro. While the CRY Pro motif records the colored dot for a match with NCRY, this is not true for the NCRY motif. In fact, while the two positions align together with a confidence score of 0.8 for the CRY Pro model taken as a template, they also align when the NCRY model is taken as the template but with a confidence score that drops at 0.6.

Phylogenetic Tree Construction for CPF, FAD, and WW Sequences

The multiple sequence alignments of CPF sequences and FAD sequences were computed using MUSCLE version v3.8.31 (Edgar 2004), and were then trimmed using trimAl version 1.4.rev22 (Capella-Gutiérrez et al. 2009) with a gap cutoff of 0.01 (i.e., columns containing more than 99% of gaps have been removed). Then, for each sequence alignment, we selected the best evolutionary model using ProtTest (version 3.4.2) (Darriba et al. 2011). More precisely, the evolutionary model best fitting the data was determined by comparing the likelihood of all models according to the Akaike Information Criterion (AIC). The model optimization of ProtTest was run using a maximum-likelihood-tree strategy and the tree generated for the best-fit model (VT+G+F) was considered as input for the construction of the final phylogenetic tree (with parameter $\alpha = 1.061$). In particular, the construction of a maximum-likelihood phylogenetic tree has been carried out with PhyML 3.0 (Guindon et al. 2010) that optimized the output tree with Subtree-Prune-Regraft (SPR) moves and considering the SH-like approximate likelihood-ratio test. Finally, branches with a support value smaller than 0.5 were collapsed. The phylogenetic tree for the set of homologous CPF sequences used to validate ProfileView is reported in supplementary figure S3, Supplementary Material online and contains 307 leaves corresponding to the 307 CPF sequences containing the FAD-binding domain. The phylogenetic tree for the set of 307 FAD sequences is reported in supplementary figure S4, Supplementary Material online.

The procedure used to generate the phylogenetic tree for WW domain sequences is the same as that used for CPF and FAD sequences. The best-fit model (computed with ProtTest) is RtREV+I+G, with parameters $\alpha = 1.647$ and p-inv $= 0.028$.

Phylogenetic and ProfileView trees have been generated with iTOL (Letunic and Bork 2019).

Output Files of ProfileView

ProfileView produces several output datasets: the model library, the ProfileView tree, the list of representative models associated with internal nodes of the tree.

Additionally, ProfileView offers the user the possibility to choose a list of representative models to compare. The first model on this list is considered a reference model. A first output provides a logo showing all conserved positions together with a list of colored dots (possibly circled) obtained after a pairwise comparison of a model in the list with the reference model (see Materials and Methods above; see for example fig. 1C). A second output provides a logo that shows an intermediate representation of the positions in the reference model, that is, it shows all conserved positions in the associated motif and all positions that are not conserved in the reference model but which are conserved in some model in the list (see e.g., fig. 4C).

Comparison with Other Tools

CUPP (Barrett and Lange 2019), eCAMI (Xu et al. 2020), and PANTHER (Mi et al. 2012, 2013) were run for comparison with ProfileView. CUPP v1.0.14 was run with CUPPclustering.py and parameter-cluster to execute the clustering (http://www.bioengineering.dtu.dk/CUPP). For the comparison on the GH30 family, we considered the "CUPP groups" predicted with the latest version of the CUPP-Predict tool (v3.2.1) available at https://cupp.info. Unfortunately, from the README, the updated "clustering" version of CUPP is still under development and CUPP results on the CPF family still refer to the application of CUPP version 1.0.14.

eCAMI was downloaded from https://github.com/zhanglabNKU/eCAMI (commit 5b00a038).

The PANTHER HMM library version 15.0 and the pantherScore2.2 tool (scoring protein sequences against the library) were retrieved at http://www.pantherdb.org. We used `pantherScore2.2.pl` with parameters `-l [ PANTHER15.0 library] -D B -n`, where `-D B` allows to visualize the best hit in the output and `-n` allows to visualize family and subfamily names in the output.

Evaluation

To assess the robustness of ProfileView classification, we compared it to functional grouping in the literature. For each protein family, we considered a set of functionally characterized sequences that have been human curated and, in most cases, tested experimentally. Each family is characterized by several functional subclasses (table 2). These datasets constitute independent sets for testing ProfileView. These test sets are available at http://www.lcqb.upmc.fr/profileview/.

We want to determine whether characterized sequences of the same functional group localize together in the ProfileView tree. For this, we identify the largest subtrees of at least 2 sequences, endowed with a representative model and comprising at least 75% of characterized sequences which belong to the same functional class. We use this overrepresentation of a functional subclass in a ProfileView subtree to label the subtree with that function and the characterized sequences as "well-classified" (denoted "W"). Note that some of the ProfileView subtrees will be labeled by a function and others may not be. Within subtrees labeled by a function, there may be characterized sequences belonging to other functional subclasses that we denote as misplaced (denoted "M"). Within subtrees that are not labeled by a function, there might be characterized sequences belonging to functional subclasses that we denote as unclassified (denoted "U"). To resume, for each known functional subgroup of characterized sequences in the dataset, we count:

1) the number of unclassified sequences (denoted "U"), that is the number of characterized sequences that do not belong to some subtree labeled by a function,
2) the number of misplaced sequences (denoted "M"), that is the number of characterized sequences that belong to a subtree labeled by a different functional class,
3) the number of well-classified sequences (denoted "W"), that is the number of characterized sequences that are over-represented in a subtree and allowed for the identification of its functional label.

These three numbers allow to evaluate whether ProfileView classifies well or not sequences of a given protein family within different functional subclasses. Also, since experimental functional accuracy might differ across functional groups, we consider subtrees comprising at least 75% of their characterized sequences within a functional subgroup which are not endowed with a representative model. For them, we count W/M/U sequences. These subtrees provide evidence of functional organization while highlighting the difficulty of identifying motifs, described by representative models that are specific to a functional class.

Our evaluation criteria were designed to highlight several possible scenarios. If a protein family is known to have *n* functional subclasses and the ProfileView tree is comprised of *n* subtrees labeled with the *n* distinguished functions, the ProfileView classification can be considered to be fully accurate for the protein family. On the other hand, if some ProfileView subtree of sequences remains unlabeled, this could suggest missing functional knowledge for the protein family. Finally, since the experimental functional accuracy between different functional groups might differ, ProfileView might suggest an alternative sequence classification.

Computing Time and Memory Usage

The most costly computational part of the ProfileView pipeline is the construction of the profile models for a protein domain sequence. The program was tested using 16 threads on a single machine equipped with an Intel Xeon E5-2670 CPU running at 2.60 GHz, with 128 GB of RAM, and a Linux operating system (CentOS release 6.5). Supplementary table S10, Supplementary Material online summarizes, for each protein family, the time complexity, and the RAM used for the model library construction and the classification step.

The time used for the model library construction depends on the number of models and the length of the domain. Once a library is constructed, it can be used for the analysis of different protein families. Note that the library constructed for the Radical SAM domain was used for both the Methylthiotransferase family and the SPASM/twitch domain containing family analyzes in supplementary text 1, Supplementary Material online.

Classification time depends on two main sub-steps: (1) HMMER annotation, which mainly depends on the number of sequence-model comparisons, rather than the sequence length, and (2) the computation of the scores used to construct the ProfileView sequence space. Step 2 is the most time-consuming as it involves the parsing of all sequence-profile alignments. ProfileView current implementation of this part of the code is single-threaded but this sub-step could be performed in parallel, possibly in future versions of ProfileView.

Note that, for the WW domain family presented in supplementary text 1, Supplementary Material online, (Tubiana et al. 2019) indicates about 1–2 days of computing time on an Intel Xeon Phi processor with 2 × 28 cores to run RBM analysis. ProfileView classifies this family in less than 9 h.

Implementation and Software Availability

ProfileView was developed and tested under a UNIX operating system, using Bash, Python, and R scripts. It exploits GNU parallel (Tange 2018), if available on the system, in order to run some jobs in parallel. It is implemented in three main parts carrying out the following pipeline modules: the construction of a single-domain model library, the generation of the ProfileView tree along with its representative models, the comparison of selected representative models and the identification of conserved positions/motifs. ProfileView is available at http://www.lcqb.upmc.fr/profileview/ under the version 2.1 of the CeCILL Free Software License.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution online.*

Acknowledgements

We thank Simona Cocco and Jérôme Tubiana for providing to us the dataset of WW sequences used in their study. LabEx CALSIMLAB (public grant ANR-11-LABX-0037-01 constituting a part of the “Investissements d’Avenir” Program ANR-11-IDEX-0004-02) (RV); the Institut Universitaire de France (AC); access to the HPC resources of the Institute for Scientific Computing and Simulation (Equip@Meso project - ANR-10-EQPX- 29-01, Excellence Program “Investissement d’Avenir”) (AC); Fondation Bettencourt-Schueller (Coups d’Élan pour la Recherche Française-2018) (AF); LabEx DYNAMO (public grant ANR-11-LABX-0011-01) (AF).

Authors’ Contributions

R.V. and A.C. conceived and designed the experiments. R.V. performed the experiments. E.L. performed the structural analysis of CPF classes. R.V., J.P.B., A.F. and A.C. analyzed the data. A.C., R.V., J.P.B. and A.F. wrote the paper. All authors read and approved the final manuscript.

Data Availability

The set of sequences for CPF, FAD, WW domain, Glyco-hydro-30 and Glyco-hydro-30-2 for GH30, HAD, and HAD_2 for Haloacid Dehalogenase, B12-binding and B12-binding_2 for B12-binding domain containing, Radical SAM for Methylthiotransferase and SPASM/twitch domain containing subgroups, and SPASM for SPASM/twitch domain containing subgroup, model libraries, phylogenetic trees, ProfileView trees are available at http://www.lcqb.upmc.fr/profileview/.

References

Akiva E, Brown S, Almonacid DE, Barber II AE, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC, *et al.* 2014. The structure–function linkage database. *Nucleic Acids Res.* **42**(D1): D521–D530.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**-(17):3389–3402.

Amato A, Dell’Aquila G, Musacchia F, Annunziata R, Ugarte A, Maillet N, Carbone A, Sanges R, Iudicone D, *et al.* 2017. Marine diatoms change their gene expression profile when exposed to microscale turbulence under nutrient replete conditions. *Sci Rep.* **7**:3826.

Barrett C, Hughey R, Karplus K. 1997. Scoring hidden Markov models. *Bioinformatics.* **13**(2):191–199.

Barrett K, Lange L. 2019. Peptide-based functional annotation of carbohydrate-active enzymes by conserved unique peptide patterns (CUPP). *Biotechnol Biofuels.* **12**(1):102.

Basu MK, Poliakov E, Rogozin IB. 2009. Domain mobility in proteins: functional and evolutionary implications. *Brief Bioinform.* **10**(3): 205–216.

Bernardes J, Zaverucha G, Vaquero C, Carbone A. 2016. Improvement in protein domain identification is reached by breaking consensus, with the agreement of many profiles and domain co-occurrence. *PLoS Comput Biol.* **12**(7):e1005038.

Björn LO. 2015. *Photobiology: the science of light and life.* New York: Springer.

Boari de Lima E, Meira W, Melo-Minardi RCd. 2016. Isofunctional protein subfamily detection using data integration and spectral clustering. *PLoS Comput Biol.* **12**(6):e1005001.

Bonetta R, Valentino G. 2020. Machine learning techniques for protein function prediction. *Proteins.* **88**(3):397–413.

Brettel K, Byrdin M. 2010. Reaction mechanisms of dna photolyase. *Curr Opin Struct Biol*. **20**(6):693–701.

Cao M, Pietras CM, Feng X, Doroschak KJ, Schaffner T, Park J, Zhang H, Cowen LJ, Hescott BJ. 2014. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics*. **30**(12):i219–i227.

Cao R, Cheng J. 2016. Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks. *Methods*. **93**:84–91.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. **25**(15):1972–1973.

Chaves I, Pokorny R, Byrdin M, Hoang N, Ritz T, Brettel K, Essen L-O, van der Horst GT, Batschauer A, Ahmad M. 2011. The cryptochromes: blue light photoreceptors in plants and animals. *Annu Rev Plant Biol*. **62**:335–364.

Clark WT, Radivojac P. 2011. Analysis of protein function and its prediction from amino acid sequence. *Proteins*. **79**(7): 2086–2096.

Crooks GE, Hon G, Chandonia J -M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. **14**(6):1188–1190.

Czarna A, Berndt A, Singh HR, Grudziecki A, Ladurner AG, Timinszky G, Kramer A, Wolf E. 2013. Structures of drosophila cryptochrome and mouse cryptochrome1 provide insight into circadian function. *Cell*. **153**(6):1394–1405.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*. **27**(8):1164–1165.

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I. 2017. Cath: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res*. **45**-(D1):D289–D295.

De Filippo C, Ramazzotti M, Fontana P, Cavalieri D. 2012. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Brief Bioinform*. **13**(6): 696–710.

Deng M, Zhang K, Mehta S, Chen T, Sun F. 2004. Prediction of protein function using protein-protein interaction data. *J Comput Biol*. **10**(6):947–960.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics*. **14**(9): 755–763.

Eddy SR. 2011. Accelerated profile HMM searches. *PLOS Comput Biol*. **7**(10):1–16.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**(5):1792–1797.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*. **8**(3):163–167.

Emmerich H-J, Saft M, Schneider L, Kock D, Batschauer A, Essen L-O. 2020. A topologically distinct class of photolyases specific for uv lesions within single-stranded dna. *Nucleic Acids Res*. **48**(22): 12845–12857.

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. 2005. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol*. **1**(5):e45.

Engelhardt BE, Jordan MI, Srouji JR, Brenner SE. 2011. Genome-scale phylogenetic function annotation of large and diverse protein families. *Genome Res*. **21**(11):1969–1980.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, *et al*. 2014. Pfam: the protein families database. *Nucleic Acids Res*. **42**(D1): D222–D230.

Fortunato AE, Jaubert M, Enomoto G, Bouly J-P, Raniello R, Thaler M, Malviya S, Bernardes JS, Rappaport F, Gentili B, *et al*. 2016. Diatom phytochromes reveal the existence of far-red light based sensing in the ocean. *Plant Cell*. **28**(3):616–628.

Furnham N, Sillitoe I, Holliday GL, Cuff AL, Rahman SA, Laskowski RA, Orengo CA, Thornton JM. 2012. Funtree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res*. **40**(D1):D776–D782.

Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet*. **14**(5):360–366.

Gaudet P, Livstone MS, Lewis SE, Thomas PD. 2011. Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Brief Bioinform*. **12**(5):449–462.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Syst Biol*. **59**(3):307–321.

Gumerov VM, Zhulin IB. 2020. Trend: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Res*. **48**(W1):W72–W76.

Hawkins T, Luban S, Kihara D. 2006. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci*. **15**(6):1550–1556.

Hirano A, Braas D, Fu Y-H, Ptáček LJ. 2017. Fad regulates cryptochrome protein stability and circadian clock in mice. *Cell Rep*. **19**(2):255–266.

Ingham RJ, Colwill K, Howard C, Dettwiler S, Lim CS, Yu J, Hersi K, Raaijmakers J, Gish G, Mbamalu G, *et al*. 2005. WW domains provide a platform for the assembly of multiprotein networks. *Mol Cell Biol*. **25**(16):7092–7106.

Jaubert M, Bouly J-P, d'Alcalà MR, Falciatore A (2017). Light sensing and responses in marine microalgae. *Curr Opin Plant Biol*, **37**, 70–77.

Karchin R, Kelly L, Sali A. 2005. Improving functional annotation of non-synonomous SNPs with information theory. In: Altman RB, Jung TA, Klein TE, Dunker AK, Hunter L, editors. Pacific symposium on biocomputing 2005. Singapore: World Scientific. p. 397–408.

Keeling DM, Garza P, Nartey CM, Carvunis A -R. 2019. Philosophy of biology: The meanings of 'function' in biology and the problematic case of de novo gene emergence. *Elife*. **8**:e47014.

Kulmanov M, Hoehndorf R. 2020. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*. **36**(2): 422–429.

Lee D, Redfern O, Orengo C. 2007. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*. **8**(12):995–1005.

Lees JG, Dawson NL, Sillitoe I, Orengo CA. 2016. Functional innovation from changes in protein domains and their combinations. *Curr Opin Struct Biol*. **38**:44–52.

Letovsky S, Kasif S. 2003. Predicting protein function from protein/-protein interaction data: a probabilistic approach. *Bioinformatics*. **19**(suppl_1):i197–i204.

Letunic I, Bork P. 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. **47**(W1): W256–W259.

Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. 2009. Protein function annotation by homology-based inference. *Genome Biol*. **10**(2):1–8.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*. **42**(D1):D490–D495.

Ma H, Holub D, Gillet N, Kaeser G, Thoulass K, Elstner M, Krauß N, Lamparter T (2019). Two aspartate residues close to the lesion binding site of agrobacterium (6-4) photolyase are required for Mg2+ stimulation of dna repair. *FEBS J*. **286**(9), 1765–1779.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the panther classification system. *Nat Protoc*. **8**(8):1551–1566.

Mi H, Muruganujan A, Thomas PD. 2012. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*. **41**(D1): D377–D386.

Mirdita M, Galiez C, Martin MJ, Söding J, Steinegger M. 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res*. **45**(D1):D170–D176.

Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*. **21**(suppl_1):i302–i310.

Otte L, Wiedemann U, Schlegel B, Pires JR, Beyermann M, Schmieder P, Krause G, Volkmer-Engert R, Schneider-Mergener J, Oschkinat H. 2003. WW domain sequence activity relationships identified using ligand recognition propensities of 42 WW domains. *Protein Sci*. **12**(3):491–500.

Pal D, Eisenberg D. 2005. Inference of protein function from protein structure. *Structure*. **13**(1):121–130.

Pazos F, Sternberg MJ. 2004. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci*. **101**(41):14754–14759.

Pham M, Lichtarge O. 2020. Graph-based information diffusion method for prioritizing functionally related genes in protein-protein interaction networks. In: Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray TA, Klein TE, editors. Pacific symposium on biocomputing 2020. Singapore: World Scientific. p. 439–450.

Ponting CP, Dickens NJ. 2001. Genome cartography through domain annotation. *Genome Biol*. **2**(7):comment2006-1.

Prakash T, Taylor TD. 2012. Functional assignment of metagenomic data: challenges and applications. *Brief Bioinform*. **13**(6): 711–727.

Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*. **9**:173–175.

Rosensweig C, Reynolds KA, Gao P, Laothamatas I, Shan Y, Ranganathan R, Takahashi JS, Green CB. 2018. An evolutionary hotspot defines functional differences between cryptochromes. *Nat Commun*. **9**(1):1138.

Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. 2005. Natural-like function in artificial WW domains. *Nature*. **437**(7058):579.

Sahraeian SM, Luo KR, Brenner SE. 2015. Sifter search: a web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res*. **43**(W1):W141–W147.

Sancar A. 2003. Structure and function of dna photolyase and cryptochrome blue-light photoreceptors. *Chem Rev*. **103**(6): 2203–2238.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*. **18**(20):6097–6100.

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*. **5**(12):e1000605.

Sharan R, Ulitsky I, Shamir R. 2007. Network-based prediction of protein function. *Mol Syst Biol*. **3**(1):88.

Shin H, Lisewski AM, Lichtarge O. 2007. Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics*. **23**(23):3217–3224.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big data: astronomical or genomical? *PLoS Biol*. **13**(7):e1002195.

Tange O. 2018. *GNU parallel 2018*. Frederiksberg: Ole Tange.

Törönen P, Medlar A, Holm L. 2018. Pannzer2: a rapid functional annotation web server. *Nucleic Acids Res*. **46**(W1):W84–W88.

Tubiana J, Cocco S, Monasson R. 2019. Learning protein constitutive motifs from sequence data. *eLife*. **8**:e39397.

Ugarte A, Vicedomini R, Bernardes J, Carbone A. 2018. A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome*. **6**(1):149.

Vazquez A, Flammini A, Maritan A, Vespignani A. 2003. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*. **21**(6):697–700.

Wan C, Jones DT. 2020. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat Mach Intell*. **2**(9):540–550.

Wass MN, Sternberg MJ. 2008. Confunc–functional annotation in the twilight zone. *Bioinformatics*. **24**(6):798–806.

Wen Z-n., Wang K-l., Li M-l., Nie F-s., Yang Y. 2005. Analyzing functional similarity of protein sequences with discrete wavelet transform. *Comput Biol Chem*. **29**(3):220–228.

Worthington EN, Kavakli İH, Berrocal-Tito G, Bondo BE, Sancar A. 2003. Purification and characterization of three members of the photolyase/cryptochrome family blue-light photoreceptors from vibrio cholerae. *J Biol Chem*. **278**(40):39143–39154.

Xu J, Zhang H, Zheng J, Dovoedo P, Yin Y. 2020. eCAMI: simultaneous classification and motif identification for enzyme annotation. *Bioinformatics*. **36**(7):2068–2075.

Zhang C, Freddolino PL, Zhang Y. 2017. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res*. **45**(W1):W291–W299.