ProfileView: multiple probabilistic models resolve protein families functional diversity

Riccardo Vicedomini[1,2,*], Jean Pierre Bouly[1,3,*], Elodie Laine[1], Angela Falciatore[1,3] and Alessandra Carbone[1,4]

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

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

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

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

alessandra.carbone@lip6.fr

July 27, 2019

[*]The first two authors share equal contribution.

Abstract

Sequence functional classification has become a fundamental bottleneck to the understanding of the myriad of protein sequences accumulating in our databases due to the recent progress in genomics and metagenomics. The large 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 living organisms and for biotechnological applications.

ProfileView is a novel computational method designed to functionally classify sets of homologous sequences. Its architecture strongly relies on the structure of biological data, and answers to the challenge of automatically partitioning datasets of protein sequences in pertinent subfamilies based on meaningful conservation patterns. It constructs a library of probabilistic models accurately representing the functional variability of protein families, and extracts biologically interpretable information from the classification process. It applies to protein families that are not necessarily large, nor conserved, whose homologs might be very divergent and for which functions should be discovered or characterised more precisely.

As a proof of concept, we apply ProfileView to the Cryptochrome/Photolyase family (CPF) and to the WW domain family, two widespread classes of proteins showing a large variety of functions and high sequence divergence. Decades of experimental studies on these families, functionally characterizing sequences and highlighting constitutive motifs, allow us to validate the two functional organisations obtained with the ProfileView approach. In addition, the method allows to identify a distinct functional group for the CPF, likely corresponding to novel photoreceptors. Thus, ProfileView appears as a powerful tool to classify protein sequences by function, screen sequences towards the design of accurate functional testing experiments and, possibly, discover new functions of natural sequences.

Software and data availability: `http://www.lcqb.upmc.fr/profileview/` **A restrained access is momentarily set. Once the article is accepted, all information will be freely accessible.**

Key words: protein sequence; functional annotation; protein classification; cryptochrome; photolyase; photoreceptor; clustering; probabilistic profile.

1 Introduction

The functional classification of biological sequences has become a fundamental bottleneck to the understanding of the ever-increasing genomic and metagenomic sequence data accumulating in our databases. This quest depends on the correct domain annotation of coding genes [1–3], which, in the past, was handled by sequence homology-, signature- and feature-based approaches. The first and most intuitive approach searches for homologous sequences to already known protein or domain sequences. However, sequence homology conveys a serious pitfall: it is defined on the basis of evolution, not function, and many homologues diversified their functions. More reliable methods use protein signatures, which are descriptions of protein or domain families derived from multiple sequence alignments. These approaches use consensus sequences to produce a probabilistic model describing the conserved characteristics of a domain across sequences in the family and use it to evaluate sequence similarity. For protein families represented by very divergent sequences, these models might provide a very weak description and become of restrained use. The third class of methods selects an appropriate set of features (like short sequence segments or wavelet decompositions) but needs the availability of a large and very diversified dataset of sequences to perform classification successfully. Novel computational approaches classifying sequences by function and overcoming the limitations intrinsic to existing methods would help screening millions of sequences to design accurate experiments directed to functional testing and to discover new functions.

We designed ProfileView, a pipeline exploring homologous domain sequences and classifying them into functional groups. ProfileView is strongly based on the understanding of the structure of the data (imposed by the evolutionary history of the sequences), it is applicable at large scale to classify hundreds/thousands of sequences, and it extracts biologically interpretable information from the classification process.

The structure of the data is learned in the first step of the approach, where the set of homologs is analysed within the larger space of natural sequences: for each homolog, a conservation profile of the homolog and the sequences in its neighbourhood (sharing high sequence identity) is constructed, as a probabilistic model, in order to reflect the structural and functional characteristics of the homolog (**Fig. 1A**).

It is the usage of a multitude of probabilistic models that makes the distinctiveness of ProfileView in the context of sequence classification. This multiplicity of models demonstrated to be very powerful in the CLADE/MetaCLADE [4,5] pipelines, where it produces very accurate domain annotations for full genomes and metagenomic/metatranscriptomic datasets, and allows for the discovery of new homologous sequences enriching protein families [6, 7]. In ProfileView, the ensemble of probabilistic models plays the role of the "expert" in classification, in the same way as a biologist, looking at sequences, finds patterns by hand, learns about meaningful characteristics from subsets of similar sequences and then uses combined information to classify new sequences. Indeed, ProfileView codes the "biologist expertise" by using all models to look at sequences to be classified. Each model will evaluate a sequence with a value representing the

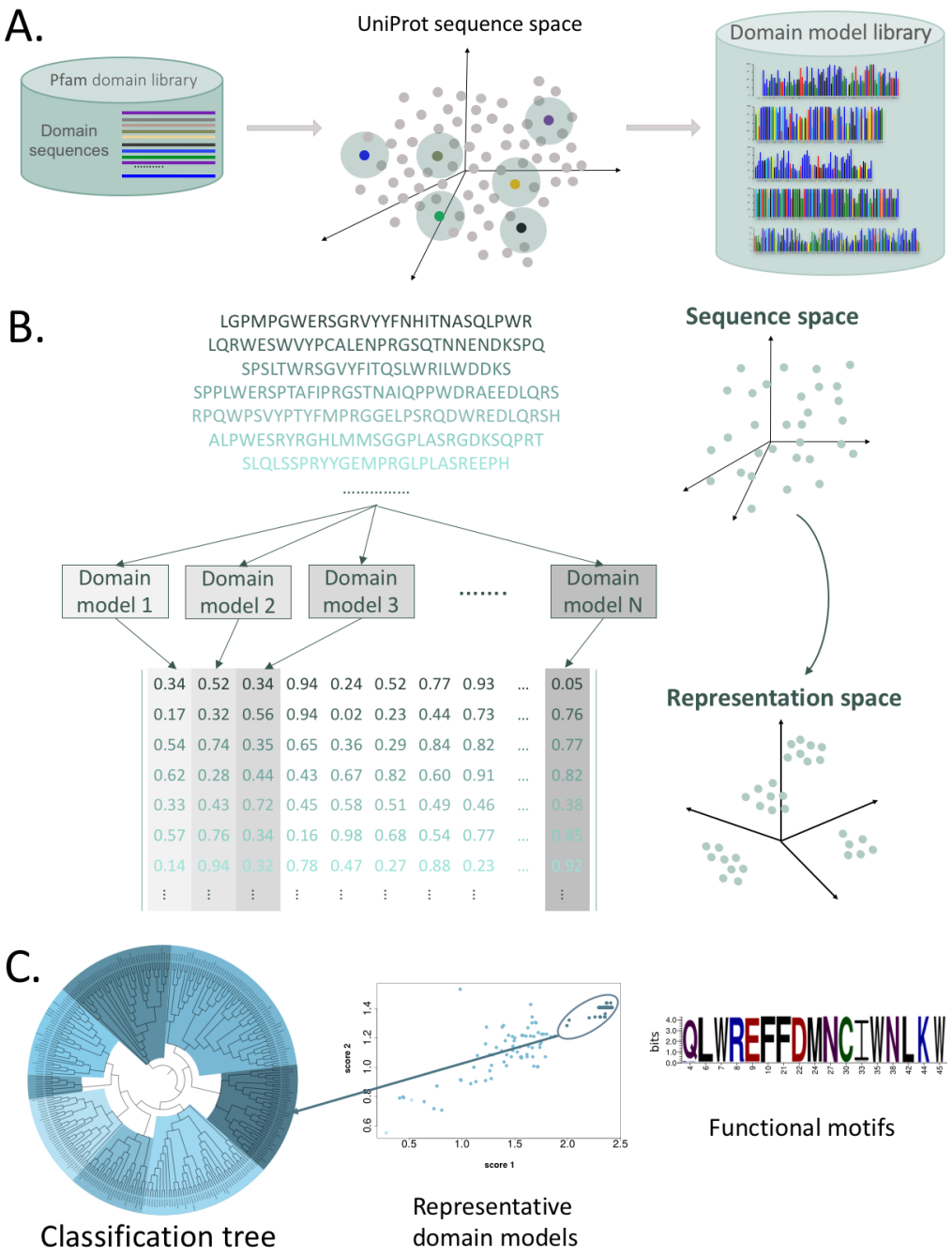


Figure 1. Schema of the ProfileView approach. A. Model library construction in ProfileView: a pool of representative sequences from the Pfam domain library are selected for the domain under study. For each representative domain sequence (coloured dots), ProfileView searches for close sequences in UniProt and constructs with HH-Blits several probabilistic models making a library of models for the domain. **B.** Sequences (dots in sequence space, top right) code for proteins with different functions. ProfileView defines a probabilistic mapping from sequences onto the representation space (bottom right) which is indicative of the function of the corresponding protein sequences. The mapping is realised through the contribution of the domain probabilistic models that evaluate the probability of their match against each sequence. Each protein sequence is mapped into a vector of real numbers (coloured row in the matrix, bottom) representing the quality of the match of all models. **C.** ProfileView clusters sequences in the representation space (B) to reconstruct a classification tree. It identifies best representative models for subtrees and their characteristic functional motifs.

probability that the sequence features the functional patterns encoded in the model. Hence, a sequence becomes a vector of values, each one provided by a different model. In more abstract terms, ProfileView transforms the space of sequences into a space of vectors of real numbers in a high dimensional space, where the number of dimensions correspond to the number of generated models (**Fig. 1B**). Sequence classification is then realised in the high dimensional space, where nearby sequences are supposed to share the same functional patterns, and where a functional tree is constructed (**Fig. 1C**, left). This simple schema, provided by a single-layer architecture of models filtering the sequences in parallel, results in a powerful approach to screen hundreds/thousands of sequences for functional classification and discovery of new functions. In particular, ProfileView applies to protein families that are not necessarily large nor conserved. Indeed, on the one hand, the construction of several models allows to classify protein families that are not conserved, and, on the other hand, the procedure to construct the models demands for a relatively small number of sequences (a minimum number of 20) allowing the method to be applied to protein families that are not represented by many sequences.

ProfileView is a general method, applicable to any protein family. To highlight its power, we applied it to the Cryptochrome/Photolyase Family (CPF), proteins widely distributed in all kingdoms of life and controlling a variety of critical biological processes. CPF members bind flavin chromophores and are characterised by peculiar photochemical and photobiological properties [8]. Although CPF proteins display similar structures and can use similar chromophores for light sensing, they show an impressive functional diversification [9–12]: cryptochromes (CRY) are mainly photoreceptors (PR) (specifically noted PR CRY in the following) involved in many biological responses to light (*e.g.* photomorphogenesis, entrainment of the circadian clock). However, some CRYs are light-independent transcriptional regulators taking part in the central circadian oscillator generating biological rhythms. On the other hand, photolyases (PL) are photoactive enzymes that use blue light to repair UV-damaged DNA (pyrimidine dimers) without base or nucleotide excision. Photolyases can mend two different types of UV-induced DNA damage, either dsDNA cyclobutane pyrimidine dimer (CPD) or (6-4) pyrimidine-pyrimidone photoproducts, and are thus classified as either CPD or (6-4) photolyases. Moreover, some photolyases can repair ssDNA cyclobutane pyrimidine dimer. The initially proposed functional separation between CRYs and PLs has however gradually started to vanish, as there are now several examples of CPF members exhibiting both functions [13–15]. In the last decade, new CPF variants, exhibiting different photobiological properties or functions, have been discovered, changing current views on their evolution [9] and functional diversification [13, 14, 16, 17]. Some CPF members have even been used for optogenetic applications [18, 19] or proposed as magnetoreceptors [20].

Although a lot of experimental progress has been made, CPF functions could not be anticipated by the analysis of domain organisation due to a very simplified architecture of the CPF sequences, nor by structural properties due to the high similarity of their protein structures, nor by primary protein sequences. Tools employed for phylogenetic reconstruction [21–24] did not allow to resolve different functions (*e.g.*, light-dependent photolyases and light-independent

transcriptional regulators) or to anticipate the function of new CPF sequences.

Here, 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, this domain non-covalently binds the FAD chromophore for light absorption or protein stabilization. It has been shown that the FAD chromophore can be found in three redox states [10] possibly associated to different functionality, and that, depending on the function of the CPF, the FAD binding domain interacts specifically either with the damaged DNA or with other domains present in CPF proteins (*e.g.*, C-ter extensions in some PR CRY) or with other partners [25] suggesting that it could contribute in different ways to the specificity of the CPF functions. ProfileView bases CPF analysis on the FAD binding domain and positively confirms our hypothesis.

The probabilistic models identified by ProfileView as best characterising CPF functional classes have been used to analyse the determinants of the functional specificity of CPF classes and to bring new insights into biochemical mechanisms underlying the functional diversity of CPFs. To validate our approach, ProfileView was assessed on known functionally characterised CPF sequences, precisely on a large number of functionally known sequence positions for characterised sequences, and on the modelling of CPF structures. It provided a functional classification for a large number of previously functionally uncharacterised sequences, and novel information on conserved amino acids that could be useful to design testing experiments.

To demonstrate its generality, ProfileView was validated on a second important protein family, the WW domain family, found in many eukaryotes. WW domains are protein modules mediating protein-protein interactions through peptide motifs recognition. Their functional classification is far from being straightforward because based on target peptide sequence motifs and their binding affinity. In particular, it was observed that the same WW domain can bind with variable affinity to multiple peptides [26–28], and that it is the modulation of binding properties that make hundreds of WW domains to interact specifically with hundreds of putative ligands in mammalian proteomes [26]. A fine classification of WW domains based on sequence analysis is of major interest for the functional annotation of novel (meta)genomic WW domain sequences. ProfileView demonstrated to recapture, in a precise manner, the experimental diversification proposed in the literature [27, 29] and to functionally classify hundreds of novel natural sequences.

2 Results

2.1 A highlight of ProfileView

Our methodological approach to sequence classification, named ProfileView, is presented for the analysis of the cryptochrome/photolyase family (CPF). It bases the analysis on the FAD binding domain, occurring in all CPF sequences, and considers the set $\mathcal{S}_{\mathrm{CPF}}$ of 397 CPF sequences spanning the whole phylogenetic tree, of which 69 are functionally characterized CPF homologs

and the remaining ones are known functionally uncharacterised sequences (**Fig. S1**, top).

ProfileView comprises four main modules (**Fig. S1**): (i) the model library construction for the FAD binding domain, (ii) a sequence filtering, (iii) a multidimensional representation space construction and (iv) a functional tree construction. We detail these steps below.

ProfileView starts with the construction of a library, $\mathcal{M}_{\mathrm{FAD}}$, of probabilistic models [30] for the FAD binding domain (**Fig. S1** and **Fig. 1**). These models, called *Clade-Centered Models* (CCM) in [4], have been introduced to annotate domains in genomes and metagenomes [5] especially characterised by very divergent sequences. Here, we build CCMs starting from homologous sequences which are much closer to each other than in previously constructed CCMs [4]. The rationale is to highlight conserved regions that might play a role towards the identification of a specific function and that might remain hidden by considering distant homologues. We constructed 4615 CCMs for the FAD binding domain.

The second step selects CPF sequences containing the FAD binding domain and filters models in the $\mathcal{M}_{\mathrm{FAD}}$ library. All CCMs in $\mathcal{M}_{\mathrm{FAD}}$ are used to select those CPF sequences among the 397 ones in $\mathcal{S}_{\mathrm{CPF}}$ that contain the FAD binding domain. Those sequences showing very little evidence of the domain presence are discarded. The remaining 386 CPF sequences are then filtered further by considering the full set of models and evaluating the strength of their hits against the sequences (see Methods). This testing is intended to discard sequences that end with just a fragment of the domain (the corresponding hits are expected to be very weak and this concerns 79 sequences; see yellow box in **Fig. S1**). For all remaining 307 CPF sequences, we extract the three models in $\mathcal{M}_{\mathrm{FAD}}$ that best match the sequence and make the union of all of them. Many CPF sequences are best matched by the same models and the final set is comprised of 240 models that best identify the presence of the FAD binding domain in CPF sequences.

The third step contains the central idea of the ProfileView method: each CCM is matched to each sequence to be classified and the scores of the hits (see columns of real numbers in **Fig. 1B**) will provide a description of how close the model is to each sequence. In its turn, a sequence can be represented by how close all models are to it through a vector of scores (see rows of real numbers in **Fig. 1B**). In this way, we define a representation space of sequences that does not reflect sequence similarity but, instead, the closeness of each sequence to each model. Since a match of a model is evaluated by two scores (see Methods), the space will be a $2n$-dimensional space (where n is the reduced number of models) and each sequence will be a point in the space.

The fourth step classifies the set of protein sequences in the $2n$-dimensional space. For the generation of our ProfileView tree, we use a hierarchical clustering algorithm which allows to build a tree that groups together the 307 sequences. We further associate several representative models to subtrees of the ProfileView tree and extract from them functional motifs that are specific of the sequences within each subtree.

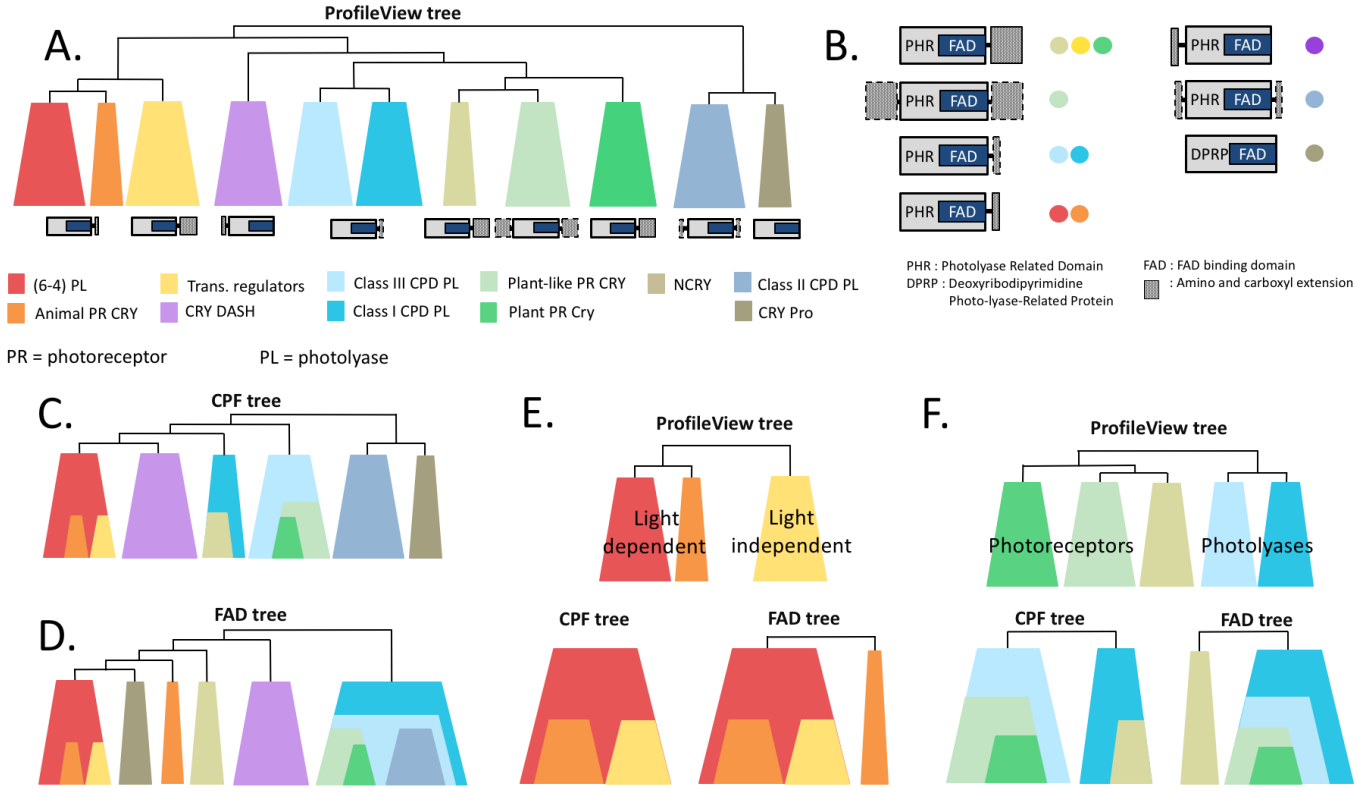


Figure 2. Topological comparison between the ProfileView tree and the canonical distance trees for the Cryptochrome/Photolyase Family and the FAD binding domain. A. Schema illustrating the main topological structure of the CPF tree constructed from the 307 FAD-binding domain sequences. Colors correspond to groups of sequences clustering together and comprising sequences with known function (bottom). The domain architectures known to be characteristic of each subtree is reported (see B for more details). **B.** Domain architectures for proteins belonging to different subtrees of A are reported (colours as in A). C- and N-terminal regions are indicated with blue boxes. Dotted border lines indicate terminal regions only present occasionally in an architecture. **C.** Scheme of the main topological structure of the CPF distance tree constructed from the 307 CPF sequences containing the FAD binding domain. Colors as in A. See the CPF distance tree in **Fig. S3**. **D.** Scheme of the main topological structure of the FAD distance tree constructed from the 307 FAD-binding domain sequences. Colors as in A. See the FAD distance tree in **Fig. S4**. **E, F.** Main topological differences between trees in A, B, C. Two zooms on subtrees involving corresponding classes of CPF sequences for the ProfileView tree in A, the CPF distance tree in C and the FAD distance tree in D. Colors as in A.

2.2 Multiple profiles organise sequences by function

We computed the ProfileView tree, the canonical distance tree for CPF sequences (CPF tree, for short) and the canonical distance tree for FAD sequences (corresponding to the FAD binding domain exclusively, see Methods; FAD tree). A comparison between these trees highlights an important topological reorganisation of major classes (**Fig. 2A**).

The model-based ProfileView tree shows a coherent functional organisation of CPF sequences in 11 different subtrees as illustrated in **Fig. 2A** (and in more detail in **Fig. S2**). Indeed, sequences known to have the same functional characterisation are clustered together in large subtrees of the ProfileView tree and this provides the first evidence of the classification power of the method.

Most importantly, at the root, the ProfileView tree topology organises large subtrees consistently with known functional classes. Namely, the ProfileView tree separates light-independent circadian transcriptional regulator CRYs from the light-dependent (6-4) PLs and animal photoreceptor cryptochromes (PR CRY; (**Fig. 2E**, top)), it reconciles classes I and III CPD PLs into a single subtree, while keeping them distinct, and it clearly separates them from plant and plant-like PR CRYs (**Fig. 2F**, top). At the best of our knowledge, these sharp separations, in agreement with known functional characterisations, have never been obtained by sequence analysis before.

Most interestingly, the ProfileView tree allowed for the discovery of a new subtree (named NCRY; see **Fig. 2A**) of proteins showing strong sequence divergence and lacking functional characterisation. This finding highlights the potential of the method to reveal the existence of novel functional classes within a protein family.

Note that some of the characterised sequences display double function. As highlighted in **Fig. S2**, their photolyase activity (either CPD or (6-4)) is consistently determined by ProfileView that groups these sequences in the photolyase subtrees.

The ProfileView tree demonstrates that the sequence similarity measure producing the canonical distance trees built from CPF sequences (**Fig. 2C** and **Fig. S3**) and FAD sequences (**Fig. 2D** and **Fig. S4**) respectively, is not adapted to functional classification. The CPF distance tree incorrectly groups sequences exhibiting disparate functions, for instance plant PR CRY and plant-like PR CRY are clustered within class III CPD PL (**Fig. 2F**, bottom). Significantly, our new NCRY subtree appears hidden within class I CPD PLs (**Fig. 2F**, bottom), which once more demonstrates that the distance tree is unfit to provide any useful insight at the functional level, in contrast to our approach. As mentioned before, separation of light-dependent and light-independent proteins have never been reached before by sequence analysis and the CPF distance tree topology in **Fig. 2E** (bottom) witnesses this fact, where animal PR CRY and circadian transcriptional regulators are clustered within (6-4) PL sequences. Furthermore, the compatibility of domain architectures associated to different functional classes of CPF sequences (**Fig. 2B**) is coherent with the ProfileView tree topology (**Fig. 2A** bottom) and much less so with the CPF distance tree. Compare, for instance, the architectures for the

classes plant-like PR CRY, plant PR CRY and NCRY, or those for classes I and III CPD PLs. All members of these classes have a PHR domain in which a specific CPF FAD binding domain is found, but C- and N-ter extensions of variable sequence or length. The architectures for plant-like PR CRY, plant PR CRY and NCRY possess N- or C-ter extensions whereas classes I and III CPD PLs only possess the PHR domain. Classes which are topologically close in the ProfileView tree preserve sequence/length characteristics of C- and N-ter regions and agree with what is expected in contrast to the subtrees of the CPF distance tree.

Similar observations can be highlighted by comparing the ProfileView tree with the FAD tree (**Figs 2EF**, bottom). See trees in **Figs S2, S3, S4** for a more precise tree reconstruction compared to the topologies in **Figs 2ACD**.

The reconstruction of the tree topology highlights three important results for CPF sequences: 1. the resolution in two functional groups of light-independent proteins (transcriptional repressor CRY) and proteins which bind the FAD chromophore and need light for their function (PL and CRY PR); 2. the resolution of classes I and III CPD PL into two distinct sibling subtrees; 3. the prediction of possible novel functions, by the identification of novel groups. We investigated these three CPF subgroups in more detail by merging functional data derived from characterised proteins and by structural modelling.

3 Representative models and conserved motifs

Previous functional and biochemical experimental studies of CPF proteins highlighted a number of important specific sequence positions for each known CPF class. We used them to validate further our ProfileView classification. For each one of the 11 subtrees highlighted by ProfileView, we considered their set of sequences, we identified the probabilistic model that best represents the sequences and we constructed from it a representative motif, based on conserved positions. Then, we tested whether or not positions in the motif match known ones. If this is the case, it means that the ensemble of sequences classified by ProfileView is indeed likely to share the same function.

A representative model for a subtree of the ProfileView tree is a probabilistic model of $\mathcal{M}_{\mathrm{FAD}}$ that separates at least 50% of the subtree sequences from all other sequences in the ProfileView tree. Interestingly, we found representative models associated to several of the internal nodes of the ProfileView tree (**Fig. S2**, where the proportion of sequences supported by a model is indicated), and many models separate subtree sequences sharply (100%). An automatic procedure identified representative models. Interestingly, all "functional" subtrees highlighted (by distinguished color) in **Fig. S2** are characterised by a representative model.

From each representative model, we extracted the set of conserved positions and univocally defined a motif for the corresponding subtree. Motifs associated to functional subtrees are reported in **Figs. S6, S7** with the exception of classes I and III CPD PL, known to share the same function, that we grouped together by considering the representative model of the minimal subtree including both classes. Experimentally validated positions within motifs are listed in

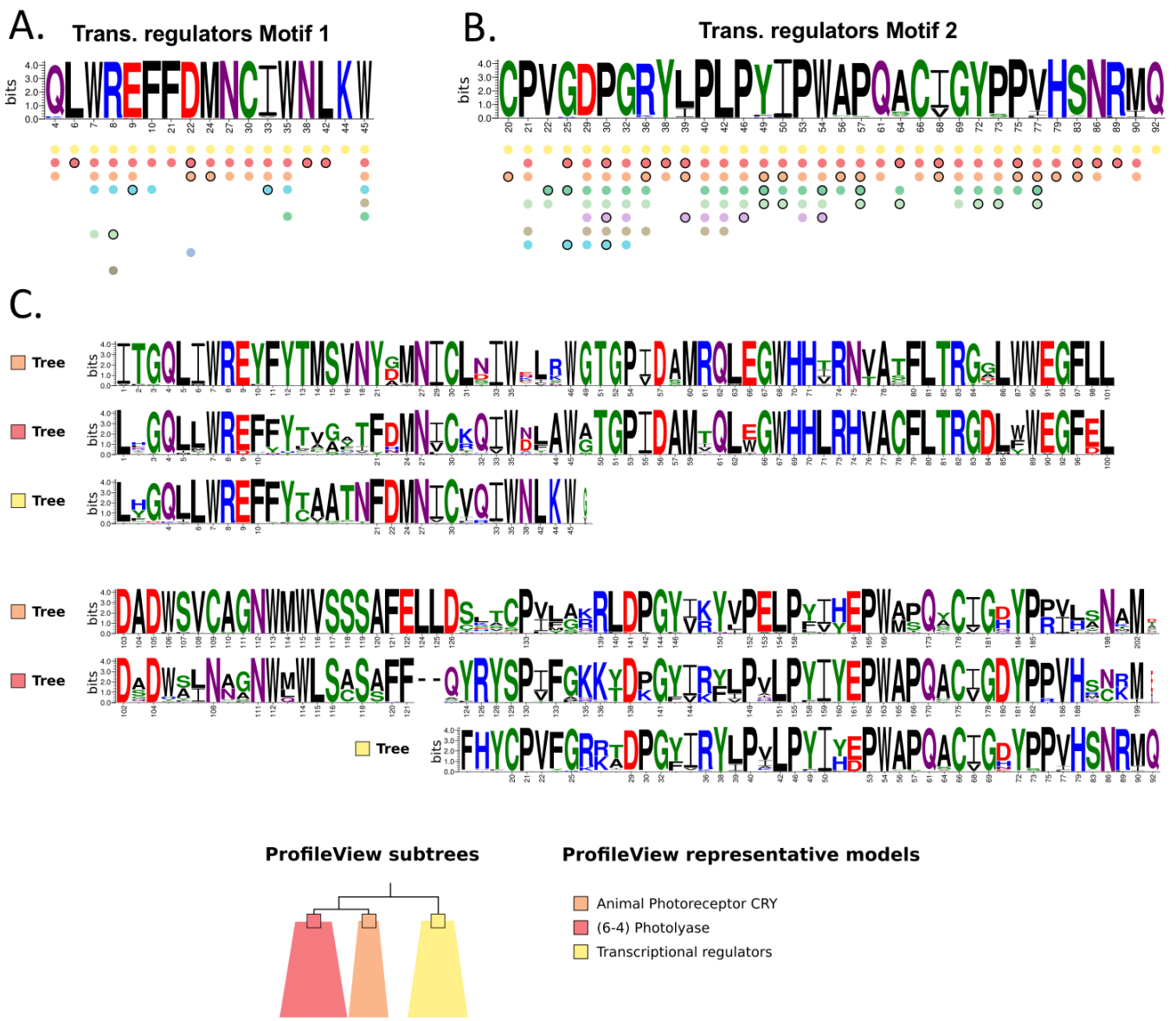


Figure 3. Trans. regulators motifs and their comparison with (6-4) PL and animal PR CRY motifs. A,B: two motifs of conserved residues present in light-independent transcriptional regulator sequences. They are extracted from two representative models of the sequences (described in **Fig. S5**) comprising the "yellow" subtree of **Fig. 2AE** (see also bottom). Numbers (under the letters) correspond to positions in a model, and they are not comparable between motifs. Coloured dots, piled below the motifs, indicate that the corresponding position is well-conserved (see Methods) for the subtrees with the same colour in **Fig. 2A**. Circled dots indicate positions that are less conserved (see Methods). For each motif, coloured dots are ordered, from top to bottom, depending on the best E-values given by `hhblits` to the pairwise alignments. **C**. Three representative motifs associated to the trans. regulators (yellow), (6-4) PL (red) and animal PR CRY (orange) subtrees of the ProfileView tree 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 trans. regulators 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.

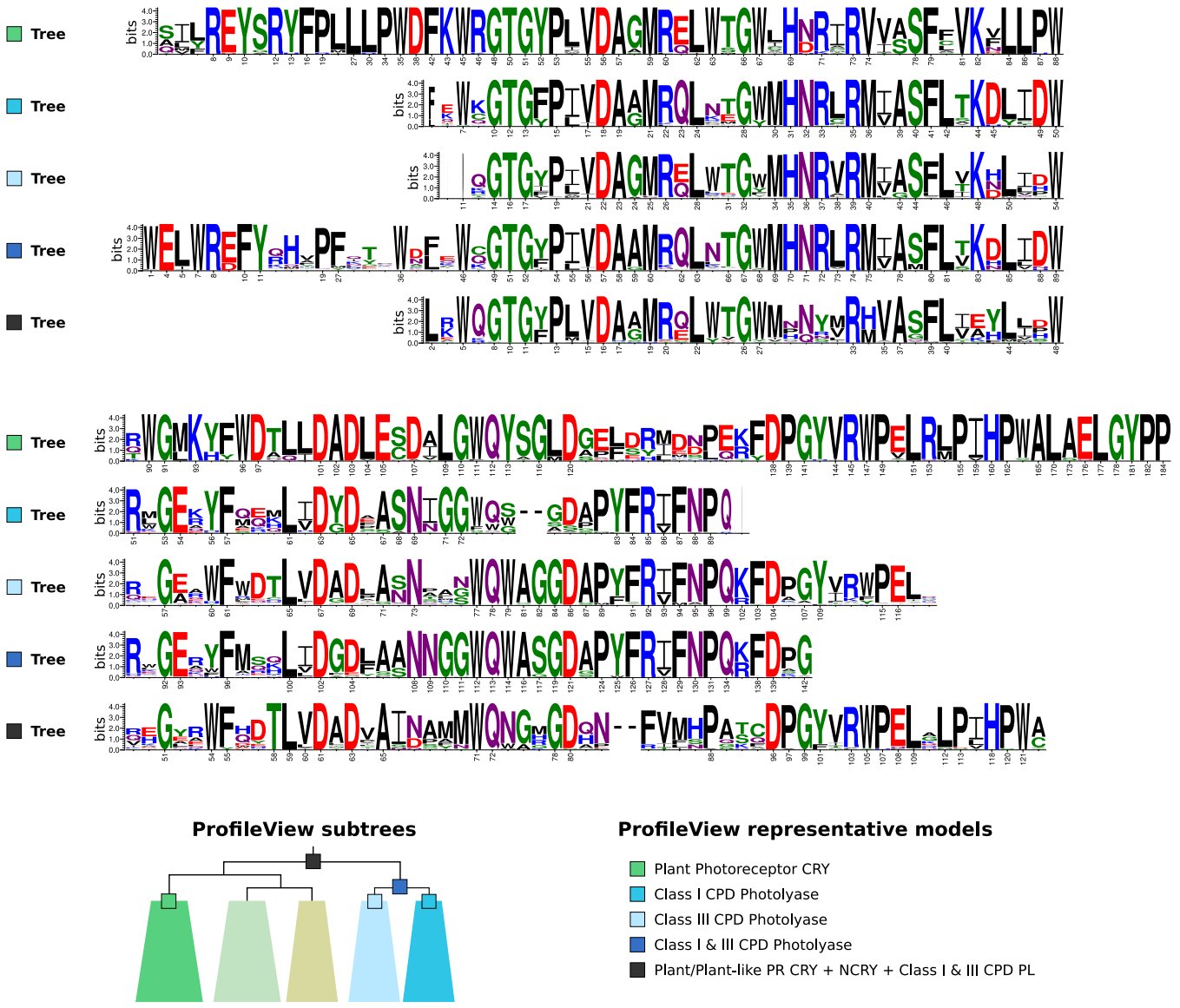


Figure 4. Five motifs for 5 subtrees in the ProfileView tree. Five representative models associated to internal nodes in the ProfileView tree are aligned. Numbered positions correspond to conserved positions belonging to the associated representative motif. The absence of the number indicates less conserved positions. The alignment has been constructed using plant PR as a template model and all others as query models. Neither plant-like PR CRY nor NCRY models were considered because no functionally characterised sequences are known for these models. The NCRY motif (associated to the beige subtree on the bottom) was not added because no functional information is available for comparison. The length of a motif depends on the length of the associated model, selected as best representing the sequences in a subtree.

the **Supplemental File**. The only subtree where we found two distinct representative motifs, covering two different regions of the FAD binding domain sequence, is the light-independent transcriptional regulator tree (**Fig. 3AB**). When comparing with the other models, these two models are the only ones which do not cover the FAD binding domainregion directly involved in proton or electron transfer to the FAD chromophore, as illustrated in **Fig. 3C** with the alignment of the two transcriptional regulator motifs, the (6-4) PL motif and the animal PR CRY motif. This alignment indirectly shows that proton/electron transfer is not involved in the function of light-independent transcriptional repressors (**Fig. 3C**) despite the importance of the FAD chromophore in their regulation [31].

To validate our motifs, we looked at whether their amino acids 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. In the **Supplemental File**, most of the reported 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 PLs or post-translational modifications in CRY, are also identified.

Some positions in a motif might be conserved also in other motifs (corresponding to other subtrees), but some positions are motif specific. An example is motif 1 associated to the light-independent transcriptional regulator sequences (**Fig. 3A** and yellow subtree in **Fig. S2**). As expected, most of the positions in the motif are conserved in other subtrees as well, notably (6-4) PL and animal PR CRY, because of the proximity of these subtrees in the ProfileView tree. Highly conserved positions in most, if not all, models are clearly identified as highly conserved also in the Pfam model (**Fig. S5**). However, four positions (L6, N38, L42 and K44 in **Fig. 3A**) appear to be specific to light-independent transcriptional regulators. Three residues belong to the same helix (α12) and two of these positions (N38, K44) are known to belong to the interaction site with a partner and to the ubiquitination site [32, 33]. The two remaining conserved positions (L6, L42), at the best of our knowledge, have not been identified before and open ways to new investigations. Similar considerations can be drawn on motif 2 associated to the light-independent transcriptional regulator sequences (**Fig. 3B**).

The analysis of the ProfileView tree can be performed for different internal nodes by using representative models and their associated motifs. By construction, representative models split sequences in subclasses supported by specific motifs, hence being susceptible to functionally characterise the associated sequences. To illustrate the great deal of information that can be extracted from these models, we considered a group of motifs, involving class I CPD PL, class III CPD PL, NCRY, plant CRY PR and plant-like CRY PR subtrees (see **Figs. 4**). Classes I and III CPD PL and plant CRY PR are well-studied families in terms of function and molecular mechanisms, and present numerous specific mutants leading to a loss of function. Remarkably, by crossing the functional characterisation of specific residues in this collections of mutants (described in **Supplemental File**) with our motifs of classes I and III CPD PL and plant CRY PR, we validate 32 and 25 of their positions, respectively.

By comparing the motif representing classes I and III CPD PL with those representing either class I CPD PL or class III CPD PL, we notice that there is almost no amino acid which is motif-specific among the three models. The strong closeness between motifs of class I and class III agrees with their shared function. We especially notice the conserved amino acids involved in FAD binding or FAD binding pocket such as R74, D102, D104, N108 (directly involved in the proton transfer to the FAD), F129 and Q134, or those involved in CPD lesion binding sites such as W7, N71, M75, W114 (where numbers refer to the motif accounting for both classes I and III CPD PL; see **Supplemental File**). This example demonstrates that the analysis of the different motifs at different nodes might be used to deduce common functions. However, it is possible to extract some differences among the two motifs, where specific amino acids such as W60 (for class III CPD PL) versus Y56 (for class I CPD PL) were suggested to make an alternative electron transfer pathway possibly important in some specific condition [34]. Other interesting differences are W29, D63 and T64 from class III CPD PL which have been identified as interacting with the MTHF in a specific binding site of MTHF from class III CPD PL [34].

By comparing the three CPD PL motifs/models with the one from plant CRY PR, we remark: 1. conserved amino acids involved in FAD binding, 2. a short motif (L104, E105, D107 and L109, where D107 is known to be directly involved in the proton transfer to the FAD) representing the specificity in the binding pocket of plant CRY PR, 3. the absence of the CPD binding sites. Interestingly, at two specific positions of the CPD binding sites (M75, W114), two specific amino acids (V74, Y114) are found in the motif of Plant CRY PR which have been involved in the ATP binding site described up to now as specific for Plant CRY PR [35, 36]. Moreover, despite very conserved amino acids in class I CPD PL model such as D45, D49 or E54 (involved in the proton transfer to the FAD), the latter ones are present but not fully conserved in class III CPD PL and are clearly absent in the plant CRY PR model. This observation suggests that these amino acids might also be involved, directly or indirectly, in the CPD repair function, and that some variability is not expected to disrupt the function.

The analysis of the motif representing plant-like, plant CRY, NCRY and classes I and III CPD PL (**Figs. 4**, bottom) on available protein structures and mutants, clearly highlights amino acids involved in the FAD binding as well as the conserved tryptophan (W) triad involved in electron transfer to the FAD, which is in agreement with available functional data on this family [37]. Moreover, despite the position of the node in the tree leading to the identification of highly conserved amino acids among the whole CPF family, this motif highlights one specific amino acid (V15) and to some extend two other positions (G78 and D80) suggesting that these three amino acids might play a critical role and might be a specific signature for the whole family.

Last, some promising new information can be extracted by the comparative analysis. Indeed, despite many studies on these PL classes, we could identify 2 specific amino acids (F27 and I55) with yet undefined function. When looking at their position in the structure, these amino acids do not seem to be directly involved neither in DNA nor in FAD binding. Nevertheless, they are highly specific suggesting their involvement in the CPD repair mechanism.

All these examples, experimentally validated by the genetic and functional analysis of selected mutations, illustrate the strength of ProfileView representative models in extracting important amino acids information from sequences that can be used to design tailored experiments for discovering new functional activities or novel biological mechanisms involving the FAD binding domain.

4 Structural analysis of NCRY as potential photoreceptor sequences

ProfileView allowed to identify a new functional subtree of previously uncharacterised sequenced, which we called NCRY. As noticed, it is positioned close to the Plant PR CRY and plant-like PR CRY, suggesting it to be a photoreceptor. In contrast, the CPF tree includes NCRY within class I CPD PL and, the FAD tree distinguishes the NCRY as a separate subtree placing it close to the animal PR CRY and CRY DASH. To our knowledge only one protein from this family has been characterised and it was shown to bind FAD but to lack photolyase activity [36]. The architecture of the NCRY protein with the presence of a C-ter extension which might prevent DNA binding as well as the absence of photolyase activity are fully in accordance with the position of this family in our functional tree.

An *in silico* structural model of the *Phaeodactylum tricornutum* NCRY sequence, called *Pt*NCRY, shows a unique FAD active site where the classical tryptophan triad, involved in the photoreduction of most CPF members, is replaced by an atypical chain of two tyrosines and one tryptophan (orange in **Fig. 5C**). In addition, a histidine is found to face the flavin N5 position (usually protonated after photoreduction). The NCRY motif obtained from the representative model characterising the set of NCRY sequences, is represented in the structural model of **Fig. 5D** (orange residues), where well-conserved positions are localised around the FAD binding site. In particular, the set of conserved positions (from 104 to 120) that are specific to the NCRY motif describe the FAD binding site (**Fig. 5D**, green) highlighting substrate binding. These structural characters support the existence of a novel functional class within the CPF, where the atypical electron and proton pathways are expected to produce quite unusual photochemical properties.

5 WW domains and their functional classification

WW domains are protein modules mediating protein-protein interactions through recognition of proline-rich peptide motifs and phosphorylated serine/threonine-proline sites. They are involved in a number of different cellular functions [29] such as transcription, RNA processing, receptor signalling and protein trafficking. WW domains have been experimentally classified in six interaction groups by Otte *et al* [27] (Y, R_a, R_b, L, poly-P, poS/poT), in 3 groups by Ingham *et al* [29] (A, B and C) and in 6 groups by Russ *et al* [28] (I, Im, I/IV, II, III, IV). Starting

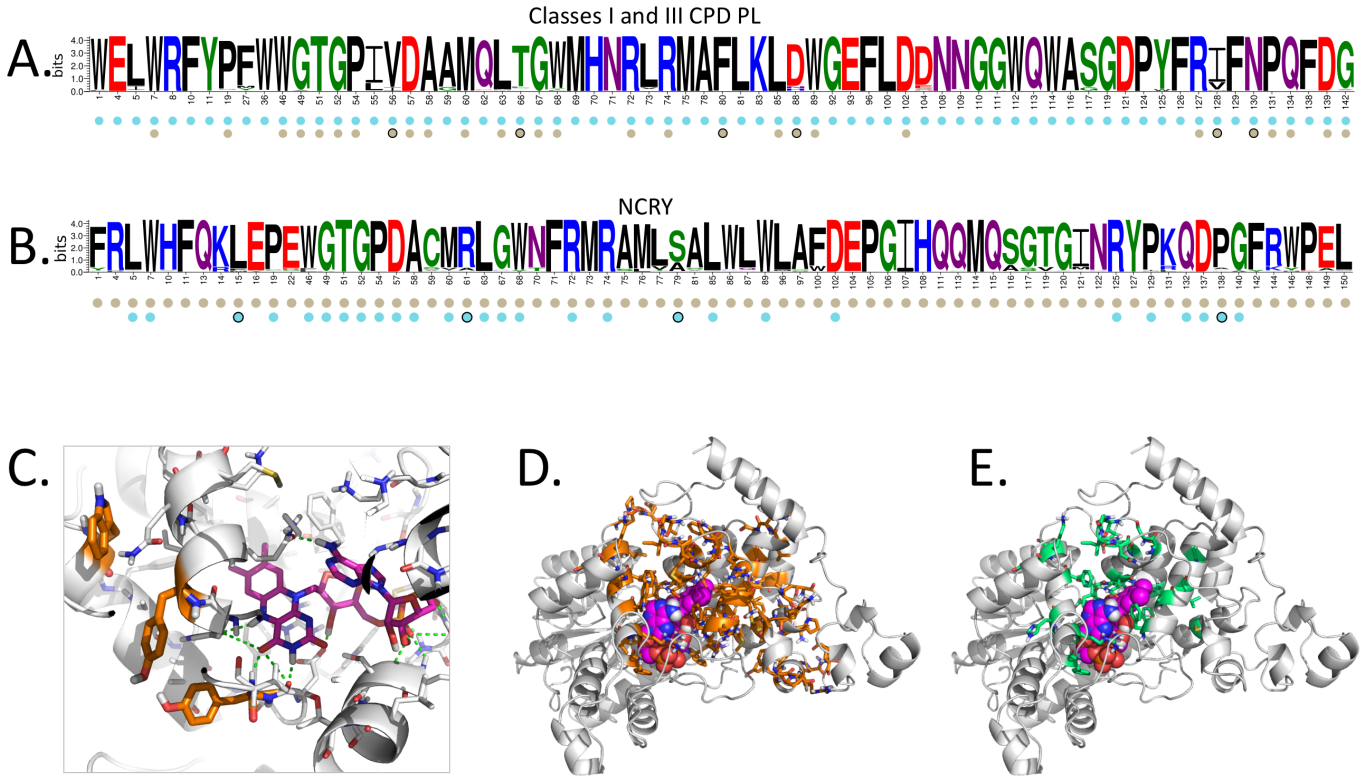


Figure 5. Motif for the NCRY sequences and structural modelling. A. Motif resulting from the representative model of class I CPD PL sequences (see **Figure 2**). The positions of this model that are also conserved in the model representative of the NCRY sequences in B are indicated with beige bullets, below the motif. **B**. Motif resulting from the representative model of NCRY sequences (see **Figure 2**). The positions of this model that are also conserved in the model representative of the NCRY sequences in B are indicated with beige bullets, below the motif. **C.** Homology model of the FAD binding site of *Pt*NCRY. The template we used is PDB: 6fn3. An atypical chain of two tyrosines and one tryptophan, possibly involved in photoreduction, is coloured orange. In addition, a histidine is found to face the flavin N5 position (usually protonated after photoreduction). H-bonds are indicated in green. The FAD cofactor is pink. **D.** Homology model of the FAD *Pt*NCRY structure where all conserved residues in the NCRY motif unique to the motif (all beige dots) are highlighted. **E.** Homology model of the FAD *Pt*NCRY structure where all conserved residues in the NCRY motif that are NCRY specific (beige dots that are not cyan) are highlighted.

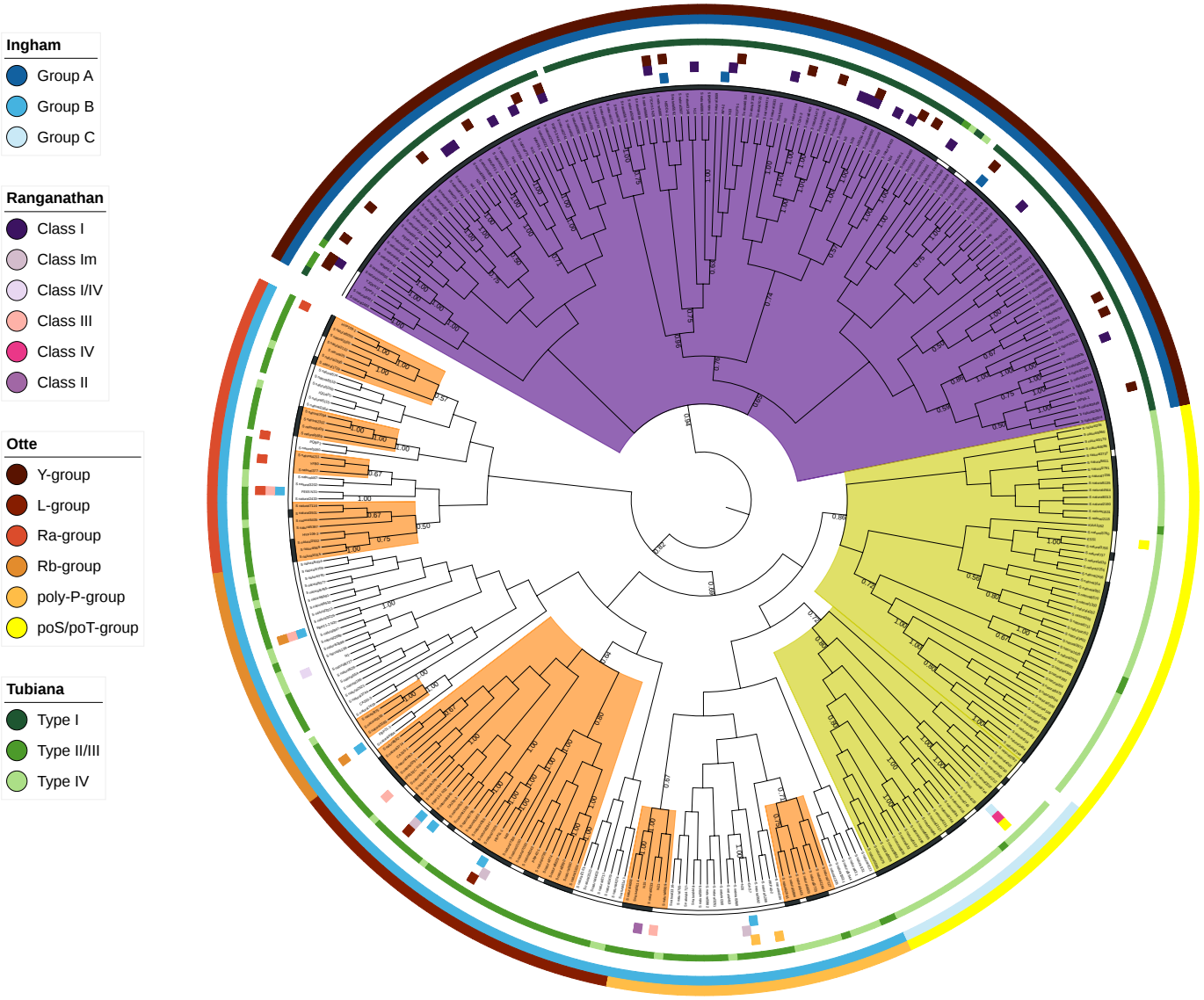


Figure 6. ProfileView tree of WW domains. The tree is constructed from natural sequences studied in [27–29, 38]. All subtrees with a representative model are indicated by a root labeled by the percentage of sequences in the subtree best matching the model (see Methods). Among these subtrees, those containing at least 3 sequences are coloured. Sequences in a coloured subtree that are matched by the associated representative model are highlighted in black, in the first circular stripe surrounding the tree. The functional group/class associated to natural sequences experimentally tested in "Otte" [27] (brown scale, see legend on the left), "Ingham" [29] (blue scale) and "Ranganathan" [28] (purple scale) are reported in small squares on the three subsequent circular stripes. "Tubiana" computational classification [38] is also reported (green scale). The two most external circular stripes report the compatibility between the grouping suggested by Otte *et al.* and Ingham *et al.* with ProfileView subtrees. The larger subtree comprising a given function, in the sense of Otte *et al.* or Ingham *et al.*, is assigned the colour of the function. Note that one of our trees is not functionally annotated by experimental data coming from Ingham *et al.*

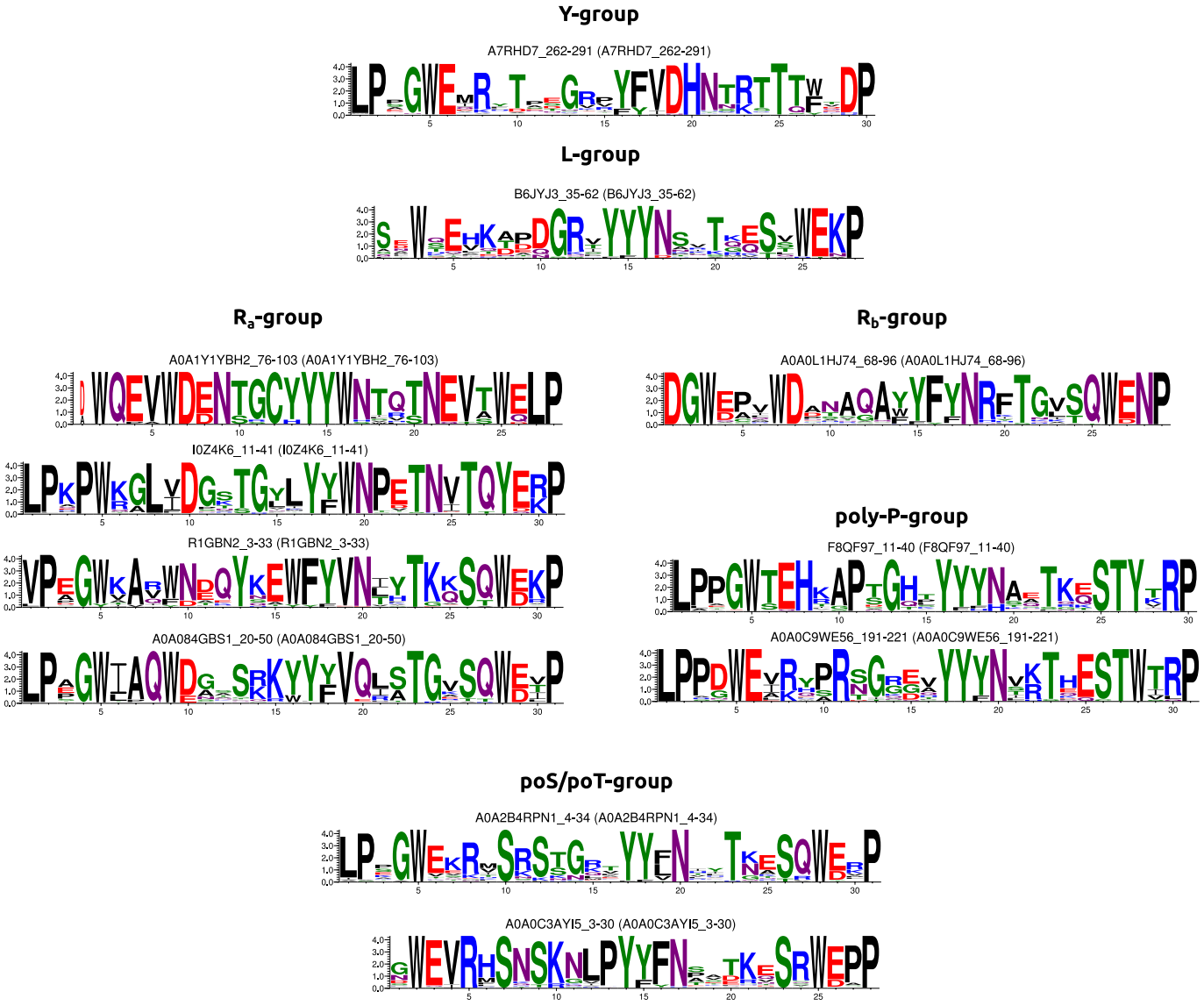


Figure 7. Representative models in ProfileView tree of WW domains. Models are representative of the sequences organised in the colored subtrees of Figure 6. For Otte's classes containing more than one representative model, the order, from top to bottom, corresponds to subtrees read anticlockwise in the outer circle (brown scale) of Figure 6, corresponding to Otte *et al* classification.

from sequence data, a computational classification in 3 groups (I, II/III, IV) describing protein binding motifs has been proposed by Tubiana *et al* [38].

All natural sequences analysed in [27–29] and upgraded with a set of natural sequences randomly selected from [38], have been considered for classification by ProfileView. ProfileView tree highlights a subtree (coloured purple in **Fig. 6**) that agrees with the four classifications above. It organises the remaining sequences in several subtrees corresponding to the remaining five groups proposed in Otte *et al* [27]. Namely, all R_a-group WW domains experimentally identified in [27] appear in the same subtree, for instance, and the same is true for all groups in Otte's classification. The ProfileView tree is also compatible with the experimental data organised in the three groups by Ingham *et al* [29], where group A corresponds to Y in Otte's classification, and C to the union of L, R_a, R_b, poly-P sequences. Due to missing experimental data in Ingham *et al* experiments, B consistently matches a part of poS/poT (see the two most external circles in **Fig. 6**). The comparison with the three groups identified computationally in Tubiana *et al* [38] is less sharp even though this grouping follows the same tendency as in Ingham *et al* (**Fig. 6**). In contrast, our computational approach has the advantage of organizing sequences in the large group II/III, indistinguishable in Tubiana *et al*, into subtrees of sequences known to bind to specific peptides, as shown in [27], providing a refined analysis of binding motifs. ProfileView tree also classifies, within its subtrees, many experimentally uncharacterized WW domain sequences, largely agreeing with Tubiana's classification but not always, as seen in **Fig. 6**.

As shown in **Fig. 6**, not all Otte's WW domain classes in the tree are associated to a unique representative model (see legend). Multiple models might be associated to the same Otte's class and describe different groups of sequences within the class. All identified representative models are reported in **Fig. 7** and their associated motifs in **Fig. S8**.

6 Discussion

The availability of large quantities of (meta)genomic data is allowing for a deeper exploration of living organisms and of the processes underpinning their genetic, phylogenetic and functional diversification. Computational approaches, able to highlight these diversities and to identify what is functionally new within the realm of sequence information, will make the first fundamental step in the discovery of new candidates to be experimentally tested for their functional activity. Moreover, due to the huge quantity of sequences to be acquired in years to come (1 zetta-bases/year are expected in 2025 [39]), there will be no more way to look into this mass of data with an "expert eye" and computational approaches will play a key role on the extraction of novel information and in functional classification.

Today, we can characterise sequences based on their similarity through distance measures modelling the evolution of entire sequences. However, as shown for CPF sequences, this computational approach is insufficient to provide insights on protein functional activities, and a large number of CPF sequences remain not yet functionally annotated. This protein family is

extremely important in medicine, biology, environment and biotechnology due to its key roles in cancer biology and DNA repair, drug delivery strategies, global circadian rhythms, optogenetics. Thanks to its key role, for decades now, experiments have accumulated a huge amount of functional information that we used to validate the ProfileView approach. Similar considerations hold for the WW domain family, of crucial importance for understanding the formation of eukaryotic protein-protein interaction networks. ProfileView functional organisation of these two families perfectly agrees with what is known.

By constructing multiple probabilistic profiles characterising different conserved motifs in CPF/WW domain homologous sequences, ProfileView captures functional signals and, by combining them, is able to successfully classify large datasets. It obtains a striking improvement compared to previous attempts. The main advantages of ProfileView approach compared to those developed before are as follows: (i) it is alignment-free and avoids errors due to the difficulty of comparing distant homologues; (ii) several probabilistic models represent more precisely than a single consensus models the functional variability of protein families; (iii) large quantities of data are not needed to learn features and run the classification; (iv) functional annotation of many sequences does not need to be known to explore with precision the space of sequences and classify them.

ProfileView uncovered a new family of photoactive proteins in CPF and provided new candidates for functional and photochemical characterisation. This study realises the preliminary step in the discovery of potentially interesting proteins whose function could be experimentally tested with the purpose of enlarging our understanding of the mechanisms exploiting light to perform functional activities in natural environments. These proteins are of interest for biotechnology and any computational approach to highlight them is desired.

ProfileView organised 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 demonstrates that a large variety of sequence motifs covers this spectrum and it identifies these motifs. Compared to Tubiana *et al* [38], a computational approach also based on sequence analysis, it describes differences among binding motifs in much greater detail, opening new avenues in the discovery of alternative binding patterns in protein-protein interaction networks.

On the methodological side, ProfileView addresses the problem of extracting biological information on protein families from the huge space of natural sequences. Starting from a set of homologous sequences, it samples hundreds of them ensuring their pairwise low sequence identity, and for each selected sequence, it samples its close neighbourhood and uses it to construct a multiple probabilistic model aimed to represent the space (**Fig. 1A**). Note that classical approaches construct a unique model as representative of the entire space, with obvious limitations when the space is characterised by many divergent homologs (see the model degeneracy for the FAD binding domain in **Fig. S5**). Sampling of distant sequences could be realised using different distance measures. This is an important direction of investigation possibly leading to more refined biological information extracted from sequences.

The resulting set of probabilistic models is employed by ProfileView in a single-layer architecture to transform sequences in sequence space into vectors of reals in the representation space to be classified (**Fig. 1B**). A possible direction of investigation is the design of multiple layers for an architecture that analyses finer motifs as well as proteins comprising multiple domains.

The fine understanding of functional mechanisms for CPF classes might need more sophisticated computational approaches than ProfileView. Indeed, 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 **Fig. S2**), the entire CPF sequence is likely to be necessary, possibly because of the interaction between domains which might have functional consequences as highlighted in [40]. In this respect, notice that the 79 sequences discarded because of the absence of a sufficiently strong signal detecting the FAD binding domain, might be retrieved and classified on a larger scale analysis, involving more sequences and more models.

Overall, ProfileView highlights that protein functional classification depends on a non-linear contribution of many probabilistic 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 is wide open today despite the clear interest in classifying protein families that have been amply studied in molecular biology, like transporters, signalling, transcription factors.

ProfileView computational method is general and applicable to contexts other than the CPF as we demonstrated with the analysis of the complex WW domain family. It is computationally efficient in screening millions of sequences in a reasonable time. When crossed with the CLADE and MetaCLADE methods, it can be applied to a very broad ensemble of sequences found to be homologous to a given domain. It can classify these sequences and discover which ones behave differently from known ones. It can provide a description of the putative functional motifs of a given domain and highlight mutational tests one might want to realise experimentally.

The majority of metagenomics and metatranscriptomics data come from organisms that cannot be cultured and that will, possibly, never be isolated. Hence, conceptual new approaches to explore their biology in complex ecosystems is desperately needed. ProfileView allows to increase knowledge on the biology of organisms whose ecological role has been recognised (*e.g.* marine phytoplankton) but that are still not accessible to functional investigations, opening a new avenue to functional exploration.

7 Methods

For clarity, this section will preferentially borrow examples from the CPF application. Information relevant to the WW domains analysis will be provided for each step of the construction.

7.1 Clade-centered models and the ProfileView space

Widely used search methods [41–43] are based on a mono-source annotation strategy, where a single probabilistic model (*e.g.*, a pHMM [30]), generated from the consensus of a set of homologous sequences, is used to represent a protein domain. The mono-source strategy usually performs well for rather conserved homologous sequences, but when sequences have highly diverged, consensus signals become too weak to generate a useful probabilistic representation and global-consensus models do not characterize domain features properly. A *multi-source* domain annotation strategy [4], in which protein domains are represented by several probabilistic models, called *Clade-Centered Models* (CCM), was implemented in CLADE [4] and MetaCLADE [5] for genomes and metagenomes/metatranscriptomes respectively.

To construct CCMs, we considered the *full* set of sequences S^i associated to a domain D^i of Pfam [44] and, for some representative sequences $s_j \in S^i$, we constructed a *clade-centered* profile HMM, in short CCM, by retrieving a set of homologous sequences close to s_j. Such a model displays features characteristic of s_j and that might differ from other domain sequences $s_k \in S^i$. The rationale is that the more s_j and s_k are divergent, the more clade-centered models are expected to highlight different features. It has been shown that CCMs significantly improve domain annotation (either for full genomes [4] or for metagenomic/metatranscriptomic datasets with MetaCLADE [5]) and, due to their closeness to actual protein sequences, they are more specific and functionally predictive than the canonical global-consensus approach. In this work, however, we build and use CCMs differently aiming at better resolve the functional organisation of sequences within a protein family (the cryptochrome/photolyase family).

ProfileView consists of four main stages (**Fig. S1**):

1. the construction of a CCM library (*e.g.* for the FAD binding domain or the WW domain);
2. a sequence selection based on matching or non matching of CCMs to a sequence (according to a posterior phylogenetic analysis, very divergent sequences are discarded) and a sequence filtering evaluating the importance of a match;
3. a model filtering identifying a set of models that best match the selected sequences and the construction of a ProfileView multi-dimensional space of sequences;
4. the construction of the ProfileView tree (a hierarchical representation of the sequences) and the identification of best representative models/motifs for functional clades.

7.1.1 CPF protein sequences dataset

The set $\mathcal{S}_{\text{CPF}}$ of CPF protein sequences was retrieved from publicly available databases such as UniProt, JGI projects (`genome.jgi.doe.gov`), and OIST projects (`marinegenomics.oist.jp`). It is comprised of 397 sequences, and was manually constructed following two main criteria: 1. it contains 69 CPF sequences known to have a specific function according to experimental evidence reported in the literature (see **Supplemental File** for bibliographical references); 2. it contains CPF sequences that span the whole tree of life; they belong to 146 species, 74 classes, and 40

phyla (see **Supplemental File** for the detailed list). The distance tree constructed from the 307 CPF sequences is reported in **Fig. S3**.

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, while a "FAD sequence" refers to the FAD binding domain sequence exclusively.

7.1.2 WW domain sequences dataset

The set of WW domain sequences was constructed by combining the datasets of natural sequences analysed in [27–29, 38]. It is comprised of 349 sequences, 60 of which have been experimentally classified [27–29] and the remaining ones have been randomly selected, in comparable proportion, from the three sets classified in Tubiana *et al* (types I, II/III, IV) [38].

7.1.3 Model library construction

To construct a library of models for a domain, we considered sequences from the Pfam v31 [44] database and, for each sequence, we built a CCM [4] by searching in UniProt (version `2017_10`) for highly significant matches of homologous sequences having at least 60% identity with the query domain sequence and covering at least 70% of it. More precisely, a multiple sequence alignment is built using the command `hhblits` of the HH-suite [43] (with parameters `-qid 60 -cov 70 -id 98 -e 1e-10` and database `uniclust30_2017_10`) and subsequently converted into a pHMM with HMMER [30] in order to perform a sequence-profile comparison. Moreover, a pHMM is considered only if it is trained with a minimum number of 20 sequences.

The set of Pfam sequences used to construct the ProfileView's model library for CPF is mainly different from the set of classified sequences: among the 240 models taken into account for the classification of the 307 sequences, just 17 of these models were built from a (Pfam) sequence in $\mathcal{S}_{\mathrm{CPF}}$ and only one is a representative model (for the (6-4) PL subtree) in the ProfileView tree. Moreover, the average identity and similarity (based on pairwise alignments) between the set of 307 sequences to classify and the set of the 240 sequences generating the models are 26.35% and 36.73%, respectively.

In order to capture conserved motifs likely to be of functional relevance for domain sequences, we built highly specific clade-centered models. Indeed, we are not concerned with improving domain annotation of divergent sequences [4] but, instead, we aim at classifying proteins sharing their domain architecture from the perspective of their function and expect their functional motifs to be represented by positions in the sequence that remain conserved on representative subsets of homologs.

In the analysis of CPF sequences, we focused on the *FAD binding domain of DNA photolyase* from Pfam version 31 (accession number `PF03441`), due to its functional importance for CPF activity. More in detail, we selected all 4615 sequences which belong to the FULL alignment in Pfam. For each one of them, a CCM has been constructed with the command mentioned above. Finally, our model library $\mathcal{M}_{\mathrm{FAD}}$ for the FAD-binding domain comprises 3735 CCMs, because

for some sequences we could not collect a minimum of 20 homologs. The pipeline for $\mathcal{M}_{\text{FAD}}$ construction is depicted in **Fig. 1**, top.

For the analysis of WW domains, we constructed models from the set of sequences in Pfam version 32 (accession number `PF00397`). They have been extracted from the reference proteome RP15 comprising 5634 sequences. We could construct 3733 models based on at least 20 homologs.

7.1.4 Sequence selection

After building $\mathcal{M}_{\text{FAD}}$, we discarded from $\mathcal{S}_{\text{CPF}}$ all sequences against which we were not able to find any domain hit (independently of the hit score). $\mathcal{S}_{\text{CPF}}$ domain annotation was carried out by considering HMMER best hits (version 3.1b2) for models in $\mathcal{M}_{\text{FAD}}$. An a posteriori phylogenetic analysis of the original set of CPF sequences has been carried out with RAxML version 8.2.11 (with parameter `-m PROTGAMMAAUTO`). We observed that the set of discarded sequences, presenting no FAD binding domain match, correspond to long branches in the tree (see red labeled sequences in **Fig. 1**). This preliminary filter led us to consider a set of 386 CPF sequences over the 397 we started with.

For the analysis of WW domains, the sequence selection step had no effect.

7.1.5 Sequence filtering

In this phase of the pipeline, each CCM belonging to the model library $\mathcal{M}_{\text{FAD}}$ is mapped against the set $\mathcal{S}_{\text{CPF}}$ of all input sequences using HMMER. Let $\mathcal{H} = \{h_{s,m} \mid s \in \mathcal{S}_{\text{CPF}},\ m \in \mathcal{M}_{\text{FAD}},\ score(h_{s,m}) > 0\}$ be the set of hits $h_{s,m}$ provided by `hmmsearch`, where s is a sequence of $\mathcal{S}_{\text{CPF}}$, m is a model of $\mathcal{M}_{\text{FAD}}$ 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,

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

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 [45].

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 concerning 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 **Fig. 1** for an illustration of the three matching types.) More formally, given a hit $h_{s,m} \in \mathcal{H}$, it belongs to $Full(\mathcal{H})$ if the aligned region of m to s (excluding gaps) is 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, $h_{s,m} \in Partial(\mathcal{H})$.

The definition of such a partition is carried out to keep only sequences that are not potentially incomplete (*e.g.*, due to assembly errors). More precisely, a sequence s is retained if any of the

following two conditions holds:

i. at most the 30% of its hits belong to $Overlap(\mathcal{H})$;

ii. at least the 50% of its hits belong to either $Full(\mathcal{H})$ or $Partial(\mathcal{H})$.

These two conditions have been introduced in order to take into account the fact that Pfam might also contain partial sequences that could lead to the construction of very short models (that could be fully aligned in potentially incomplete sequences). For the CPF analysis, we remained with 307 sequences corresponding to the leaves of the ProfileView tree.

Moreover, in order to restrict our analysis to a set of representative models, we kept in $\mathcal{M}_{\text{FAD}}$ only those models that achieve one of the k best scores for at least one sequence of $\mathcal{S}_{\text{CPF}}$ (setting $k = 3$). The rationale of this model filtering is to get rid of "noisy" models and, at the same time, significantly reduce the size of $\mathcal{M}_{\text{FAD}}$. From 3 735 CCMs, the number of models reduces to 240 and we refer to this reduced set $\mathcal{M}^*_{\text{FAD}}$.

Finally, let L_s be the length of the region in s aligned to m. For each hit $h_{s,m}$ we define the following two scores:

- a normalized bit-score $ns(h_{s,m}) = \frac{score(h_{s,m})}{L_s}$;
- a normalized weighted bit-score $nws(h_{s,m}) = \frac{Wscore(h_{s,m})}{L_s}$, where $Wscore(h_{s,m})$ is the sum of bit-scores over the positions in the sequence-profile alignment where the bit-score is greater than 3 (that is, the positions where m and s strongly agree). More formally, let $\sigma(s_i, m_j) = \log_2 \frac{e(s_i,m_j)}{bg(s_i)}$ 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 [30]. 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 $Wscore(h_{s,m}) = \sum_{z=1|\sigma(s_{i_z},m_{j_z})\geq 3}^{K} \sigma(s_{i_z}, m_{j_z})$.

They will be used to construct the ProfileView space of sequences.

For the analysis of WW domains, three sequences have been filtered out using the criteria described above and a total of 346 sequences were retained for classification. To reduce the number of models in the library, we selected the 5 best matching models for each of the 346 sequences and obtained a total of 1244 CCMs considered for the construction of the ProfileView space of WW domain sequences. The number of best matching models was increased from 3 (value used for the CPF analysis) to 5 to consider more WW domain models, that is 1244 versus the 845 obtained with k=3. The idea being that when the dataset of sequences to be classified is very diversified, as in the case of the WW domain family, the number of models should be large (> 1000) to explain diversity, eventually through non linear effects captured by multiple conserved motifs defining the ProfileView space.

7.1.6 The construction of a ProfileView space of sequences

For each sequence $s \in \mathcal{S}_{CPF}$, we construct a vector v_s, where the dimension of v_s is $2|\mathcal{M}^*_{\text{FAD}}|$ and $|\mathcal{M}^*_{\text{FAD}}|$ is the number of models in $\mathcal{M}^*_{\text{FAD}}$. The vector v_s contains the pairs of values $ns(h_{s,m})$ and $nws(h_{s,m})$, for each $m \in \mathcal{M}^*_{\text{FAD}}$. If a model m does not have a hit on the sequence $s \in \mathcal{S}_{CPF}$, 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 $2|\mathcal{M}^*_{\text{FAD}}|$-dimensional space, where each dimension is associated to either the normalized bit-score or the normalized weighted bit-score for some model $m \in \mathcal{M}^*_{\text{FAD}}$. Each sequence is a point in $\mathcal{PV}$ and its position reflects the proximity of the sequence to CCMs in $\mathcal{M}^*_{\text{FAD}}$.

7.1.7 The ProfileView tree construction

After the construction of the ProfileView space $\mathcal{PV}$ for sequences $s \in \mathcal{S}_{CPF}$, Principal Component Analysis (PCA) is performed in order 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 99% of the variance for the set $\mathcal{S}_{CPF}$. In practice, from 480 dimensions, the reduction produces a space of 37 dimensions. Sequences are thereby clustered in $\mathcal{PV}^*$ using a hierarchical agglomerative strategy. Namely, we considered the euclidean distance between vectors and Ward's minimum variance method for cluster merging. The rationale of this criterion is to select, at each step, the pair of clusters which minimizes the total within-cluster variance after the merging. Starting from all clusters being singletons, this bottom-up algorithm completes in $|\mathcal{S}_{\text{CPF}}| - 1$ 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 every internal node defines a cluster of two or more elements (according to the chosen merge criterion). Moreover, in such a tree, distances/dissimilarities between merged clusters are encoded as edge weights. The ProfileView tree built for the CPF sequences is depicted in **Fig. S2**, where internal colours are identified by representative models and external strips are associated to known functions (according to the literature, see **Supplemental File** for the detailed list of publications).

For the WW domain analysis, we started from a space in 2488 dimensions (equivalent to 2×1244, where 1244 is the number of models), performed PCA analysis and reduced the number of dimensions to 11 by explaining the 80% of the variance. Due to the diversity of the 346 WW domain sequences, we relaxed the constraints on the variance, compared to the CPF analysis, to obtain a few tens of dimensions versus the 206 obtained with a threshold of 99% used for CPF.

Representative models for clustered sequences. To better explore subtrees in the ProfileView tree, potentially associated to known functions, we associated a *representative model* to the sets of sequences labelling 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 denoted $\mathcal{S}_{\text{CPF}} \setminus \mathcal{C}$) in the ProfileView space $\mathcal{PV}$. 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 $\mathcal{S}_{\text{CPF}} \setminus \mathcal{C}$. This has to hold for at least one of the metrics – ns and nws – defining $\mathcal{PV}$ (see section "Sequence filtering"). 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, then we choose the model m that provides the best separation, that is the model maximizing the distance between the centroids of the sets $\mathcal{C}_m^*$ and $\mathcal{S}_{\text{CPF}} \setminus \mathcal{C}$ (again, computed according to the metrics ns and nws). If $\mathcal{C}$ is the set of sequences of a subtree T of the ProfileView tree (which is not the whole tree), then a *representative model m for $\mathcal{C}$ is associated to the root of T* when the following two conditions are satisfied: 1. $\mathcal{C}_m^*$ comprises at least half of the sequences in $\mathcal{C}$ and 2. $\mathcal{C}_m^*$ contains at least a sequence from each one of the child subtrees of T. Note that a node in the ProfileView tree might remain without representative model. **Fig. S2** indicates which nodes of the CPF tree are represented by a model. When a ProfileView outputs a representative model for a node of the tree, it also outputs a list of suboptimal models which cover either the same amount of sequences $|\mathcal{C}_M^*|$ or the 90% of $|\mathcal{C}|$.

Model logos were built using the python package of Weblogo [46] (version 3.7) which allowed us to easily export sequence logos [47].

Motifs identified from representative models. A motif extracted from a representative model is defined to be the set of all amino acids characterizing well-conserved columns (*i.e.* match states) in the sequence alignment associated to the model according to `hhblits`' definition. Namely, 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.

For each motif illustrated in the text, we highlighted, by a coloured "dot", positions in it found to be 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 notice that given two models and a 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 one of the models is considered as a template, the scores assigned to the same position might vary (the confidence values are obtained from the posterior probabilities calculated in the Forward-Backward algorithm of `hhblits`). In particular, `hhblits` is warning that the confidence score for an aligned position depends on the confidence on the alignment of the close by region. As a consequence, certain conserved positions might see their alignment score to decrease because of

the presence of a very variable region in their vicinity, possibly containing gaps. This explains why, for aligned positions of two motifs, we might miss to indicate related positions or we might display different color dots. An illustrative example of missing related positions is made by position 102 in the NCRY motif and position 103 in the plant PR CRY motif. The two motifs clearly diverge within the region just following position 102/103, justifying a difficult model alignment and a low confidence score for 102/103. A second example, illustrating the asymmetricity of the coloured dots, is position 102 in the NCRY motif aligned with position 95 in CRY Pro. While the CRY Pro motif records the coloured dot for a matching with NCRY, this is not true for the NCRY motif. Indeed, 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 together when the NCRY model is taken as the template but with a confidence score dropping at 0.6.

7.1.8 Distance tree construction for CPF and FAD sequences

The multiple sequence alignments of CPF sequences and FAD sequences were computed using MUSCLE version v3.8.31 [48], and were then trimmed using trimAl version 1.4.rev22 [49] with a gap cutoff of 0.01 (*i.e.* columns containing more than 99% of gaps were removed). Then, for each sequence alignment, we selected the best evolutionary model using ProtTest (version 3.4.2) [50]. 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 optimisation 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 [51] 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 distance tree for $\mathcal{S}_{\mathrm{CPF}}$ is reported in **Fig. S3** and contains 307 leaves corresponding to the 307 CPF sequences containing the FAD binding domain. The distance tree for the set of 307 FAD sequences is reported in **Fig. S4**.

Phylogenetic and ProfileView trees have been generated with iTOL [52].

7.1.9 Output files of ProfileView

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

Also, ProfileView provides to the user the possibility to choose a list of representative models to be compared. The first model of this list is considered as a reference model. A first output describes and provides the logo reporting all conserved positions together with a list of coloured dots (possibly circled) obtained after a pairwise comparison of a model in the list with the reference model (see Methods above; see for example **Fig. 3AB**). A second output describes and provides the logo reporting an intermediate representation of the positions in the reference

model, namely reporting all conserved positions in the associated motif and all positions that are not conserved in the reference model but that are conserved in some other model in the list (see for example **Fig. 3C** and **Fig. 4**).

7.1.10 Implementation and software availability

ProfileView has been developed and tested under a UNIX operating system, using Bash, Python, and R scripts. It exploits GNU parallel [53], if available on the system, in order to perform some jobs in parallel. It is implemented in three main parts carrying out the following modules of the pipeline: 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. **A restrained access is momentarily set. Once the article is accepted, all information will be freely accessible.**

7.1.11 Data accessibility

The set of CPF sequences, FAD domain sequences, WW domain sequences, model libraries, distance trees, ProfileView trees are available at `http://www.lcqb.upmc.fr/profileview/`.

Acknowledgments

LabEx CALSIMLAB (public grant ANR-11-LABX-0037-01 constituting a part of the "Investissements d'Avenir" program - reference : 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"). We thank Simona Cocco and Jérôme Tubiana for sending us the dataset of WW sequences used in their study.

References

[1] Ponting CP, Dickens NJ. Genome cartography through domain annotation. Genome biology. 2001;2(7):comment2006–1.

[2] Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications. Briefings in bioinformatics. 2012;13(6):711–727.

[3] De Filippo C, Ramazzotti M, Fontana P, Cavalieri D. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Briefings in bioinformatics. 2012;13(6):696–710.

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

[5] Ugarte A, Vicedomini R, Bernardes J, Carbone A. A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. Microbiome. 2018 Aug;6(1):149. Available from: `https://doi.org/10.1186/s40168-018-0532-2`.

[6] Fortunato AE, Jaubert M, Enomoto G, Bouly JP, Raniello R, Thaler M, et al. Diatom phytochromes reveal the existence of far-red light based sensing in the ocean. The Plant Cell. 2016;p. tpc–00928.

[7] Amato A, Dell'Aquila G, Musacchia F, Annunziata R, Ugarte A, Maillet N, et al. Marine diatoms change their gene expression profile when exposed to microscale turbulence under nutrient replete conditions. Scientific Reports. 2017;7.

[8] Björn LO. Photobiology: The science of light and life. Springer; 2015.

[9] Jaubert M, Bouly JP, d'Alcalà MR, Falciatore A. Light sensing and responses in marine microalgae. Current Opinion in Plant Biology. 2017;37:70–77.

[10] Sancar A. Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. Chemical reviews. 2003;103(6):2203–2238.

[11] Brettel K, Byrdin M. Reaction mechanisms of DNA photolyase. Current opinion in structural biology. 2010;20(6):693–701.

[12] Chaves I, Pokorny R, Byrdin M, Hoang N, Ritz T, Brettel K, et al. The cryptochromes: blue light photoreceptors in plants and animals. Annual review of plant biology. 2011;62:335–364.

[13] Coesel S, Mangogna M, Ishikawa T, Heijde M, Rogato A, Finazzi G, et al. Diatom PtCPF1 is a new cryptochrome/photolyase family member with DNA repair and transcription regulation activity. EMBO reports. 2009;10(6):655–661.

[14] Heijde M, Zabulon G, Corellou F, Ishikawa T, Brazard J, Usman A, et al. Characterization of two members of the cryptochrome/photolyase family from Ostreococcus tauri provides insights into the origin and evolution of cryptochromes. Plant, cell & environment. 2010;33(10):1614–1626.

[15] Franz S, Ignatz E, Wenzel S, Zielosko H, Putu EPGN, Maestre-Reyna M, et al. Structure of the bifunctional cryptochrome aCRY from Chlamydomonas reinhardtii. Nucleic acids research. 2018;46(15):8010–8022.

[16] Fortunato AE, Annunziata R, Jaubert M, Bouly JP, Falciatore A. Dealing with light: the widespread and multitasking cryptochrome/photolyase family in photosynthetic organisms. Journal of plant physiology. 2015;172:42–54.

[17] Essen LO, Franz S, Banerjee A. Structural and evolutionary aspects of algal blue light receptors of the cryptochrome and aureochrome type. Journal of plant physiology. 2017;217:27–37.

[18] Ozkan-Dagliyan I, Chiou YY, Ye R, Hassan BH, Ozturk N, Sancar A. Formation of Arabidopsis cryptochrome 2 photobodies in mammalian nuclei application as an optogenetic DNA damage checkpoint switch. Journal of Biological Chemistry. 2013;288(32):23244–23251.

[19] Liu H, Gomez G, Lin S, Lin S, Lin C. Optogenetic control of transcription in zebrafish. PloS one. 2012;7(11):e50738.

[20] Rodgers CT, Hore PJ. Chemical magnetoreception in birds: the radical pair mechanism. Proceedings of the National Academy of Sciences. 2009;106(2):353–360.

[21] Chaves I, Yagita K, Barnhoorn S, Okamura H, van der Horst GT, Tamanini F. Functional evolution of the photolyase/cryptochrome protein family: importance of the C terminus of mammalian CRY1 for circadian core oscillator performance. Molecular and cellular biology. 2006;26(5):1743–1753.

[22] Lucas-Lledó JI, Lynch M. Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. Molecular biology and evolution. 2009;26(5):1143–1153.

[23] Mei Q, Dvornyk V. Evolutionary history of the photolyase/cryptochrome superfamily in eukaryotes. PloS one. 2015;10(9):e0135940.

[24] Ozturk N. Phylogenetic and functional classification of the photolyase/cryptochrome family. Photochemistry and photobiology. 2017;93(1):104–111.

[25] Czarna A, Berndt A, Singh HR, Grudziecki A, Ladurner AG, Timinszky G, et al. Structures of Drosophila cryptochrome and mouse cryptochrome1 provide insight into circadian function. Cell. 2013;153(6):1394–1405.

[26] Sudol M, Hunter T. NeW wrinkles for an old domain. Cell. 2000;103(7):1001–1004.

[27] Otte L, Wiedemann U, Schlegel B, Pires JR, Beyermann M, Schmieder P, et al. WW domain sequence activity relationships identified using ligand recognition propensities of 42 WW domains. Protein Science. 2003;12(3):491–500.

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

[29] Ingham RJ, Colwill K, Howard C, Dettwiler S, Lim CS, Yu J, et al. WW domains provide a platform for the assembly of multiprotein networks. Molecular and cellular biology. 2005;25(16):7092–7106.

[30] Eddy SR. Profile hidden Markov models. Bioinformatics (Oxford, England). 1998;14(9):755–763.

[31] Hirano A, Braas D, Fu YH, Ptáček LJ. FAD regulates CRYPTOCHROME protein stability and circadian clock in mice. Cell reports. 2017;19(2):255–266.

[32] Hirano A, Yumimoto K, Tsunematsu R, Matsumoto M, Oyama M, Kozuka-Hata H, et al. FBXL21 regulates oscillation of the circadian clock through ubiquitination and stabilization of cryptochromes. Cell. 2013;152(5):1106–1118.

[33] Schmalen I, Reischl S, Wallach T, Klemz R, Grudziecki A, Prabu JR, et al. Interaction of circadian clock proteins CRY1 and PER2 is modulated by zinc binding and disulfide bond formation. Cell. 2014;157(5):1203–1215.

[34] Scheerer P, Zhang F, Kalms J, von Stetten D, Krauß N, Oberpichler I, et al. The class III cyclobutane pyrimidine dimer photolyase structure reveals a new antenna chromophore binding site and alternative photoreduction pathways. Journal of Biological Chemistry. 2015;290(18):11504–11514.

[35] Orth C, Niemann N, Hennig L, Essen LO, Batschauer A. Hyperactivity of the Arabidopsis cryptochrome (cry1) L407F mutant is caused by a structural alteration close to the cry1 ATP-binding site. Journal of Biological Chemistry. 2017;292(31):12906–12920.

[36] Brautigam CA, Smith BS, Ma Z, Palnitkar M, Tomchick DR, Machius M, et al. Structure of the photolyase-like domain of cryptochrome 1 from Arabidopsis thaliana. Proceedings of the National Academy of Sciences. 2004;101(33):12142–12147.

[37] Aubert C, Vos MH, Mathis P, Eker AP, Brettel K. Intraprotein radical transfer during photoactivation of DNA photolyase. Nature. 2000;405(6786):586.

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

[39] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big data: astronomical or genomical? PLoS biology. 2015;13(7):e1002195.

[40] Rosensweig C, Reynolds KA, Gao P, Laothamatas I, Shan Y, Ranganathan R, et al. An evolutionary hotspot defines functional differences between CRYPTOCHROMES. Nature communications. 2018;9(1):1138.

[41] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997 09;25(17):3389–3402.

[42] Eddy SR. Accelerated Profile HMM Searches. PLOS Computational Biology. 2011 10;7(10):1–16. Available from: `https://doi.org/10.1371/journal.pcbi.1002195`.

[43] Remmert M, Biegert A, Hauser A, Sding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 2011 Dec;9:173–. Available from: `https://doi.org/10.1038/nmeth.1818`.

[44] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Research. 2014;42(D1):D222–D230. Available from: `http://dx.doi.org/10.1093/nar/gkt1223`.

[45] Barrett C, Hughey R, Karplus K. Scoring hidden Markov models. Bioinformatics. 1997;13(2):191–199. Available from: http://dx.doi.org/10.1093/bioinformatics/13.2.191.

[46] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004;14(6):1188–1190.

[47] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic acids research. 1990;18(20):6097–6100.

[48] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004 Mar;32(5):1792–1797. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15034147.

[49] Capella-Gutirrez S, Silla-Martnez JM, Gabaldn T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–1973. Available from: http://dx.doi.org/10.1093/bioinformatics/btp348.

[50] Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics (Oxford, England). 2011 Apr;27(8):1164–1165. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21335321.

[51] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010;59(3):307–321. Available from: http://dx.doi.org/10.1093/sysbio/syq010.

[52] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research. 2019;.

[53] Tange O. GNU Parallel 2018. Ole Tange; 2018. Available from: https://doi.org/10.5281/zenodo.1146014.

SUPPLEMENTARY FIGURES

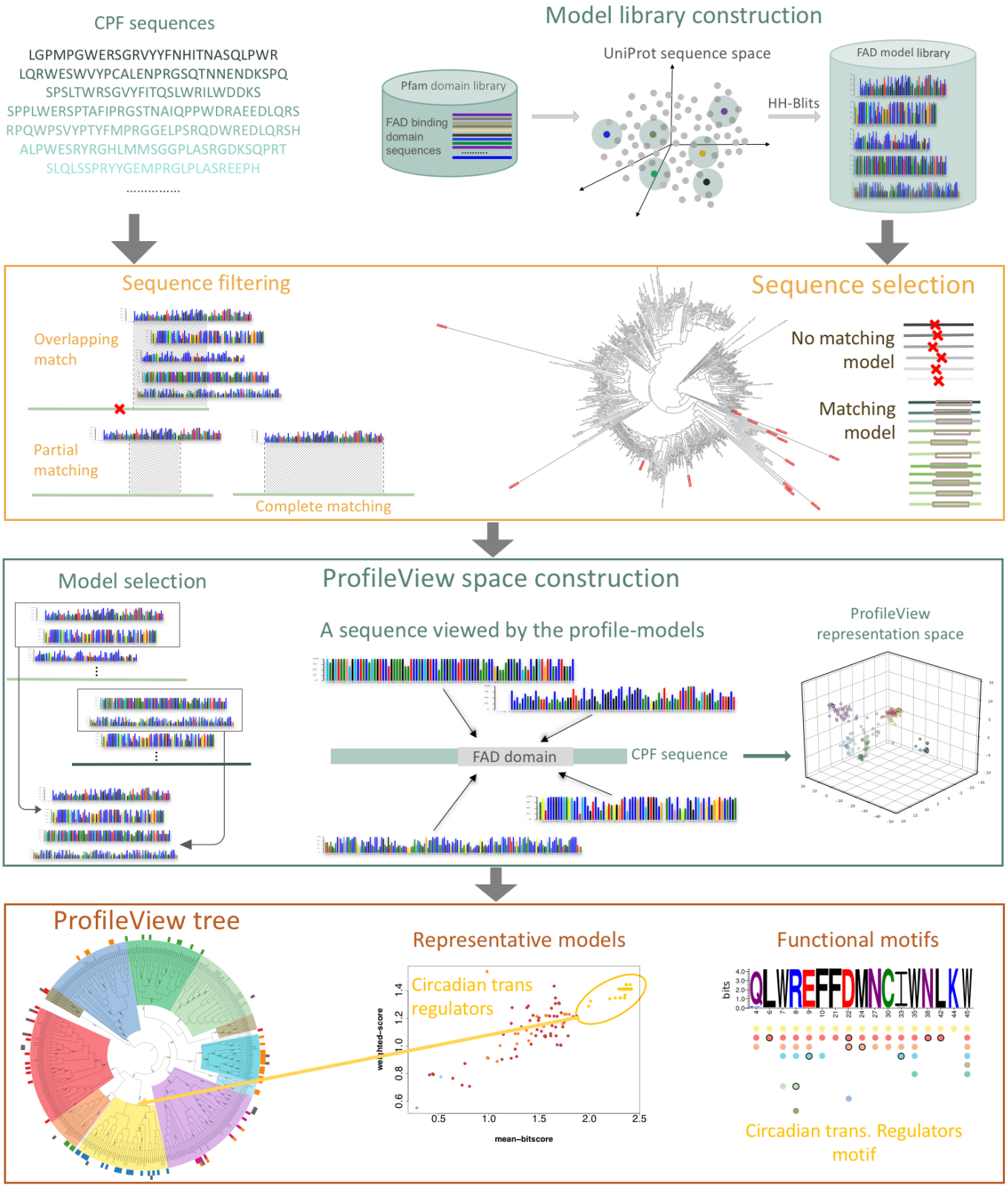


Figure S1. ProfileView pipeline applied to CPF sequences. The ProfileView pipeline is organised in four main steps: (i) the model library construction for the FAD binding domain and the collection of CPF sequences to classify, (ii) a sequence selection based on matching/unmatching of the models to a sequence and on sequence filtering evaluating the importance of a match, (iii) a model filtering step reducing model redundancy and the construction of a ProfileView space of sequences, (iv) the construction of the ProfileView tree and the identification of both the best representative models for functional clades and their characteristic functional motifs.

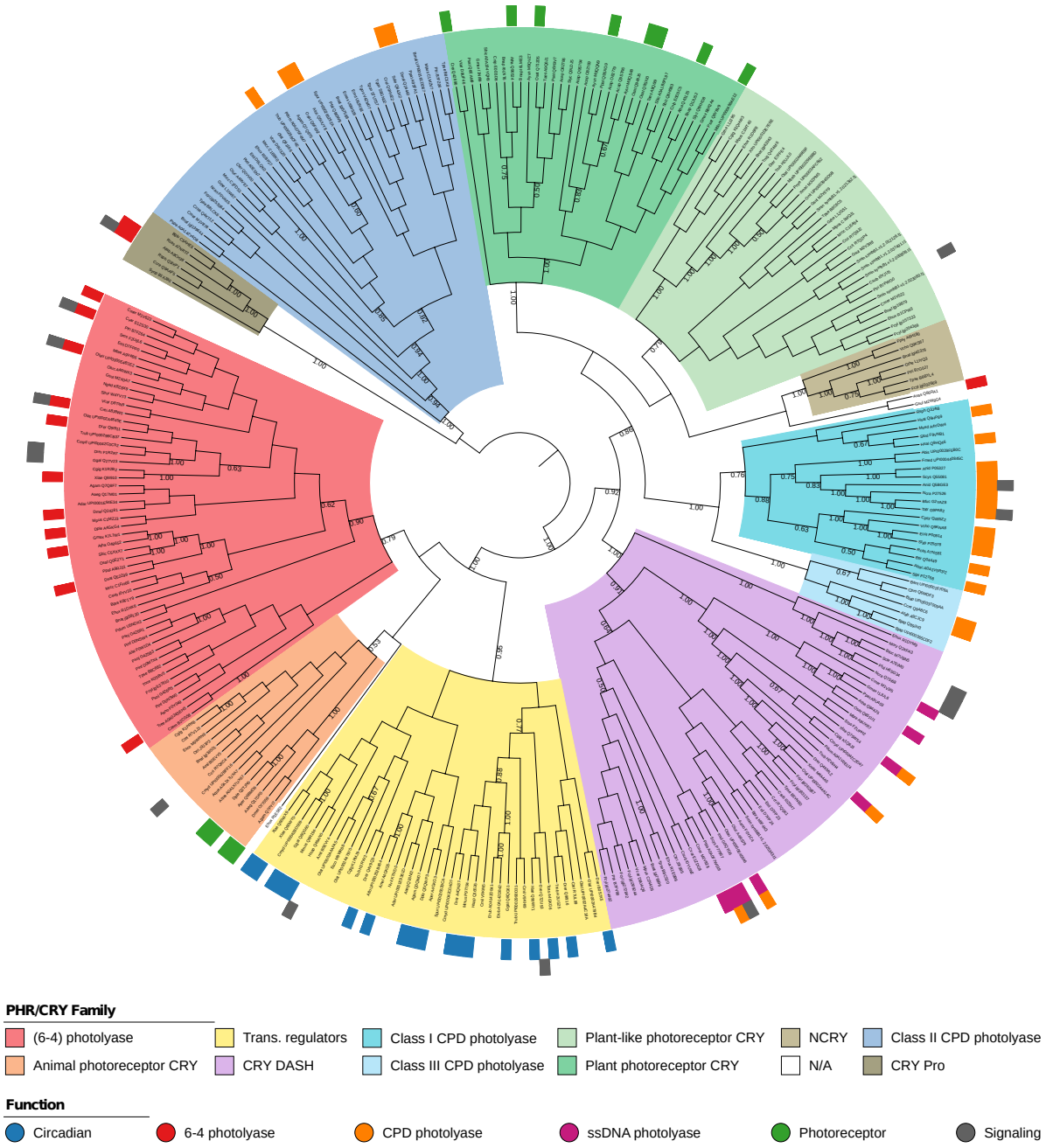


Figure S2. ProfileView tree of 307 FAD-binding domain sequences built from FAD-binding domain (PF03441) models from Pfam v31 using a hierarchical agglomerative clustering strategy. Colors of subtrees are identified by representative models and correspond to known CPF classes, with the exception of the Ncry subtree. External coloured labels define known functions for the sequences. Some of the 307 sequences are known to hold multiple functions and are labelled by two colors. The function "signalling" (grey) refers to signalling 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 best representative model occurring in the model library.

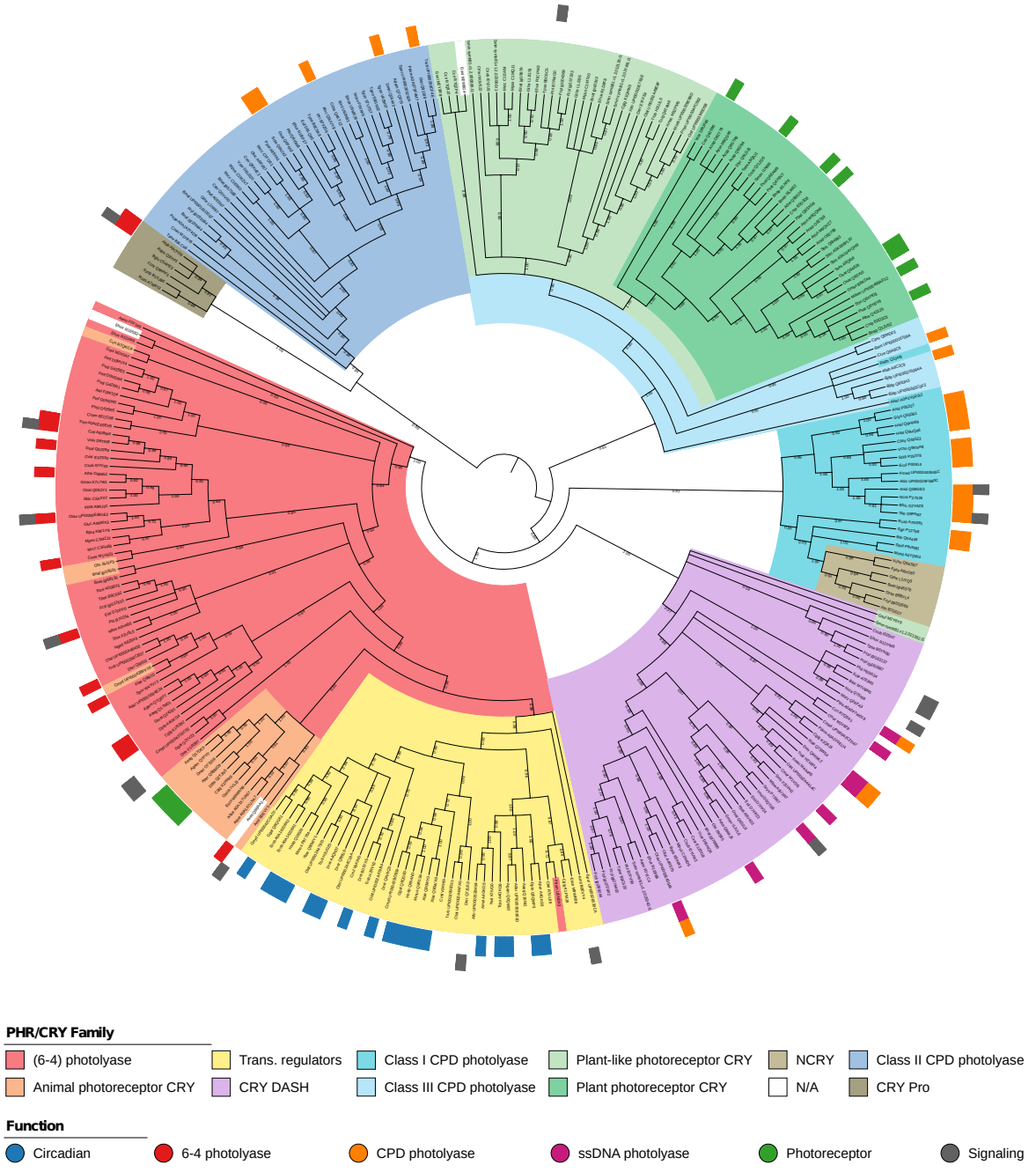
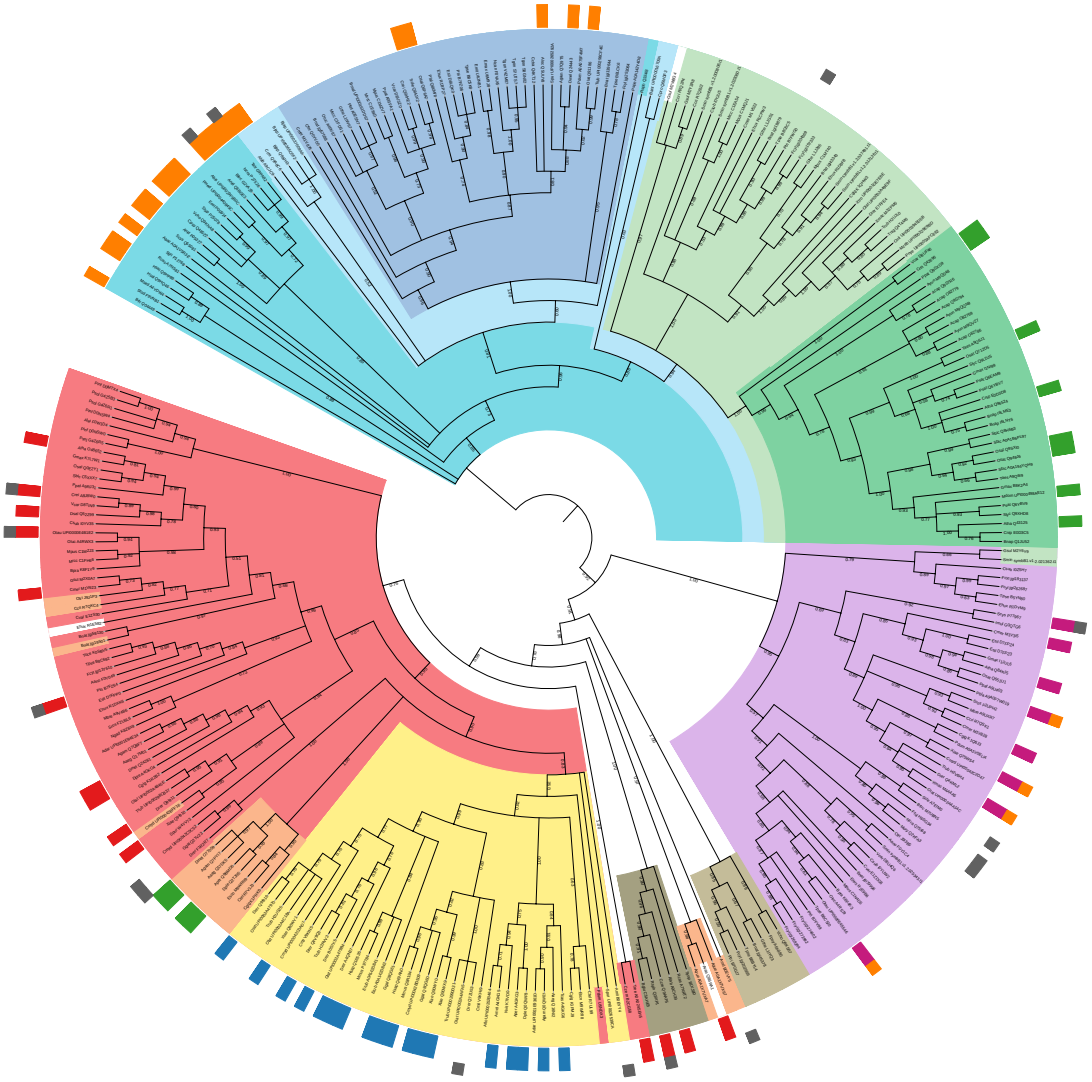


Figure S3. Distance tree constructed from 307 CPF sequences. Each sequence in the distance tree is coloured as in the ProfileView tree. Colors of internal subtrees are induced by sequence coloring. External labels report known functions for the sequences (see legend of **Fig. S2**). Numbers on the branches are bootstrap values.



PHR/CRY Family

- (6-4) photolyase
- Trans. regulators
- Class I CPD photolyase
- Plant-like photoreceptor CRY
- NCRY
- Class II CPD photolyase
- Animal photoreceptor CRY
- CRY DASH
- Class III CPD photolyase
- Plant photoreceptor CRY
- N/A
- CRY Pro

Function

- Circadian
- 6-4 photolyase
- CPD photolyase
- ssDNA photolyase
- Photoreceptor
- Signaling

Figure S4. Distance tree constructed from 307 FAD-binding domain sequences. Each sequence in the distance tree is coloured as in the ProfileView tree. Colors of internal subtrees are induced by sequence coloring. External labels report known functions for the sequences (see legend of **Fig. S2**). Numbers on the branches are bootstrap values.

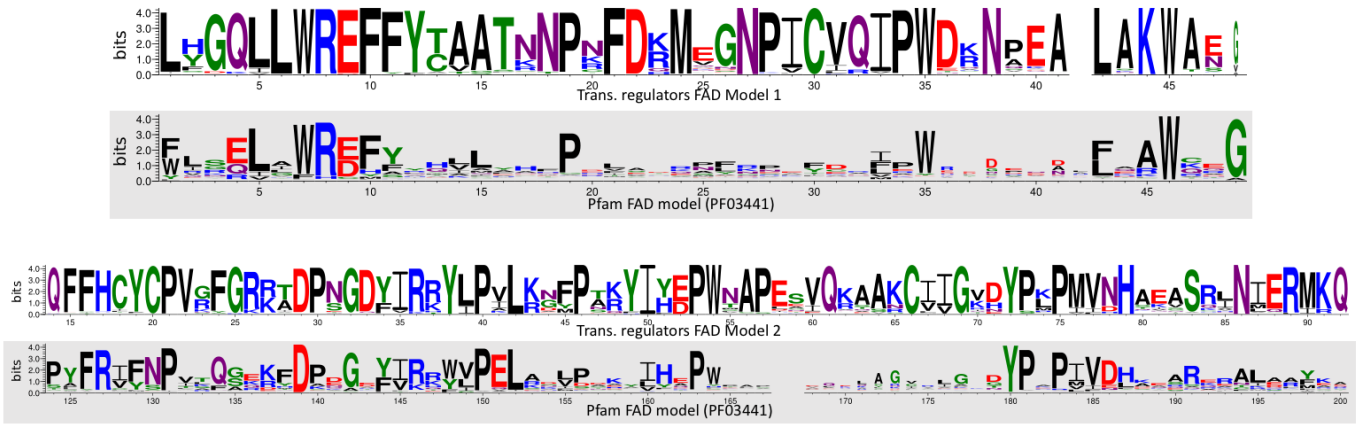


Figure S5. Comparison between trans. regulators models and Pfam models. Alignment of the two full models, corresponding to the trans. regulators motif 1 in A (top) and the trans. regulators model 2 in B (bottom), on two distinct regions of the Pfam FAD model PF3441 (grey background). The regions do not overlap.

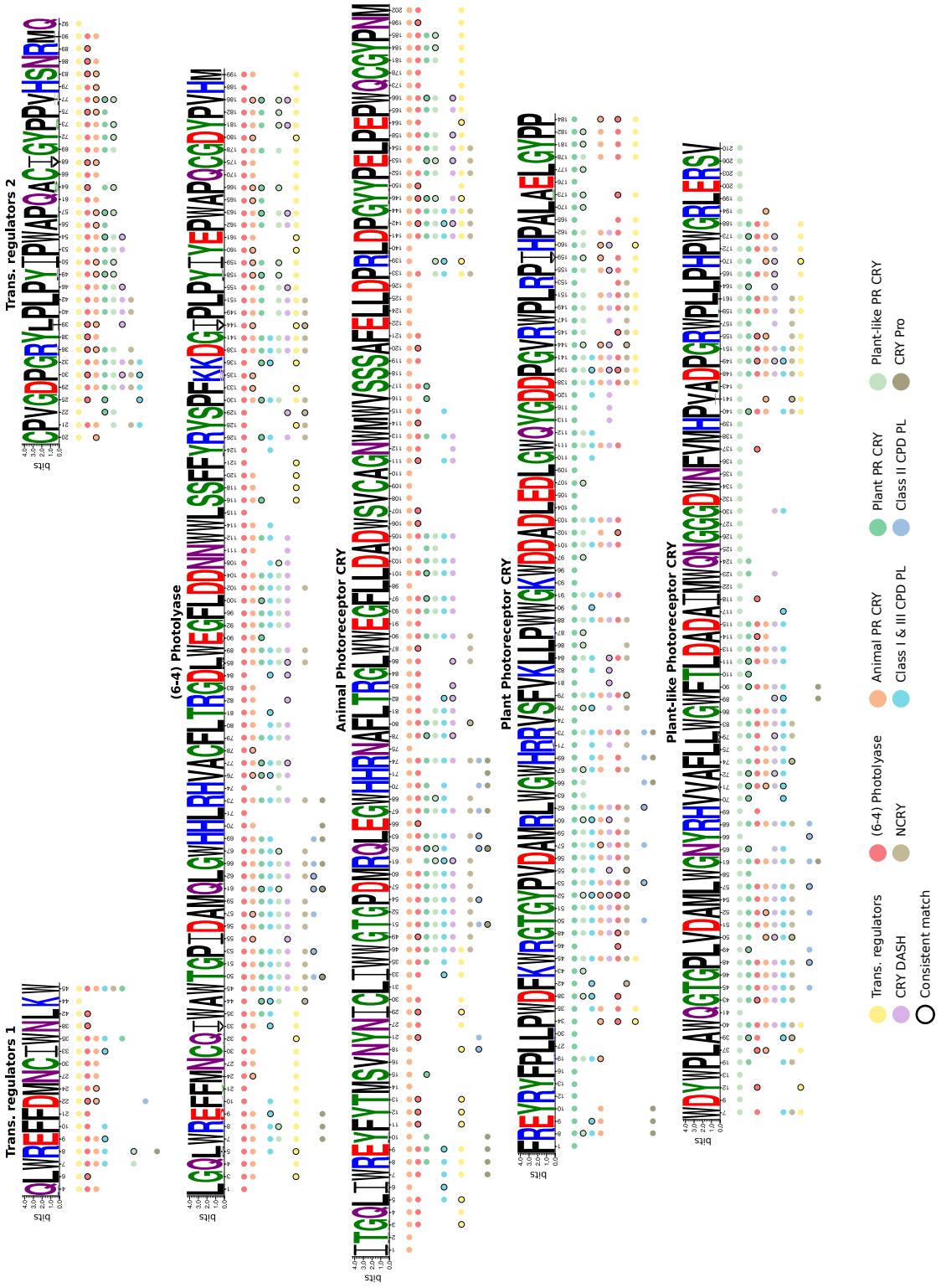


Figure S6. Eleven motifs for 10 subtrees in the ProfileView tree of CPF. Each motif for a class is represented by the most conserved positions in the corresponding representative model, that is positions showing > 60% frequency in the associated alignment (see Methods). Below each position, the coloured dots indicate that the position is well-conserved in other motifs (after their alignment; see Methods). Circled dots indicate that the position in the motif is not conserved as much in another motif (see Methods). The possible asymmetric distribution of color dots or an absence of dots between comparable positions in motifs is explained in Methods. Specific positions in a motif have no additional dot. The transcriptional regulators' subtree, represented by two distinct representative models, is provided with two independent motifs. For each motif, coloured dots are ordered, from top to bottom, depending on the best E-values given by `hhblits` to the pairwise alignments.

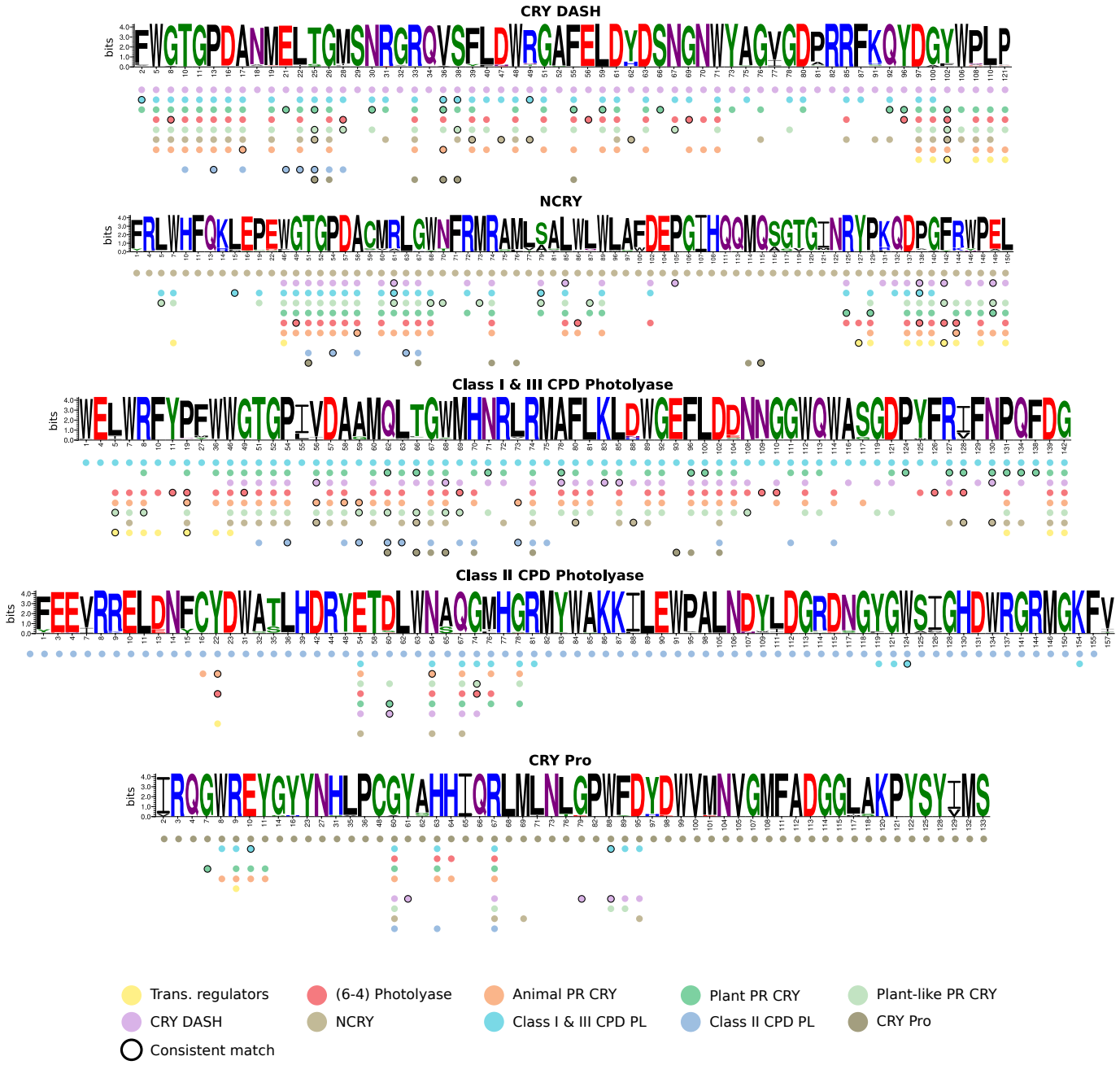


Figure S7. Eleven motifs for 10 subtrees in the ProfileView tree (continued). See legend in **Fig. S6**.

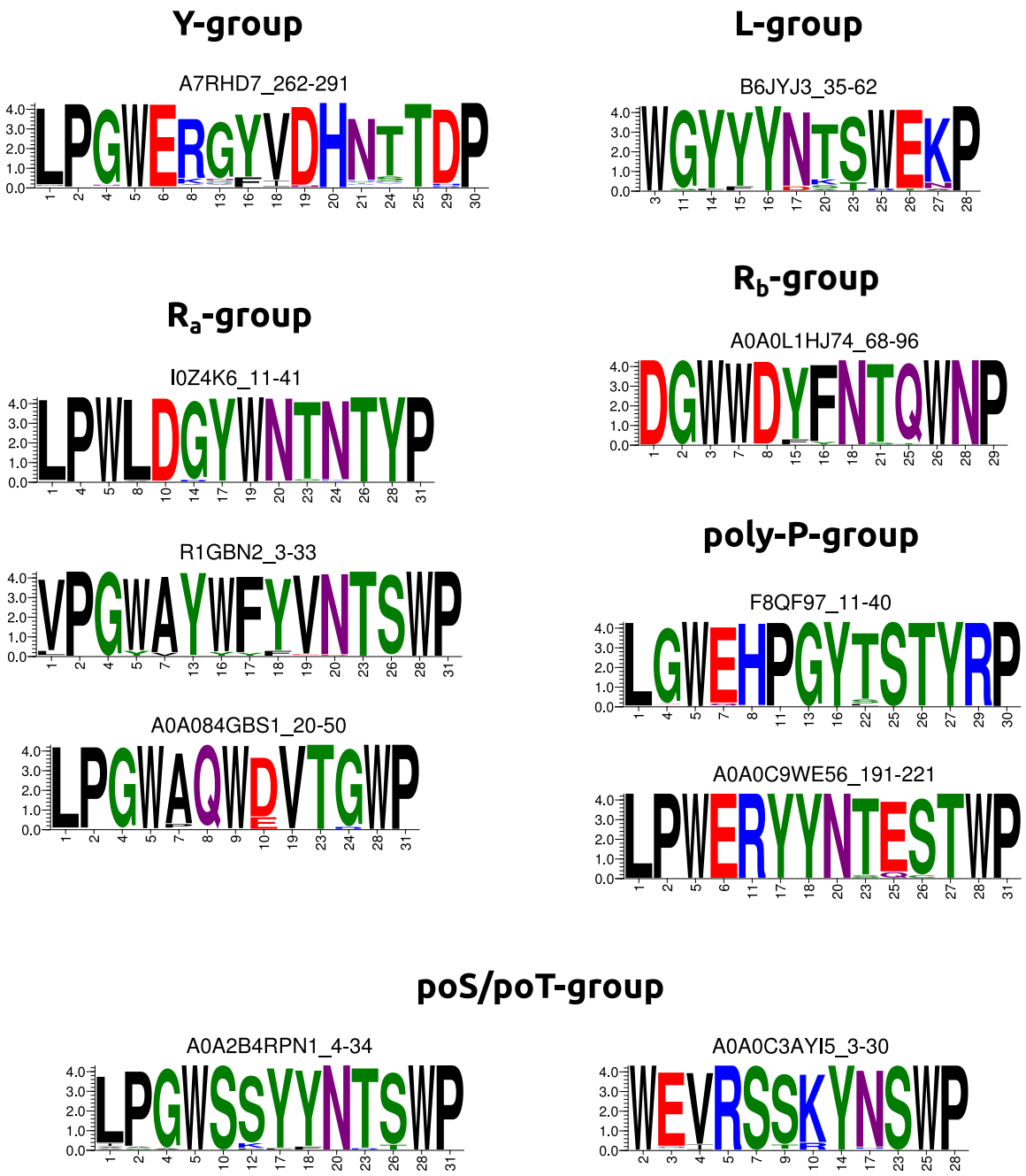


Figure S8. Ten motifs for 11 subtrees in the ProfileView tree of WW domains. Each motif for a group is represented by the most conserved positions in the corresponding representative model, that is positions showing $> 60\%$ frequency in the associated alignment (see Methods). Notice that by using the threshold of $> 60\%$ frequency from `hhblits`, one of the R_a models does not provide any conserved motif.