BMC Genomics

RESEARCH

Open Access

Searching for α–solenoid proteins involved in organellar gene expression

Céline Cattelin[1,2], Rebecca Goulancourt[1], Emanuel Chatellet[1], Alexis Astatourian[1], Charles H. Robert[2], Francis-André Wollman[1] and Ingrid Lafontaine[1*]

Abstract

Background In photosynthetic eukaryotes of the green lineage, the expression of the chloroplast genome is mainly regulated post-transcriptionally, by RNA-binding proteins encoded in the nuclear genome termed organelle trans-acting factors. Most of those identified to date belong to two families of α–solenoid proteins - the pentatrico-peptide repeat (PPR) and octatrico-peptide repeat (OPR) families - and interact with specific sequences on their target mRNAs through a domain composed of repeated motifs, allowing their maturation, splicing, editing, stabilization and translation activation.

Results To identify new organelle trans-acting factors, we developed three approaches for annotating α-solenoid proteins targeted to the chloroplast or the mitochondria. One to identify distant homologs of existing organelle trans-acting factors families, and two others (decision tree and random forest classifiers) to identify new organelle trans-acting factors families. The combined approaches efficiently retrieve previously annotated organelle trans-acting factors in 2 model organisms. It identified 1067 OPR proteins and 4983 PPR proteins in 43 proteomes of Archaeplastida. Our analysis also identified chimeric proteins composed of both OPR and PPR domains. Finally, our results identified 3300 other α-solenoid candidates which are likely to participate as new regulators of organelle gene expression. In particular, we identified new candidates in species in which the regulatory mechanisms of plastid gene expression are still understudied, such as in the glaucophyte *Cyanophora paradoxa* and the red alga *Porphyridium purpureum*.

Conclusion Our study contributes to the extensive description of organelle trans-acting factors by providing valuable new tools to decipher their repertoire and new candidates for experimental characterization in the entire eukaryotic tree of life.

Keywords Regulation of gene expression, Chloroplast, Mitochondria, Alpha-solenoid, Post-endosymbiotic innovations, Annotation

*Correspondence:
Ingrid Lafontaine
ingrid.lafontaine@sorbonne-universite.fr
[1]UMR7141, Photobiology and Physiology of Plastids and Microalgae, P3M, Sorbonne Université, CNRS, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris F-75005, France
[2]UMR8266, Laboratoire de Biochimie Théorique, Université Paris Cité, CNRS, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Background

Eukaryotic photosynthesis is ensured by plastids, organelles originally acquired *ca.* 1.5 billion years ago from a primary endosymbiosis involving a protist host and a cyanobacterial ancestor, which gave rise to the extant green algae and land plants (together Viridiplantae), to rhodophyte and glaucophyte algae [1]. These endosymbiotic events were followed by massive gene transfers from the plastid progenitors to the nucleus of the host cell, as had already occurred during mitochondrial endosymbiosis, which probably involved an Archaeal host and an α-proteobacterial ancestor, *ca.* 1.8 billion years ago [1]. Most of the organelle proteome - either mitochondrion or plastid-, is now nucleus-encoded, translated in the cytosol and imported into the organelle. However, these energy-providing organelles have retained a tiny genome, therefore most of the major protein complexes associated with their bioenergetic membranes are genetic mosaics with subunits encoded in two different genomes. To ensure cell viability and acclimation of the organelle activity (energy production and metabolism), expression of organelle genomes became, during evolution, closely interconnected with that of the host cell. While the expression of nuclear genes is regulated at multiple levels (transcriptionally and post-transcriptionally as well as via epigenetic marks), plastidial and mitochondrial genomes are known to date to be mainly regulated post-transcriptionally by nucleus-encoded RNA-binding proteins, as demonstrated by pioneering studies in Viridiplantae and in yeast [2–4]. Most of the factors described up to now, hereafter named OTAF (for organellar trans-acting factor), contain a succession of either PPR motifs of 35 residues (pentatrico-peptide repeat) or OPR motifs of 38 residues (octatrico-peptide repeat) [5–8]. They belong to a large class of proteins containing repeated motifs forming anti-parallel α-helices, like mTERF (mitochondrial transcription termination factors), HEAT, Ankyrin, Armadillo and Pumilio repeats, which confer a rod-shape like structure with concave surface where ligands can bind [9–12]. PPR and OPR motifs are degenerated and form pairs of antiparallel α-helices. The variable number of repeats of the motif form an α-solenoid shape with a positively charged surface that binds to the mRNA in a sequence-specific manner [7, 13]. PPR repeats are related to TPR repeats [5, 14] that mainly mediate protein-protein interactions and are involved in a variety of cell processes [15]. Sel1-like repeats are themselves linked to TPR repeats, while HAT repeats are "Half-A-TPR" repeats, some of which, like PPR, bind to RNA [16]. The PPR and OPR protein repertoires are remarkably diverse between organisms, with land plants containing several hundreds of PPR and few OPR, whereas the green alga *Chlamydomonas reinhardtii* encodes only a dozen PPR [17, 18] but more than a hundred of OPR [6, 19–23]. PPR and OPR proteins also contain additional domains, like the DYW domain in PLS-type PPR, involved in editing [7, 24, 25] and the RAP domain with a probable endonuclease activity [26] at the C terminus of several OPR in *C. reinhardtii* [6, 22, 27].

To tackle the issue of the PPR motifs degeneracy, profile-based methods have been developed, like TPRpred that identifies TPR motifs and related PPR and Sel1-like motifs [28]; PPRFinder based on plant homologs [17] and SCIPHER based on yeast homologs [29]. These approaches use taxonomically restricted profiles of either plants or yeasts to perform motif search. Given the highly biased distribution of PPR and OPR repeats across the eukaryotic tree, it is likely that those families are still incomplete, especially for OPR.

Here, we present three procedures to complete the catalogue of known OTAFs and discover new OTAF families, which we applied to a representative set of Archaeplastida species. The first approach is an iterative profile-based similarity procedure to retrieve distant OTAFs homologs **p**redicted to be **t**argeted to **o**rganelles (*pto*), in order to fully describe their distribution. Applied to OPR and PPR, we show that OPR expansions were restricted within Chlorophyta and that outside of Viridiplantae, PPRs and OPRs are few in number, suggesting that other players in the regulation of gene expression in organelles remain to be discovered. These differences likely reflect genetic adaptation to different lifestyles or ecological niches. We also present two machine learning procedures to retrieve new families of nuclear-encoded candidates likely to be involved in organelle genome expression, *i.e. pto* proteins adopting an α-solenoid shape with similar physico-chemical properties as known OTAFs. We thus identified several dozens of new *pto* α-solenoid candidates, including a family of α-solenoid proteins tandemly duplicated in *C. reinhardtii*, whose experimental characterization would be relevant to understanding their possible contribution to chloroplast gene expression.

The tools developed in this study will allow the identification of OTAFs outside Archaeplastida, which is of genuine interest owing to the scarce knowledge on the regulation of gene expression in organelles beyond what has been reported in Opistokhonta and Viridiplantae.

Results

Procedures for the identification of OTAF candidates

We developed procedures (see Methods for details) to identify distant homologs of known OTAF families, and to identify novel OTAF candidates. These procedures were applied to establish the OTAF catalogues within a set of 43 proteomes from all three taxa of Archaeplastida: the only available one in Glaucophyta, the 5 available ones in Rhodophyta, and within Chloroplastida, the

23 Chlorophyta reference proteomes available at Uniprot, plus the JGI proteome of *C. reinhardtii*, and 12 proteomes of Streptophyta including *A. thaliana* (Additional file 2: Table S1).

Iterative profile-based procedure to retrieve homologs with known motifs

To retrieve distant homologs of known OTAFs, we developed an Iterative Profile-Based Procedure (IPB) (Fig. 1, left panel). Profiles built from a defined set of motifs (for example PPR or OPR motifs) are used to search against the proteomes of interest to identify new motifs. Proteins with significant hits from one or more profile are selected if they are predicted to be localized in mitochondria or chloroplast (pto) by at least 2 over 4 prediction algorithms (see Methods). The procedure can be iterated a defined number of times, or until no new *pto* protein is identified. After each iteration, all motifs (newly identified or already existing) are clustered according to their similarity into $\text{clusters}_{\text{motifs}}$ with Markov Clustering (see Methods). A new profile is built for each $\text{cluster}_{\text{motif}}$, and all obtained profiles will be used for the next iteration. In this study, we built profiles from 107 motifs from 12 published OPR proteins and 155 P-motifs, the most abundant type of PPR motifs [13] from 11 PPR proteins (Additional file 2: Table S2). With the OPR profiles, 10 iterative steps were performed to reach convergence, i.e. until no new *pto* protein was retrieved. With the PPR motifs, 10 iteration steps nearly reached convergence (only one protein was retrieved after the last iteration).

In order to find new OTAF candidates with physicochemical and structural properties similar to those of the known OTAF but not necessarily evolutionary related to OPR and PPR families, we developed two classification procedures based on predicted protein properties, one based on a decision tree and one based on a random forest. These approaches thus retrieve potential α-solenoid proteins composed of repeated motifs and pairs of α-helices targeted to an endosymbiotic organelle.

Decision tree procedure

The decision-tree procedure (DT) is based on both the detection of repeated sequences (sequence repeat) within the protein and of predicted anti-parallel α-helix pairs (structural repeat), because the sequence similarity between the structural repeats could be undetectable and because a sequence repeat does not necessarily fold into an anti-parallel α-helix pair (Methods and Fig. 1, central panel). All proteins containing a predicted transmembrane helix by TMHMM (see Methods) are excluded, as most published studies describe OTAFs as soluble proteins. Only those candidates with at least two repeated sequences ($>=26$ amino acids), *i.e.* two motifs predicted by RADAR, were selected. Selected

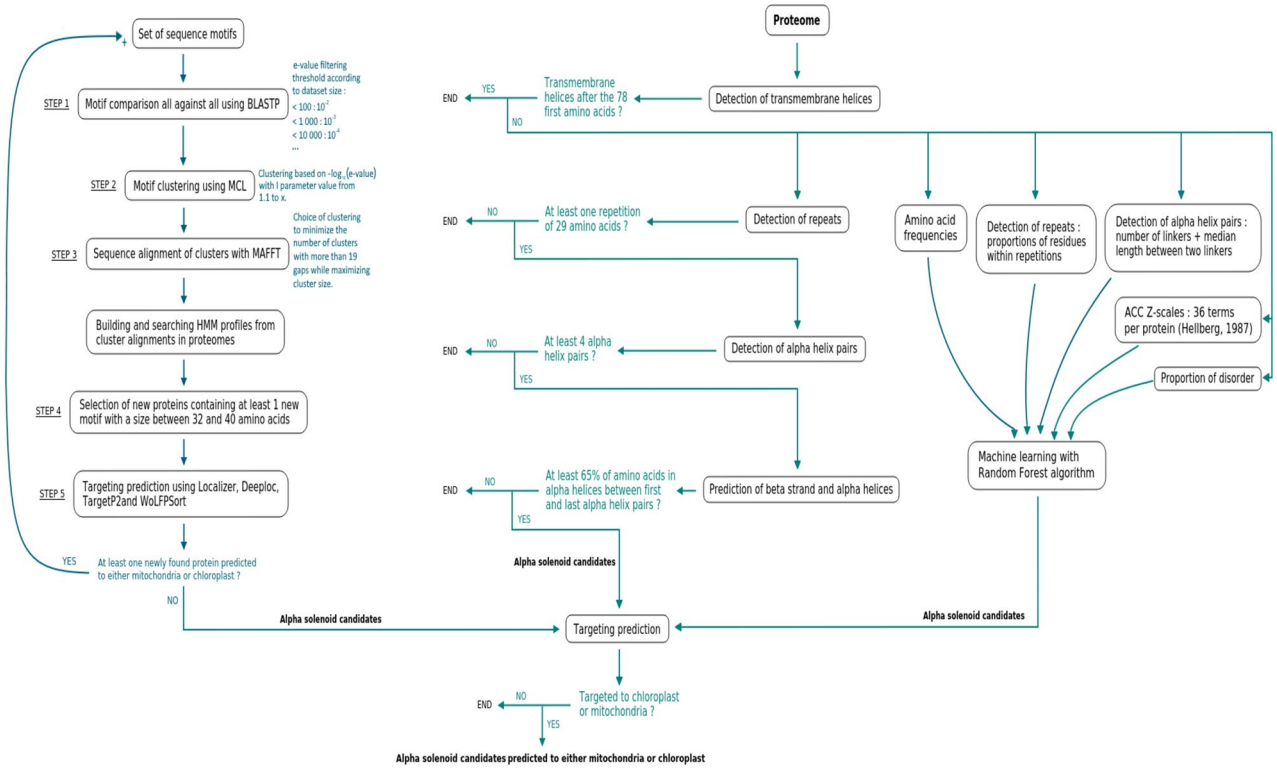


Fig. 1 Workflow of the three developed procedures. IPB, left; DT, middle; RF, right

candidates at the third step are those that also contain at least 4 linkers (separated by 32 up to 400 amino acids) predicted by ARD2 [30]. A linker is defined as the amino acid(s) between the two α-helices of a structural repeat unit. Four linkers implies that there are four structural repeats, not necessarily detected as sequence repeats, or three structural repeats if ARD2 detects a linker between structural units. The maximum distance of 400 amino acids between two linkers implicitly takes into account the fact that ARD2 can miss some linkers [30]. In that perspective, 400 amino acids would correspond broadly to the maximum sequence distance between the first and the last repeat of a suite of 10 structural repeats.

The parameterization of DT was performed on a training set composed of 426 α-solenoid proteins (positive control) and 286 non α-solenoid proteins (negative control) classified based on their experimentally resolved or predicted 3D structure (see Methods). Threshold values for RADAR and ARD2 were determined to maximize the precision and recall of the detection on the training set (Table 1), but the F1-score of DT remains modest (0.46), with a rather good specificity 0.73 at the expense of recall (0.34) (Table 1).

Random forest procedure

To improve the sensitivity of the similarity-free detection approach and to avoid the definition of threshold values for selecting candidates as in the DT approach, we also developed a random forest (RF) classifier (Fig. 1, right panel), in which composition and physico-chemical properties are also considered. Each protein is thus described by 61 variables. Four variables directly describe α–solenoid properties: the proportion of amino acids in repeats, the median length of the repeats, the number of linkers between two α-helices and the median length between the predicted linkers. The remaining general variables are the 20 amino-acid frequencies proportion of amino acids with disorder propensity, and 36 Auto-Cross Correlation terms between the Z-scales physico-chemical descriptors for each amino-acid [31] as described in [32]. The RF classifier was trained on the same training set as the DT procedure (see Methods). Over 1000 iterations of the model on our validation test (those 10% proteins from the positive and negative sets that were never used to train the model), on average the precision is 0.97, sensitivity 0.98 and specificity 0.96 (Table 1). For detection, RF was run 1000 times and candidates were retained if they were retrieved in at least 90% of the iterations.

The importance of each variable in the classifier is shown in Additional file 1: Figure S1. The most important one is the proportion of residues within repeat regions, and the second ones are the AAC term z1.2.lag4, that reflects amphipathic constraints and the frequency of Proline, which is known to destabilize the conformation when occurring in the middle of an α-helix, but which has a stabilizing effect at their N-terminus [33]. The median length of the repeats plays only a limited role. This was quite expected given the wide distributions in both positive and negative controls (see Methods). The importance of variables related to the α-helices (number of linkers and median length between linkers) and the proportion of disorder is lower than 32/60 other variables, making RF notably different from DT.

Table 1 Performances of IPB and DT on the training set and average performance of 1000

Procedure	Specificity	Recall	Precision	F1-score
	TN/(TN + FP)	TP/(TP + FN)	TP/(TP + FP)	
DT	0.82	0.34	73	0.46
RF	0.96	0.98	0.97	0.97

Comparative ability of the three procedures to retrieve those OPR and PPR proteins, previously annotated in model organisms

We determined how many of the annotated PPR in *A. thaliana* and OPR in *C. reinhardtii* were retrieved by each procedure.

Iterative profile-based

From P-class PPR motifs, IPB retrieved 95% (451/475) of the PPR proteins listed in the PPR database at University of Western Australia [13] with a UniProtKB identifier in the *A. thaliana* proteome (named PPR UWA hereafter) and 3 PPR not listed at UWA, but annotated as PPR in Swissprot. IPB also retrieved 10 of the 14 PPR annotated in *C. reinhardtii* (Table 2).

Regarding OPR proteins, IPB retrieved the unique one (ATRAP, AT2G31890) in *A. thaliana* and 119 already annotated OPR in *C. reinhardtii*. Among those OPR in *C. reinhardtii*, 56 (out of 58) are published OPR proteins and 57 (out of 60) additional proteins annotated as "OctotricoPeptide Repeat Protein" in the annotation v5.6 version available at JGI, Phytozome13 (Table 2). There are 5 OPR that IPB failed to retrieve, most probably because their OPR motifs are too distant from the ones used: the published OPR Raa3 [34] and NCL18 [27] and three additional OPR annotated at JGI (Cre07.g347500 annotated as OPR114, Cre02.g146900 and Cre07.g336500). However, they were retrieved by RF and or DT (see below). The heptatrico-peptide repeat (HPR) proteins described in [35] are recognized by IPB as harboring OPR motifs, confirming that HPR proteins belong to the OPR family (Additional file 2: Table S3a).

In summary, the IPB procedure has 95% precision and 93% sensitivity (Table 3).

Table 2 Number of known PPR and OPR detected by the three methods, IPB, DT and RF in *A. thaliana* and *C. reinhardtii*

	A. thaliana	**Fraction detected**	***C. reinhardtii***	**Fraction detected**
PPR UWA	475	-	14	-
detected by IPB	451	0,95	10	0,71
detected by DT (not by IPB)	197 (2)	0,41	9 (2)	0,64
detected by RF (not by IPB or DT)	454 (13)	0,96	10 (1)	0,71
total detected	466	0,98	13	0,93
PPR Swissprot	449	-	1	-
detected by IPB	434	0,97	0	0,00
detected by DT (not by IPB)	197 (3)	0,43	0	0,00
detected by RF (not by IPB or DT)	438 (10)	0,98	1	1,00
total detected	447	1,00		
published OPR	1	-	58	-
detected by IPB	1	1	56	0,97
detected by DT (not by IPB)	0	0	41 (1)	0,71
detected by RF (not by IPB or DT)	0	0	54 (0)	0,93
total detected	1	1	57	0,98
additional annotated OPR at JGI	-	-	60	-
detected by IPB	-	-	57	0,95
detected by DT (not by IPB)	-	-	50 (2)	0,81
detected by RF (not by IPB or DT)	-	-	54 (1)	0,87
total detected			62	1,00
Total described OPR	1		120	
detected by IPB	1	1	115	0,96
detected by DT (not by IPB)	0	0	92 (3)	0,77
detected by RF (not by IPB or DT)	0	0	109 (2)	0,91
total detected			119	0,99

PPR annotated at Swissprot are all the proteins with the term pentatricopeptide or PPR in their names. 3 PPR are annotated by Swissprot but absent from UWA: At1g06143 (PPR15), At1g19525 (PPR51) and At2g13420 (PP150). 33 PPR are unique to UWA. Note that among the PPR list at UWA, 24 are annotated as TPR in TrEMBL. This is expected as the TPR and PPR are evolutionary related

Table 3 Candidates retrieved by IPB unpublished and unreferenced as OPR with at least 2 motifs (OPR or repeat$_{radar}$)

locus ID	**mito/chlo**	**repeat radar**	**opr motif**	**α-solenoid shape**	**Annotation Phyto v5.6**	**Annotation Phyto v6.1**
Cre01.g025808	2	12	1	yes	-	OctotricoPeptide Repeat protein 122**
Cre01.g033250	2	8	1	yes	-	-
Cre02.g109200	3	5	1	yes	-	OctotricoPeptide Repeat protein 123**
Cre03.g145867	3	11	1	yes, 7ptk	-	mitochondrial ribosome protein ml113
Cre03.g168400	0	10	3	yes*	-	OctotricoPeptide Repeat protein 134**
Cre03.g201103	2	9	1	yes*	RAA7, psaA mRNA trans-splicing factor	psaA mRNA trans-splicing factor
Cre04.g217909	2	13	3	yes	-	OctotricoPeptide Repeat protein 124**
Cre04.g228550	1	40	1	yes *	-	-
Cre06.g251400	3	3	1	no	MME6, NADP-dependent malic enzyme	NADP-dependent malic enzyme
Cre09.g395880	2	11	2	yes	-	OctotricoPeptide Repeat protein 11**
Cre13.g579100	0	na	1	yes *	APC1, Anaphase promoting complex subunit 1	APC1, Anaphase promoting complex subunit 1
Cre13.g605850	2	33	1	yes *	-	OctotricoPeptide Repeat protein 127**
Cre13.g605900	2	28	1	yes	-	OctotricoPeptide Repeat protein 128**
Cre16.g651900	2	12	1	yes	CRI1, carotenoid isomerase	OctotricoPeptide Repeat protein 129**
Cre16.g662951	0	35	1	yes	-	-

locus ID: locus identifier in the v5.6b annotation. targ: number of predictions for chloroplast/mitochondrial targeting; radar repeat: number of repeats predicted by RADAR. OPR motif: number of OPR motif found by IPB. α-solenoid shape: Yes/No (AF2 predictions). When an experimental 3D structure is available the PDB accession number is given. * indicates average pLDDT quality scores (< 0.5) for the AF2 predictions. ** [36]

Decision tree

In the *A. thaliana* proteome, DT retrieved only 197 PPR, 2 of which were not retrieved by IPB (Table 2). DT also found 3 out of the 8 TPR annotated in SwissProt in *A. thaliana*. In the *C. reinhardtii* proteome, 41/50 of the published/additional OPRs were retrieved, all but one (Cre09.g392579) were not retrieved by IPB; DT also retrieved 9 of the 14 PPR present in the *C. reinhardtii* proteome.

Random forest

In the *A. thaliana* proteome, the RF procedure retrieved 454 PPR, 13 of which were not retrieved, by either IPB or by DT (Table 2). In *C. reinhardtii*, it retrieved 54/54 of the published/additional OPR. Raa3 and Ncl18, the sole published OPRs that were not retrieved by IPB, were retrieved in 3/1000 and 571/1000 RF iterations, respectively. RF also retrieved an additional OPR annotated at JGI (Cre07.g347500, OPR31) that was not retrieved by IPB nor by DT (Table 2). As shown in Table 2, altogether a same subset of 197 PPR candidates in *A. thaliana* and 91 OPR candidates in *C. reinhardtii* were identified by the three procedures. By cumulating IPB, DT and RF results, 98% and 99% of the PPR and OPR were detected, respectively.

New OPR candidates

No new PPR have been identified in *A. thaliana* nor in *C. reinhardtii* with IPB, but 15 new OPR have been identified in *C. reinhardtii*, which were not annotated as such in the v5.6 version of genome annotation (Table 3). Among them, 11 are *pto* candidates. Only three of these new OPR candidates have a gene name in the v5.6 annotation available at Phytozome13: (i) Cri1 (Cre16.g651900), with 11 repeats predicted by RADAR (repeats$_{RADAR}$) and annotated as a carotenoid isomerase, adopts an α-solenoid shape predicted by AlphaFold2, (hereafter named AF2pred.); (ii) Mme6 (Cre06.g251400) with 3 repeats$_{RADAR}$ is likely a false positive since its AF2pred 3D structure of low-quality displays α-helices not adopting an α-solenoid shape. (iii) Raa7 (Cre03.g201103), the psaA mRNA trans-splicing factor [37], with 7 repeats$_{RADAR}$. Although other trans-splicing factors Raa1, Raa8 and Rat2 contain OPR motifs [38], Raa7 has not been described as an OPR. Its AF2pred. 3D structure although of average plDDT quality score below 0.5, displays pairs of α-helices and could fold into an α-solenoid conformation. The 8 other *pto* candidates are uncharacterized and adopt an α-solenoid shape according to AF2pred. Only 2 of the *pto* candidates have more than one OPR motif, but they all contain at least 3 repeats$_{RADAR}$. Thus, some OPR motifs have diverged upon recognition by profile similarity, compared to the motifs selected to initiate the IPB procedure (Additional file 1: Figure S2), but they are still detected by RADAR. Note that Cre03.g145867 is part of the large subunit of the mitoribosome (PDB entry 7pkt, chain M).

New OTAF candidates

Because of the low recall of DT and the low specificity of RF on OPR and PPR from *C. reinhardtii* and *A. thaliana*, we considered the *pto* candidates identified by both DT and RF (*pto* DTRF) as the most robust newly identified OTAF candidates. All candidates retrieved by IPB with 2 OPR/PPR motifs and all *pto* DTRF candidates are listed in Additional file 2: Table S4.

Among the 45 *pto* DTRF candidates found in *A. thaliana*, 5 are PPR UWA and were not detected by IPB. Among the 189 *pto* DTRF candidates in *C. reinhardtii*, there is one OPR (OPR14) and one PPR (PPR11) that were not detected by IPB (Additional file 2: Table S4). To estimate the quality of the DTRF candidates, they were classified according to their 3D structure predictions (Additional file 2: Table S5) in 12 selected proteomes: 3 representative Embryophyta species among Streptophyta (the Bryophyta *Physcomitrium patens*, the Tracheophyta *A. thaliana* and the Marchantiophyta *Marchantia polymorpha*), 3 representatives Chlorophyta species (*Chlorella sorokiniana* in the Trebouxiophyceae, *Ostreococcus tauri* in the Mamiellophyceae and *C. reinhardtii* in Chlorophyceae) and the proteomes of the sole Glaucophyta species and the reference proteomes of the 5 Rhodophyta species. 66% of the *pto* DTRF candidates in those 12 representative proteomes have an α-solenoid structure predicted by AF2 with good quality. If candidates with regions of low-quality predicted structure (plDDT < 0.5) containing only α-helices and disordered regions are considered to potentially adopt an α–solenoid shape, this number raises to 77% (Additional file 2: Table S5).

Functional annotation of OTAF candidates

Functional annotation of the *pto* DTRF candidates was performed with InterProScan against the PFAM database (v37.4) (Additional file 2: Table S6). More than half of the *pto* DTRF candidates have a PFAM domain (1719/3300, 51.1%). Table 4 lists the annotations found in more than 1% of the candidates with a detected PFAM domain. Because InterPro (IPR) annotations summarize results of different databases, IPR annotations are given instead of PFAM domain description to avoid redundancy. The most frequently found IPR annotations for *pto* DTRF candidates are the MYND-type zinc finger domains (169), Ankyrin repeats (63) and Armadillo repeats (53). Note that there are also 129 *pto* DTRF candidates containing PPR/TPR related repeats motifs that were not detected by IPB with the PPR motifs (Table 4). Lastly, 40 *pto* DTRF candidates contain a mTERF domain (Table 4) and 23 contain a domain from the MTERF superfamily.

Table 4 The number of candidates with a detected PFAM motif are given, as well as the subgroup of those candidates with an IPR annotation

Types		number candidates	%total
DTRF		3300	
with PFAM domain		1719	52,1
with PFAM domain and IPR annotation		1528	46,3
	IPR002893/Zinc finger, MYND-type	169	5,1
	IPR002110/Ankyrin repeat	63	1,9
	IPR000225/Armadillo	52	1,6
	IPR001611/Leucine-rich repeat	46	1,4
	IPR002885/Pentatricopeptide repeat	46	1,4
	IPR000719/Protein kinase domain	44	1,3
	IPR003690/Transcription termination factor, mitochondrial/chloroplastic (MTERF)	40	1,2
OPR		1067	
with PFAM		221	20,7
with PFAM domain and IPR annotation		219	20,5
	IPR013584/RAP domain	163	15,2
PPR		4983	
with PFAM		4979	99,9
with PFAM and IPR		4979	99,9
	IPR002885/Pentatricopeptide repeat	4936	99,1
	IPR046848/E motif	1886	37,8
	IPR032867/DYW domain	945	19,0
	IPR046849/E2 motif	486	9,8
	IPR033443/PROP1-like, PPR domain	446	9,0

To avoid redundancy, the most frequently found annotations are the IPR ones instead of PFAM descriptions

Among those 63 candidates are 5 mTERF proteins in *A. thaliana*, and three hypothetical proteins in *C. reinhardtii.* Notably, none was found in Rhodophyta and only 2 were found in *C. paradoxa* (Table S6). Note that RF alone was able to detect 31 of the 35 mTERF proteins addressed to either mitochondria or chloroplast in *A. thaliana* (Additional file 2: Table S3b).

The vast majority (99%) of the PPR candidates have described PPR domains and associated DYW and E domains (Table 4). Because the OPR motif is not described in protein domain databases, the vast majority of OPR candidates lacks characterization. Only 219 out of the 1067 OPR candidates have a PFAM domain (Table 4), of which 163 have a RAP domain, including 10 NCL OPR-RAP tandemly duplicated in *C. reinhardtii* [27]. Note that according to Uniprot annotations, many OPR candidates are composed of disordered regions and regions with compositional biases, either towards acid and basic or polar residues.

Distribution of OPR, PPR and new OTAF candidates across Archaeplastida

Figure 2 shows the distribution of the 6050 candidates retrieved by IPB with at least 2 OPR/PPR motifs and 3300 *pto* DTRF candidates in the 43 studied proteomes. PPR are present in all phyla. There are less than 10 PPR per species in Rhodophyta, Glaucophyta and Chlorophyta and several hundred copies per Streptophyta species, except in the two early-diverging Streptophytes *Mesostigma viride* and *Chara braunii*.

OPR are the most abundant in Chlorophyta, with at least 10 copies, and no less than 48 OPR copies per species in Chlorophyceae, but very few are found in other phyla, except in the early diverging unicellular green alga *Mesostigma viride* (Streptophyta), which contains 17 OPR candidates and in *P. purpureum* which contains 10 OPR candidates. In *P. purpureum*, OPR are automatically annotated as Tbc2 translation factor, chloroplastic, except A0A5J4Z7V5 that remains uncharacterized. If the encoded proteins have the same function as in *C. reinhardtii*, this would suggest that the PsbC mRNA encoding the CP43 subunit of the photosystem II is particularly regulated [39]. The distribution of *pto* DTRF candidates, as for OPR and PPR, is highly contrasted in the different lineages. For example, there are more than 150 *pto* DTRF candidates in *Cyanophora paradoxa* compared to only 22 PPR and 3 OPR. On the contrary, the situation in Rhodophyta is more balanced with only twice as many *pto* DTRF candidates as PPR. Among the 13 *pto* DTRF candidates in *Gracilariopsis chorda*, two are annotated as Mbb1, the PsbB mRNA maturation factor. This is also the case of 3 *pto* DTRF candidates in *P. purpureum*. While in Streptophyta there are 38 new candidates per species in average, the most contrasted distributions of new

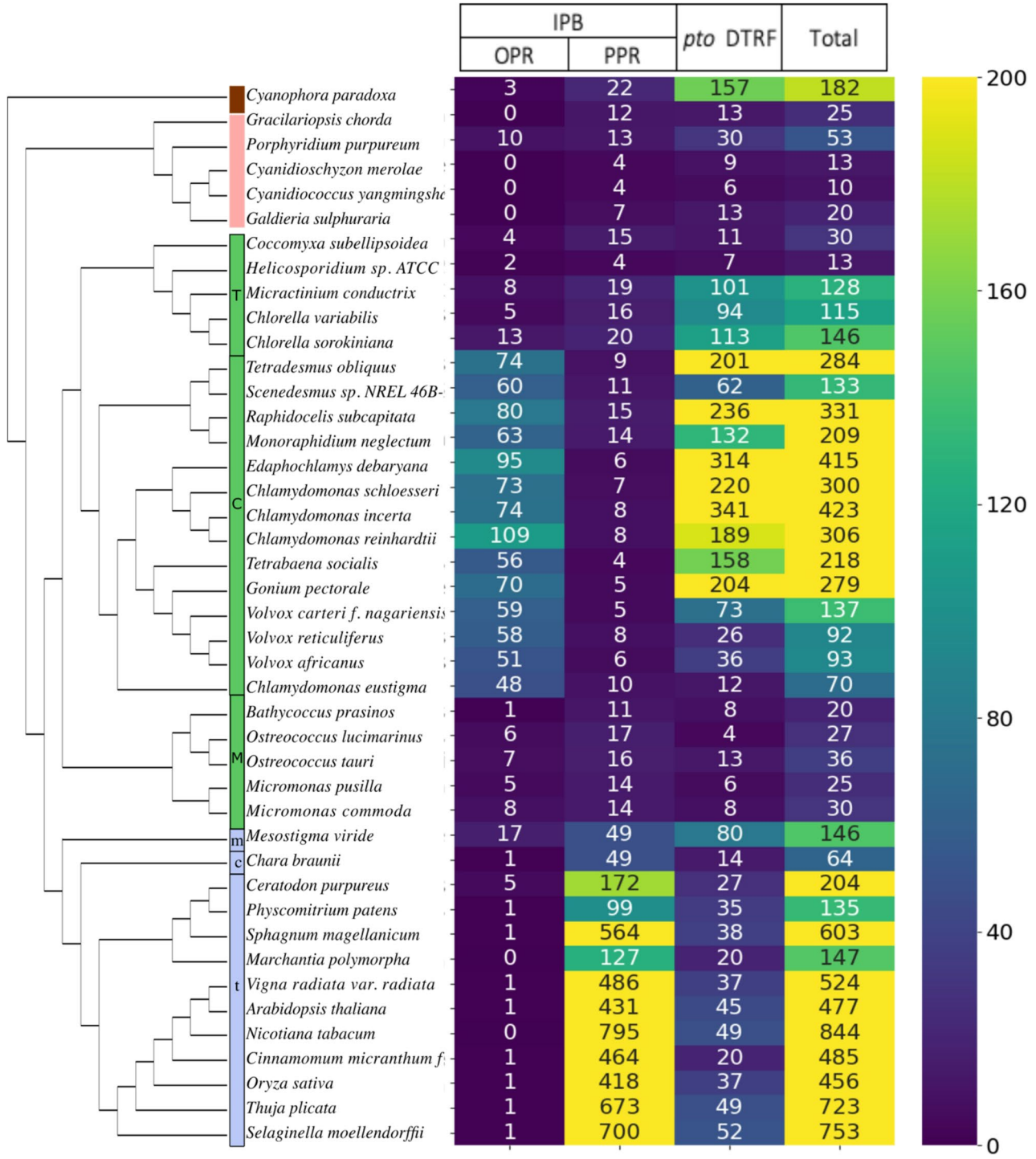


Fig. 2 Candidates across the 43 proteomes of Archaeplastida species detected by IPB (with at least 2 OPR or PPR motifs, whatever the number of predictions to be addressed to organelles) and new *pto* OTAF candidates detected by DT and RF, classified according to their NCBI taxonomy and reference phylogenies (see Methods). The last column indicates the total number of retrieved candidates (IPB+DTRF). The color column indicates the different phylum: Glaucophyta (brown), Rhodophyta (red), Chlorophyta (green) and Streptophyta (blue). Sub-clades are indicated within Chlorophyata (T: Trebouxiophyceae, C: Chlorophyceae, M: Mamiellophyceae) and Streptophyta (m: Mesostigmatales, c: Charophyceae, t: Tracheophyta). The whole list of candidates is available in Additional file 2: Table S4

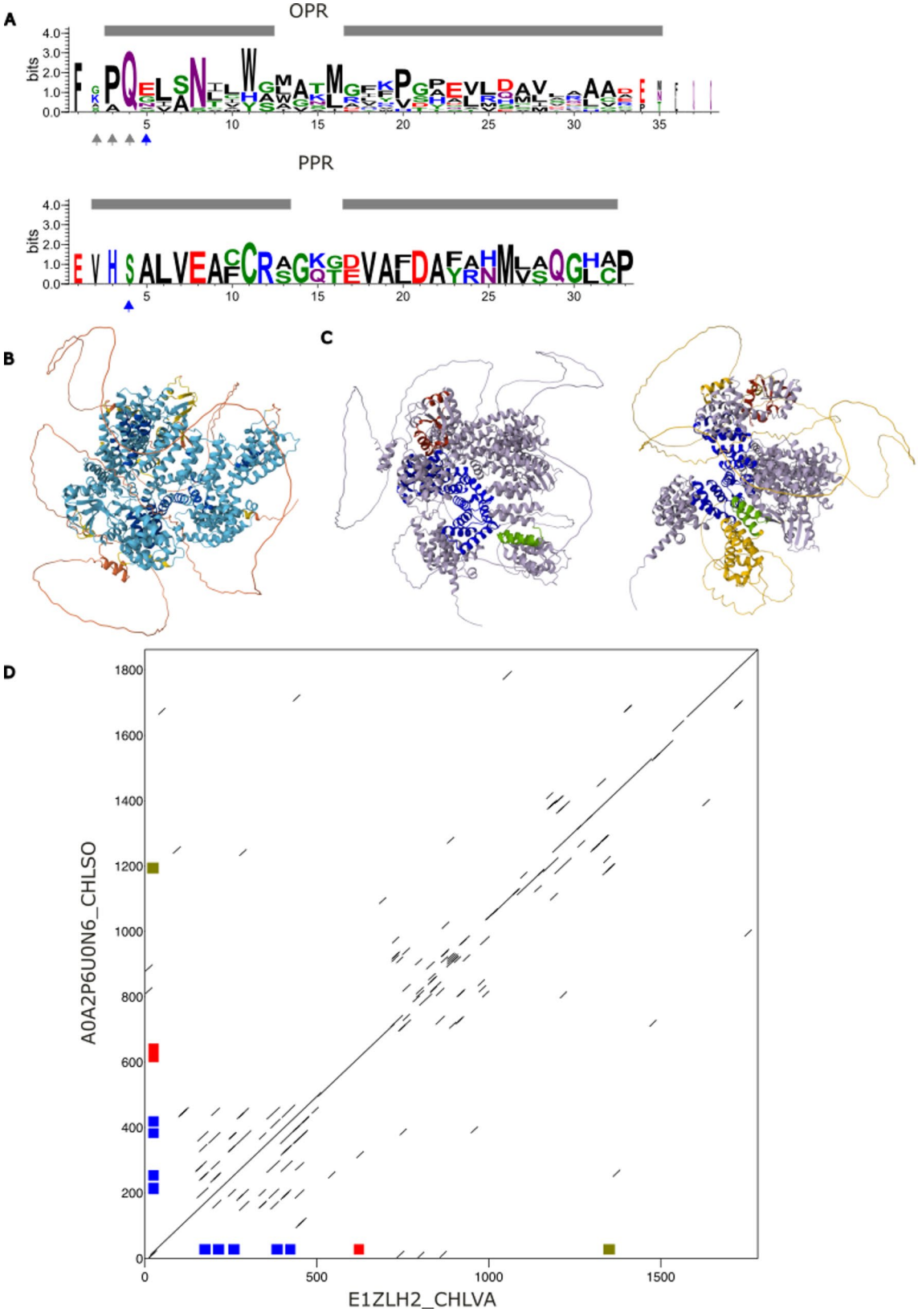


Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Structure of the composite OPR-RAP/PPR proteins. **A** Sequence logos of the 13 OPR motifs and 2 PPR motifs detected by IPB in the 3 detected orthologs. The positions expected to be determinant for RNA recognition are indicated in blue as in [7, 40]. Note that the 35th key position is absent from the PPR logo. For OPR, positions indicated in grey are also expected to contribute to RNA recognition. The two putative helices of the AF2pred. within the OPR and PPR sequence logos are shaded in grey. **B** AF2pred. of A0A2P6U0N6_CHLSO colored according to pLDDT quality score (predicted aligned error plot is given in Additional file 1: Figure S2B). **C** Different orientations of the structure in B. Approximate positions of repeats detected by IPB are in blue (OPR) and olive green (PPR) and in red for the PFAM RAP domain. The region between the RAP domain and the PPR repeat is colored in yellow in the left panel. **D** Dotplot of A0A2P6U0N6 against against E1ZLH2 (made with Dotmatcher at https://www.ebi.ac.uk/jdispatcher, windowsize 20, threshold 30). Approximate positions of the detected OPR and PPR motifs, and of the RAP domain are indicated as in **C**

candidates (*pto* DTRF) are in Chlorophyceae (ranging from 4 in *Ostreococcus lucimarinus* to 341 in *Chlamydomonas incerta*) (Fig. 2).

Potential composite OPR-RAP/PPR candidates

In *Chlorella sorokiniana and* in *Chlorella variabilis*, there are 2 *pto* orthologous candidates A0A2P6U0N6 _CHLSO and E1ZLH2_CHLVA (see below, *Origin of candidates OTAF*) in which IPB detected 4 OPR motifs as well as 1 PPR motif (Fig. 3). The sequence logos of the detected OPR and PPR motifs are given in Fig. 3A, in which the conserved residues of the canonical motifs are retrieved (see Fig. 7 for canonical motifs). The 3D model of A0A2P6U0N6 _CHLSO in Fig. 3B, shows two α-solenoid domains that are predicted with high confidence (Additional file 1: Figure S2B). The OPR repeats are part of the first one, followed by a RAP domain (in blue and red respectively in Fig. 3C). The detected PPR repeat is part of the second α-solenoid domain (green in Fig. 3C). The two domains are separated by a disordered region of approximately 400 amino acids (yellow region, Fig. 3C, left panel). There is a third ortholog in *Micractinium conductrix* for which the PPR repeat was not detected and 2 additional orthologs in *Chlorella ohadii* and *Chlorella vulgaris*, not considered in the present study, but part of the similarity cluster Uniref50_E1ZLH2 which group sequences sharing at least 50% identity (Additional file 2: Table S7). The domain organization is conserved in these 6 proteins, as exemplified by the dotplot of A0A2P6U0N6 _CHLSO and E1ZLH2_CHLVA (Fig. 3D). They would be the first described composite OPR-RAP/PPR candidates. Note that RADAR detected many more repeats in these candidates (around 20 in each) than were detected as belonging to OPR repeats. This again shows that some PPR/OPR motifs could have diverged upon recognition by profile similarity, compared to the motifs selected to initiate the IPB procedure (Additional file 1: Figure S2).

Origin of candidates OTAF

To date the dynamics of expansion and reduction of the PPR, OPR and *pto* DTRF candidates, they were clustered into protein families by MCL clustering (I = 50) on their pairwise similarity network (Additional file 1: Figure S3). In the present study, we choose to focus on families composed of proteins that did not undergo domain rearrangements since their common ancestor and hence probably conserved similar functions (sequence similarity hits over more than 70% of the length of both aligned proteins). For clarity, the clusters of candidates in the sole 12 representative proteomes are given in Fig. 4 and the complete clustering is provided in Additional file 2: Table S8. Each family was classified according to the distribution of its members into species-specific, phyla-specific or inter-phyla family (Fig. 4). There are 68% (6373/9350) of candidates clustered with at least another candidate, and the majority of them (5113) are distributed into 274 PPR families, 58 *pto* DTRF families and 37 OPR families of more than 3 members. There are 25/9 *pto* DTRF candidates clustered within PPR/OPR protein families, thus corresponding to remote PPR/OPR homologs, as mentioned earlier. In agreement with PPR frequency and the similarity network (Additional file 1: Figure S3), the three largest protein families comprise a total of 3080 PPR candidates in all 4 phyla studied. In contrast, the largest *pto* DTRF and OPR families are restricted to Chlorophytes, and are much smaller (78 and 32 homologs for *pto* DTRF and OPR respectively).

Compared to OPR and DTRF candidates, PPR are the most frequently clustered into inter-phyla families (57%) and into Streptophyta-specific families (33%). PPR would thus be the most ancient families, or at least the most conserved ones, with the lowest frequency of singletons (4,8%). At the opposite, there is only one OPR family with members in more than one phylum. The majority (89%) are either within Chlorophyta-specific families (52%) or as singletons in Chlorophyta (41%), suggesting that OPR emerged more recently or evolved faster than PPR.

There are 35 families of *pto* DTRF candidates with at least 5 members across the 43 studied proteomes (Table S7). 70% of *pto* DTRF candidates are singletons, suggesting that most of the families among DTRF candidates emerged more recently or evolved faster than both OPR and PPR (Additional file 2: Table S8). The largest families are restricted to Chlorophyta, but 3 large ones are shared between the phyla: one of 11 ankyrin repeats shared by 10 species in all phyla except in Rhodophyta, one of 9 peptidases shared by 8 species in all phyla but Glaucophyta and on of 8 TPR undetected by IPB in Chlorophyta and Rhodophyta. Higher inflation (I) values for MCL clustering would split the largest PPR families into several smaller ones, thus inverting the proportion of inter-phyla families (from 55% at I = 50 to 24% at I = 100)

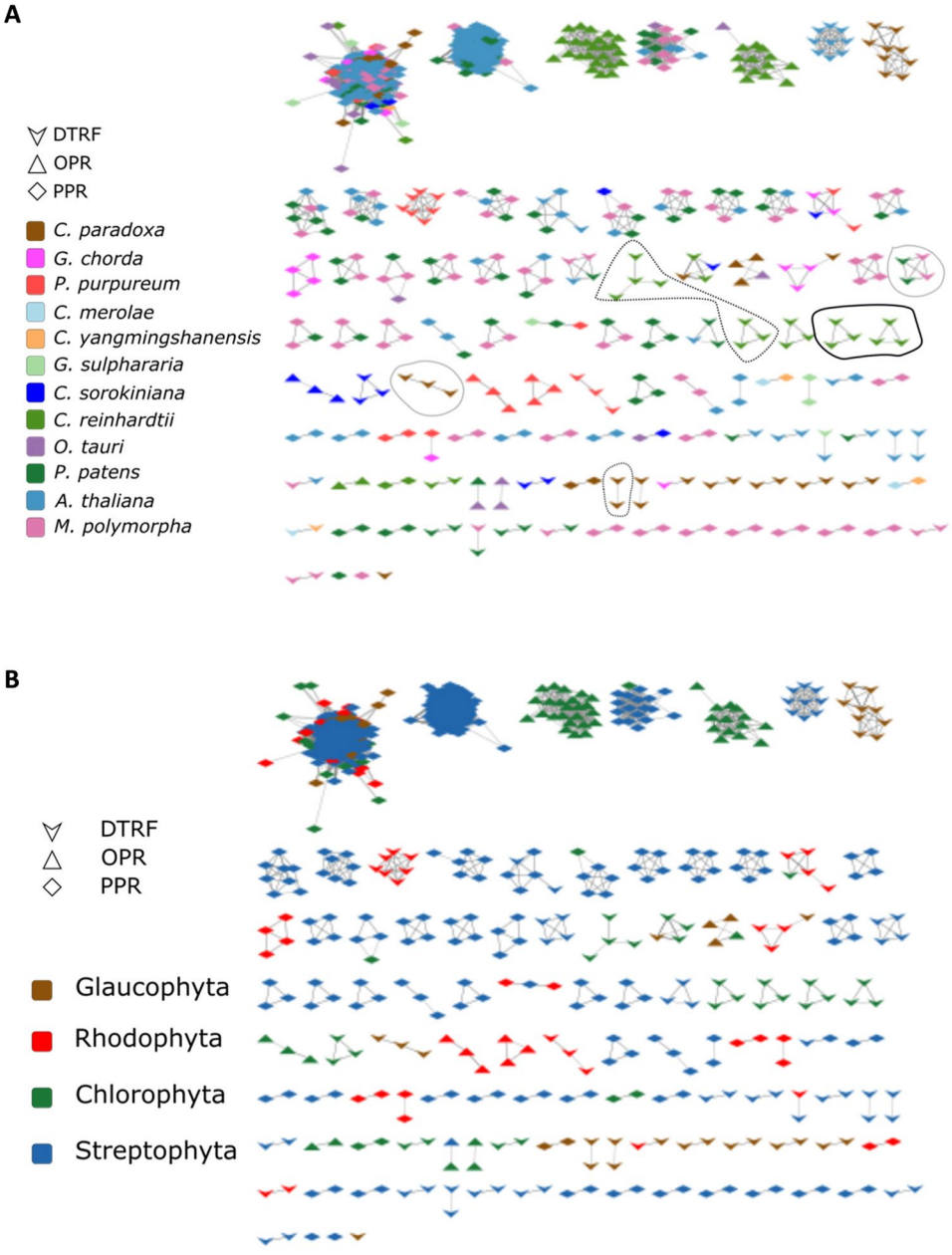


Fig. 4 **A** Clusters of IPB and *pto* DTRF candidates found in Glaucophytes and Rhodophytes, as well as in 3 representatives in both Streptophytes and Chlorophytes. *Pto* candidates are colored per species and their shape indicates the detection method (legend box). *Pto* candidates linked by grey edges are in the same cluster and share significant similarity (E-value < 10e-6) over at least 70% of their length. Clusters of *pto* DTRF candidates containing different type of repeats are surrounded: TPR repeats (PFAM domains, grey), Ankyrin repeats (dotted black, no PFAM domain but CATH superfamily) and Mynd-type clusters (black, no PFAM domain but CATH superfamily). **B** Same as in A, with candidates colored according to their taxonomical group

and phyla-specific families (from 33% at I = 50 to 59% at I = 100, Additional file 2: Table S8), but it has only very minor effect on OPR and *pto* DTRF clustering. In consequence, the general tendencies of the comparative analysis of PPR, OPR and *pto* DTRF families are robust: PPR are the most shared between phyla and have the fewest proportion of singletons.

Species-specific families, likely reflecting recent duplications or fast evolving candidates, are observed in all species, with the highest duplication rates in the moss *Selaginella moellendorffii* (Streptophyta) for PPR, in *Edaphochlamys debaryana* (Chlorophyta) for OPR and in *Cyanophora paradoxa* for *pto* DTRF candidates (Additional file 2: Table S8). There are 11 species-specific expansions of at least 5 *pto* DTRF paralogs found in each phylum. In those families, we note the high frequency of proteins containing Zinc fingers MYND-type (Table 5).

Table 5 Species-specific protein families

Cluster id	# *pto* DTRF paralogs	Phyla	Species	InterPro annotation
63	9	Glaucocystophyceae	*Cyanophora paradoxa*	na
99	7	Rhodophyta	*Porphyridium purpureum*	IPR000477/Reverse transcriptase domain
96	7	Streptophyta	*Arabidopsis thaliana*	IPR005630/Terpene synthase, metal-binding domain
51	10	Chlorophyta	*Edaphochlamys debaryana*	na
178	5	Chlorophyta	*Edaphochlamys debaryana*	IPR002893/Zinc finger, MYND-type
179	5	Chlorophyta	*Edaphochlamys debaryana*	na
180	5	Chlorophyta	*Edaphochlamys debaryana*	IPR002893/Zinc finger, MYND-type
50	10	Chlorophyta	*Tetrabaena socialis*	IPR002893/Zinc finger, MYND-type
175	5	Chlorophyta	*Tetrabaena socialis*	IPR002893/Zinc finger, MYND-type
176	5	Chlorophyta	*Tetrabaena socialis*	IPR002893/Zinc finger, MYND-type
177	5	Chlorophyta	*Gonium pectorale*	IPR002893/Zinc finger, MYND-type

Note that Zinc fingers MYND-type domains have been found only in Chlorophyta (Additional file 2: Table S6).

There is a species-specific expansion of interest in *C. reinhardtii*. It contains 7 paralogs tandemly duplicated on chromosome 10, comprised of 2 *pto* DTRF candidates, one other *pto* DTRF candidate and 4 non-*pto* RF candidates (Fig. 5A). The 7 paralogs in that duplicated region have an average pairwise identity of 68% over at least 85% of their sequence and adopt an α-solenoid fold (Fig. 5B). No PFAM domain is detected within these candidates. There are also 3 homologs of that family in *C. incerta* and 2 homologs in *C. schloesseri* (cluster 94, Additional file 2: Table S8.1). In each of these two genomes, the paralogs are not tandemly duplicated, in agreement with an expansion restricted to *C. reinhardtii*.

Molecular and physico-chemical properties of OTAFs candidates

Number of repeats

The number of OPR and PPR motifs detected by IPB is given in Fig. 6A. Half of the PPR candidates found by IPB have more than 7 motifs and in *C. reinhardtii*, the number of OPR motifs per candidates follows a bimodal distribution with a pic at 2 and the other at 7 motifs per proteins. As already mentioned, some OPR and PPR motifs diverged upon recognition but are detected by RADAR (Additional file 1: Figure S2). This explains why the PPR motif per protein detected by IPB is lower than the median number (9–10) of PPR motifs described among P-type PPR proteins found in land plants [13], and of course lower than the median number of PPR motifs [12] among all PPR UWA.

The number of repeats detected either by IBP or RADAR in each type of OTAF candidates among the 4 phyla is shown in Fig. 6B. The number of OPR and PPR repeats per *pto* candidates is similar in each phylum. Inversely, the number of repeats$_{RADAR}$ in the *pto* DTRF candidates vary between phylae, with the highest number being in Chlorophyta. This suggests, in line with the duplication/loss rates described above (Additional file 2: Table S8) a higher diversity of protein families and a higher propension to form repeats in Chlorophyta.

Clustering of OPR and PPR motifs

The similarity networks between OPR and PPR motifs found in the 12 representative proteomes is shown in Additional file 1: Figure S3A. While most of the PPR motifs form a unique network, OPR form one big network with distinct sub-networks by species and 6 smaller networks of *C. reinhardtii* OPR motifs. Motifs were clustered (see Methods) and classified as species-specific, phyla-specific or shared among different phyla (Additional file 2: Table S9). Motifs within a cluster$_{motifs}$ share on average 55% identity while motifs between two different clusters$_{motifs}$ share on average 0.1% identity (Additional file 1: Figure S4B).

Most of the cluster $_{motifs}$ are phyla-specific. In agreement with the distribution of the protein families, PPR clusters$_{motifs}$ are mostly found in Streptophyta proteins and OPR clusters$_{motifs}$ in Chlorophyta proteins. Species-specific clusters and singletons also exist (36% for PPR motifs and 12% for OPR motifs). This suggests a heterogeneity of evolutionary rates among clusters. PPR motifs are globally more conserved, but there are also rapidly evolving PPR motifs within Streptophyta species and rapidly evolving OPR motifs in Chlorophyta. Fewer inter-phyla cluster$_{motifs}$ are found for PPR and OPR (ca. 4% and 3%, respectively). The existence of PPR/OPR clusters $_{motifs}$ with no *C. reinhardtii*/*A. thaliana* members confirms the ability of IPB to retrieve motifs diverged from the ones used for the first iteration step.

On average, all the motifs within a given cluster$_{motifs}$ belong to different proteins. In other words, the different motifs that compose an OPR/PPR protein are part of different clusters$_{motifs}$. Since our study spans 43 proteomes, this could reveal greater similarity between homologous motifs in homologous OPR/PPR than between the motifs that make up a given protein. Alternatively, this could reveal a positional bias of the different clusters$_{motifs}$ along the protein. To investigate this hypothesis,

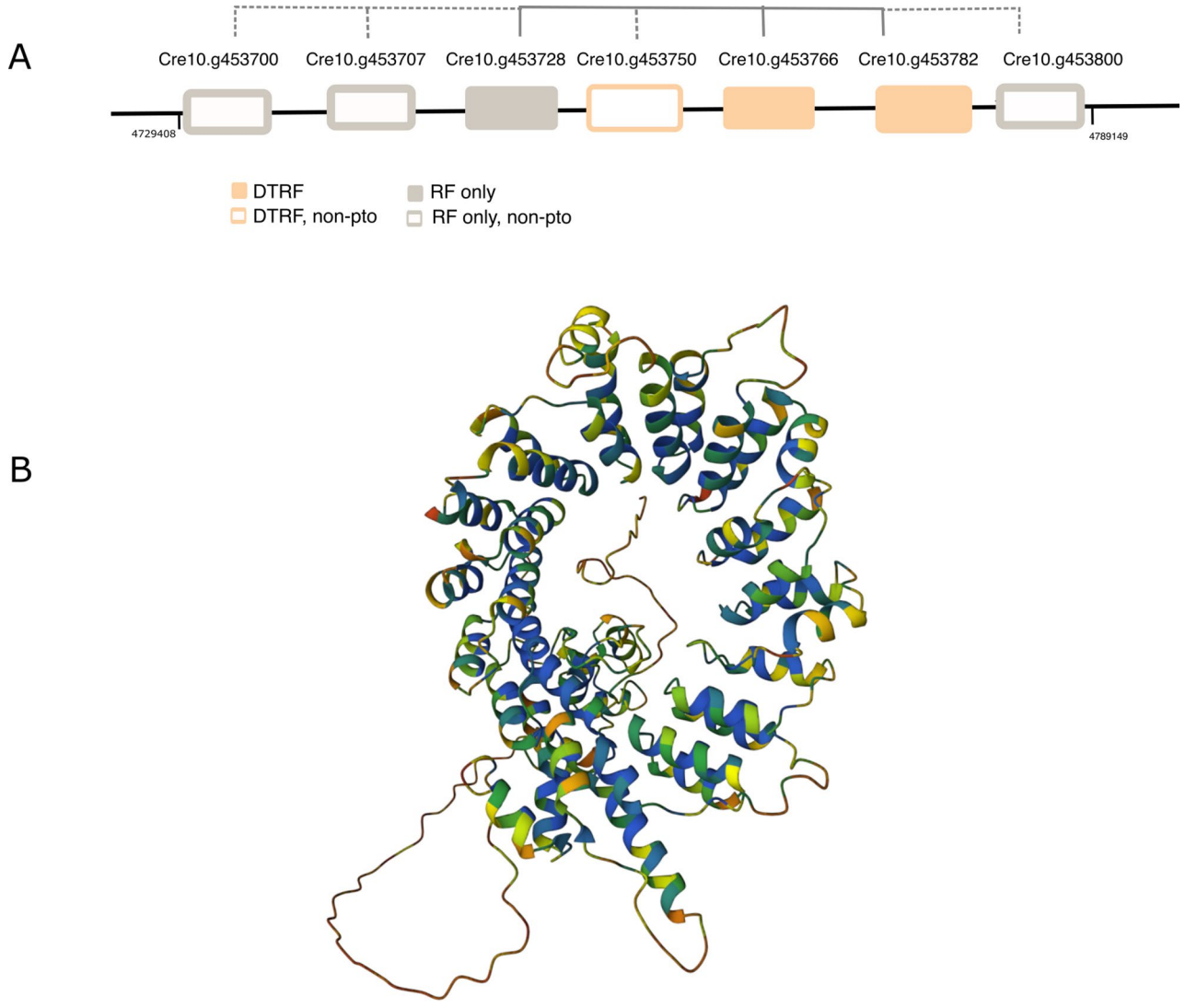


Fig. 5 Tandems of DTRF candidates on chromosome 10 of *C. reinhardtii*. **A** Solid grey lines relate proteins in the same cluster of *pto* candidates (Fig. 4). Dashed grey lines relate non-*pto* to *pto* candidates. **B** AF2pred of Cre10.g453800. All proteins are classified in the CATH Superfamily 1.25.40.20 termed ankyrin repeat-containing domain (InterPro IPR036770) but they do not contain any PFAM Ankyrin or ankyrin-related repeats

we determined the order of appearance of each cluster$_{motifs}$ along the proteins (repeat relative position). For this analysis, a MCL clustering was performed on the similarity network of the 706 OPR motifs and the 3346 PPR motifs found in *C. reinhardtii* and *A. thaliana* only, to exclude the orthologs of the 41 other species. For each cluster, only one representative among the different paralogs of a given protein family were considered. The left panel of Fig. 7A shows the heatmaps of the frequency distribution of clusters$_{motifs}$ (>4 motifs) along PPR proteins. For example, the PPR motifs of cluster$_{motif}$ 1 are found as the first repeat of PPR proteins less than 10% of the time, and they appear at similar frequencies at other relative positions along the protein. Thus, the frequency distribution of cluster$_{motif}$1 is not biased. However, three PPR cluster$_{motifs}$ present a strong positional bias (at least 80% of the motifs at a given relative position) along PPR candidates. PPR motifs from cluster 12 and 18 are mostly found as the first repeat and from cluster 29 as the second repeat along PPR proteins. The relevant motifs logos are shown on the right panel of Fig. 7A. The logo of all PPR motifs described at UWA (UWA logo) is given at the top. It is mostly similar to the logo of the P motif, which is the most frequent PPR motif. The strong similarity of the logo for all motifs found by IPB to the UWA logo confirms the performance of IPB. Interestingly, the conserved Glycine at the 15th position of the UWA logo is substituted in the clusters 12 and 29 with positional bias: (G-> S) within cluster$_{motifs}$ 12, biased at the first position and (G-> K) for cluster$_{motifs}$ 29. Cluster$_{motifs}$ 29 also have a substitution (G-> Y) at position 31 of the UWA logo. Position 15th lies within the linker between

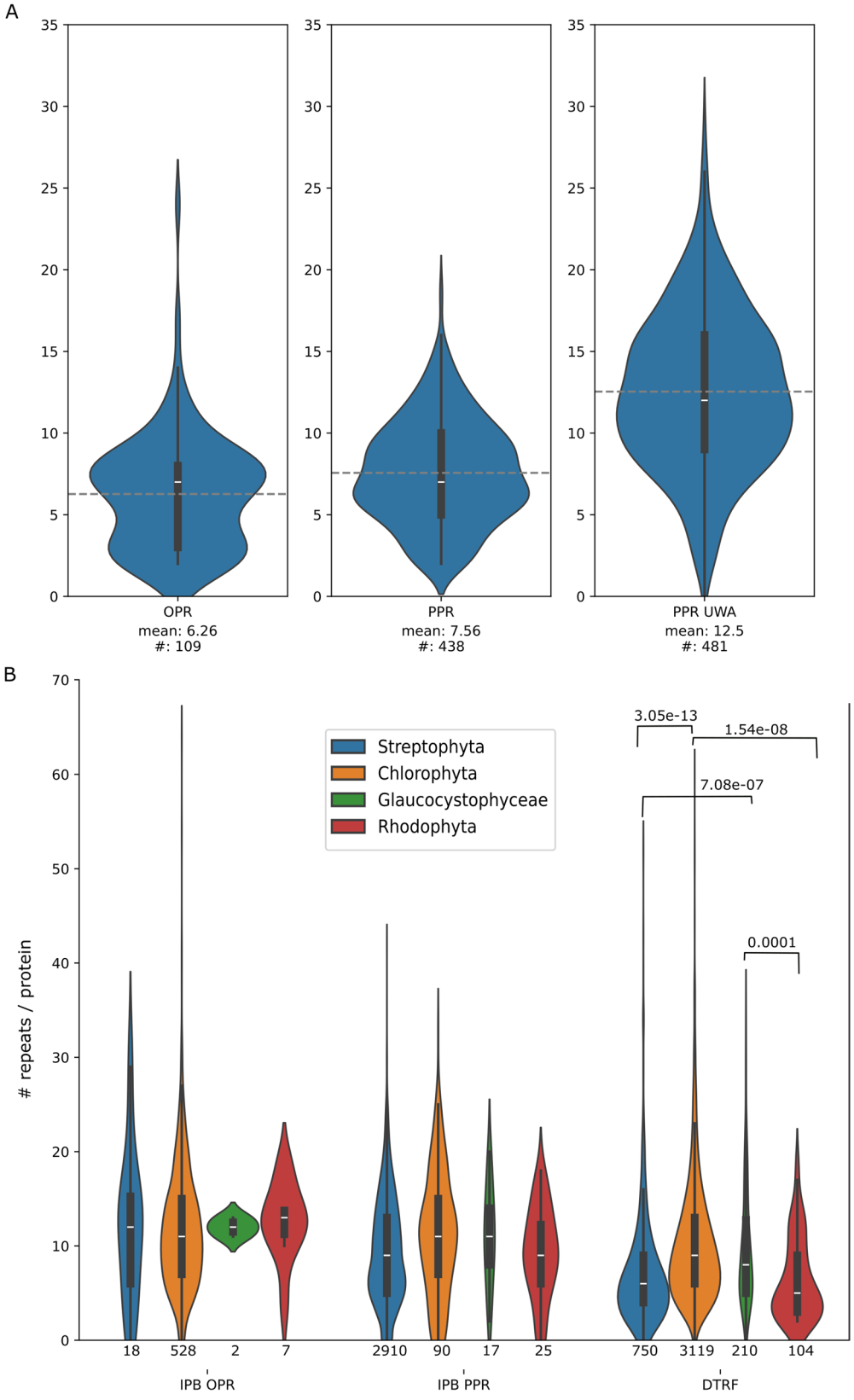


Fig. 6 **A** Distribution of motifs. Left and middle panels: Distribution of the number of OPR/PPR motifs in IPB candidates with at least 2 detected motifs in *A. thaliana* and *C. reinhardtii*. Right panel: Distribution of the number of PPR motifs per PPR UWA. **B** Distribution of repeats per protein detected by RADAR for the IPB and pto DTRF candidates in the 4 taxons studied. The number of *pto* candidates is indicated below and significant p-values (post-hoc Tukey test, p-value < 0.05) above the violinplots

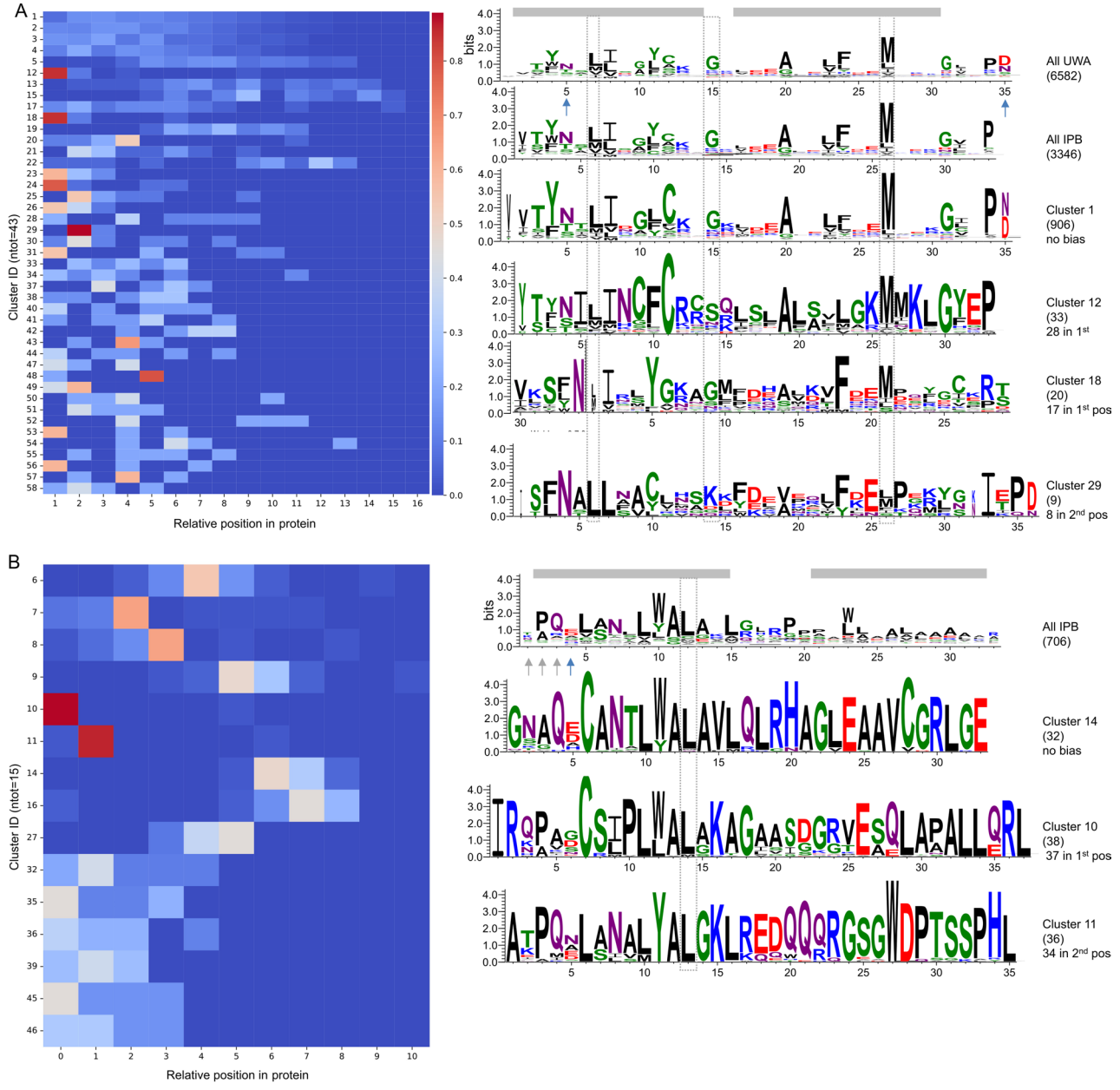


Fig. 7 Distribution of cluster$_{motifs}$. **A** PPR motifs. Left panel, heatmap of the frequencies of each cluster$_{motifs}$ (motifs > 4) according to their relative position along the PPR protein. Right panel, logos of remarkable cluster$_{motifs}$. On top, logos of PPR motifs described at UWA and all PPR motifs found by IPB. The amino acid positions are those obtained from the multiple alignment and for PPR motifs at UWA it coincides to those proposed based on structural data in [41]. All logos are aligned on the Methionine at position 27 of PPR motifs at UWA. Conserved residues important for the structure of the motifs are boxed in dotted grey. The N-terminus of first helix of cluster$_{motif}$ 18 (position 30 to 35) was found after the second helix in the logo. In the figure, those residues have been moved at the beginning of the motif. As in [7], the two helices of the PPR motifs are shaded in grey and the positions shown to be the primary determinants of RNA-binding specificity are indicated by arrows. A selection of conserved positions in the three regions of the PPR motif are boxed in dotted dark. For each logo, the number of motifs in the cluster is indicated in parenthesis. When the distribution is biased, the occurrence of motif at the biased position is given. Amino acids are colored according to their chemistry (green, polar; purple, neutral; blue, basic; black, hydrophobic). There are on average 9, 9, 5 and 8 PPR motifs per proteins in clusters 1, 12, 29 and 18 respectively. **B** Same for OPR motifs. All logos are aligned on position 12 of the logo of all detected OPR motifs, boxed in dotted grey. The two putative helices of the AF2p OPR sequence logo are shaded in grey and the position expected to be the primary determinant of RNA-binding specificity as in [40] is indicated by a blue arrow. The three positions immediately upstream, indicated by grey arrows, are also likely contributing to the RNA recognition. There is are on average 8 OPR motifs per proteins in clusters 6, 10 and 11

the two α-helices and position 31 lies just after the second α-helix. These substitutions might thus provide different flexible properties to the repeat. Clusters$_{motif}$ with positional biases have also more hydrophobic residues in the first part of the second helix that might also provide different structural conformations, like the higher global hydrophobicity of the first positions within cluster$_{motifs}$ 12. Also note that cluster 29 shows no variability in the residue (N only) at the first key positions for mRNA binding specificity, suggesting that these biased motifs repeat may binds to a Cytosine depending on which sequence position is considered as the second key residue [7]. Also note that the biased clusters do not match the sequence logo of the variant of the PPR motif L and S [7]. Interestingly, the linker region of the consensus PPR motif from residue 17 up to the residue 30 in the second helix contains mainly hydrophobic amino acids, with 5 negatively charged ones evenly distributed. The negative charge of this region is completely lost in cluster$_{motif}$ 12, in which the corresponding positions contains either positively charged residues (two) or neutral ones (three residues). In cluster$_{motif}$ 18, the negative charge also decreases, as positively charged residues at position 13 and 16 are the most frequently observed.

Figure 7B shows the heatmap and logos for OPR motifs. The motifs in Cluster$_{motif}$ 10 are mostly found as the first repeat in OPR proteins and cluster 11 are mostly found as the second repeat. The linker region in the consensus OPR motif (positions 16 to 20) is overall positively charged (+1), as well as in the cluster with no positional bias (+2), while in the cluster with positional bias towards the first or the second position within the proteins, the overall charge becomes negative (-1). Also, in the cluster$_{motif}$ 11 biased towards the second position, the overall hydrophobicity is lost.

Physico-chemical properties

All OTAF candidates were described with the variables used in the RF procedure. Figure 8 shows the principal component analysis (PCA) of the 1551 candidates within the set of selected 12 proteomes used to evaluate the quality of AF2pred. of *pto* DTRF candidates. Overall, candidates of a given phyla are close together in the two-dimensional space defined by the first two principal components (PC1) and PC2 (Fig. 8A). They differ mainly along the PC1 axis, with OPR proteins on the right (Fig. 8C) and PPR proteins on the left (Fig. 8D). OPR and PPR thus differ mainly by their hydrophobic and steric properties (Additional file 1: Figure S5), which is also visible on their sequence logos (Fig. 7). Some DTRF candidates co-localize in the same area as OPR/PPR candidates retrieved by IPB, indicating that the similarity-free detection procedures succeed at identifying the defined properties of known OTAFs (Fig. 8B). Other *pto* DTRF candidates explore a wider space through the upper-right part of the graph, away from OPR and PPR, indicating higher amphipathic properties (Additional file 1: Figure S5), that could also reflect contrasting structural properties and thus functions.

Discussion

Critical assessment of IPB

Our iterative similarity-based procedure (IPB) efficiently retrieves the OPR (including their HPR subset) and PPR proteins previously identified in *C. reinhardtii* and *A. thaliana*. IPB even identified new remote homologs in these two species, seven of which have been annotated as new OPR [36] in version 6.1 of *C. reinhardtii* [42]. Starting from motifs found in a single species, the iterative process allowed us to capture specific signatures in diverse taxonomic groups to identify OPR and PPR homologs across Archaeplastida species. On average we found fewer PPR motifs by PPR candidates than those reported in the PPR database at the University of Western Australia (UWA). This discrepancy is mainly due to the fact that we used a very limited number of PPR P-motif in IPB. In contrast, PPRFinder, used to establish the PPR UWA catalog [17], searches a dozen of profiles of PPR sub-motifs that were expertly defined combining analyses of both sequence and structural similarity based on the analysis of 41 land plant proteomes [13]. Despite its simplicity, IPB proved to be an efficient approach for detection of conserved motif at a broad taxonomy scale, with no a priori knowledge on the motif diversity and its tridimensional structure. Its sensibility may be improved by increasing the number of motifs used to build the original profiles; by evaluating the quality of multiple alignments obtained by several methods with metrics as proposed in [43] and also by evaluating the quality of the similarity clusters obtained by alternatives to MCL for clustering (such as the Louvain algorithm).

Critical assessment of DT and RF

The DT procedure is very specific but poorly sensitive (0.34), which explains why it failed to detect several of the known OTAFs. In contrast to DT, the RF procedure is very sensitive but lacks specificity on the real datasets. However, it was able to retrieve the majority of mTERF proteins in *A. thaliana*, confirming its ability to detect other types of α-solenoid OTAFs than OPR and PPR. RF could be improved by using additional tools for single residue 2D predictions such as NetSurfP [44] for repeat *ab initio* detection and α–solenoid structures. As with any supervised classifier, RF candidates are strongly determined by the training set. There are also many possibilities to improve the RF procedure, starting by testing other protein descriptors. Some attempts to develop machine learning procedures have been already made

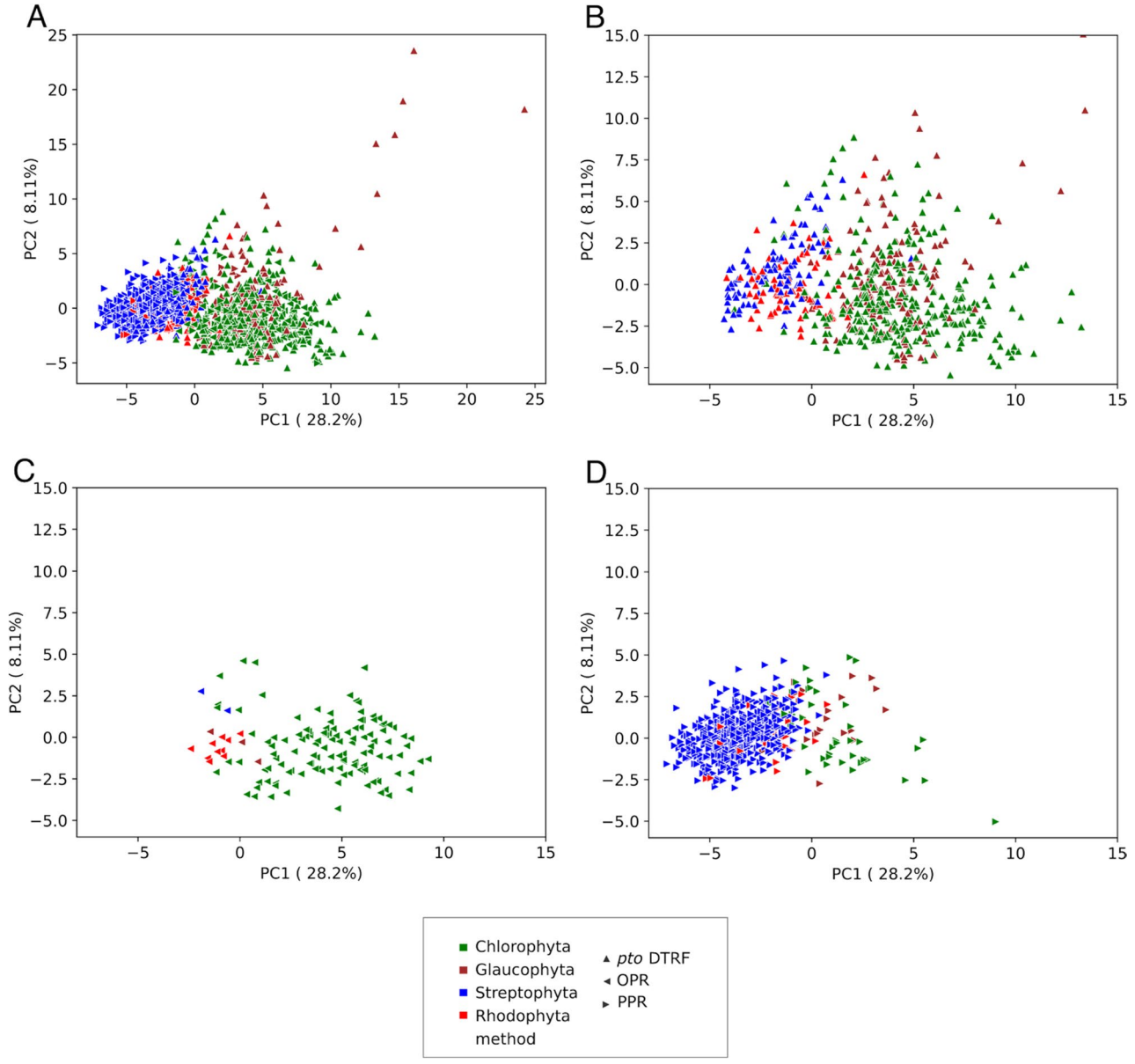


Fig. 8 Principal component analysis of candidates with at least 2 OPR/PPR motifs and *pto* DTRF candidates in 12 representative proteomes of Archaeplastida described by properties used by the RF procedure. **A** All candidates along principal components 1 and 2. **B** *pto* DTRF candidates (same PCA as in A). **C** OPR candidates (same PCA as in A). **D** PPR candidates (same PCA as in A). Correlation circles are given in Additional file 1: Additional file 1: Figure S5

for PPR discovery, based on protein features extracted from PPR as a positive training set and non-PPR as negative training sets as in [45–47] but the performance of these methods has been estimated only on the training sets: they were never used to search for PPR in complete proteomes, and therefore it is not known how they compare to profile PPR-based methods in such tasks and how they cope with taxonomical bias. Indeed, deep learning approaches using natural language processing models and neural networks models should also be tested. However, when taken together, DT and RF identified new *pto* α–solenoid candidates (DTRF), of yet unknown OTAFs families. We showed that hydrophobic and steric constraints on the same face of an α-helix (Additional file 1: Figure S1) have a crucial role for the RF classification. Such properties could reflect amphipathic constraints and hence the amphipathic nature of the helices, necessary to allow the formation of a hydrophilic cavity within the α-solenoid structure, in which the positively charged mRNA could bind. Based on our estimation 66% of *pto* DTRF candidates have an actual α-solenoid shape. Note that this evaluation could be automated in the future by using pipelines like the recently developed

SOLeNNoID [48] to detect solenoid residues into protein 3D structures.

Completed catalog of OPR and PPR across Archaeplastida

PPR are the most conserved OTAFs between phyla and have the fewest proportion of singletons, compared to OPR and *pto* DTRF candidates. Massive expansions of PPR candidates occurred in Tracheophyta, which is probably linked to the expansion of mRNA editing in this lineage among land plants [17]. The relative paucity of PPR in green microalgae could be associated to the absence of RNA editing in Chlamydomonas [49], as well as, to our knowledge, in Rhodophyta and Glaucophyta. If such events were to be identified in those clades, they would be mediated by players other than PPRs, in agreement with the proposal that eukaryotic editing was probably acquired multiple times from ancestral bacterial toxin deaminase [50]. IBP allowed us to provide the first extensive catalog of OPR protein across Archaeplastida. OPR are mainly found in Chlorophyta, with a few members in Streptophyta, but also in Rhodphyta and Glaucophyta, confirming the trends already found by comparing micro-algae and land plants. The HPR described in [35] are recognized by IPB as harboring OPR motifs, confirming that HPR proteins belong to the OPR family (Additional file 2: Table S3a).

A potential bias in the distribution of the repeated motifs along the OPR and PPR proteins

The clustering of the motifs provided by IPB also pointed to apparent bias in the distribution of the clusters of PPR and OPR motifs along a protein sequence, with no periodicity of occurrence along the protein, such as that seen for the PLS subfamily of PPR proteins, composed of repetitions of P, L and S consecutives motifs [13]. The analysis of the sequence logo of the biased OPR and PPR motifs shows that the overall charge of the linker regions in those motifs (and the second helix for PPR) varies compared to the consensus motif. To our knowledge, such a bias has not been reported previously. It calls for further investigations to assess the contribution of such bias for particular position in mRNA target recognition and/or protein structural properties, such as folding and flexibility. As the linker region lies on the external surface of the α-solenoid, this could play a role for binding of other protein partners.

Have composite OPR/PPR proteins appeared in an ancestral Trebouxiophyceae?

Thanks to IPB, we identified 5 orthologous proteins composed of both an OPR-RAP and a PPR α-solenoid domain in Trebouxiophyceae. The fact that those proteins were identified in 5 different species makes the probability of 5 independent sequencing or assembly error quite low. Such a composite protein has never been observed before in OTAFs. If they arose by a chromosomal rearrangement in an ancestral Trebouxiophyceae, their conservation suggests that they could have been selected during evolution. Further investigations will be performed to determine if they are real proteins and, if so, to determine their function.

New α-solenoid OTAFs candidates available for functional characterization across Archaeplastida

We report on average *ca*. one hundred of new α-solenoid OTAFs candidates in Chlorophyta and in the Glaucophyta *Cyanophora paradoxa*, and fewer in Streptophyta and Rhodophyta, with 29 candidates in *A. thaliana* and 19 in *Porphyridium purpureum*. We thus provide this -most probably incomplete- tentative list of new OTAF candidates, for further experimental characterization. As expected, among the *pto* DTRF candidates, some display conserved domains typically defining α–solenoid proteins like the mTERF regulators of plastid gene expression [51–53], but also armadillo and ankyrin repeats, some of which have been found in land plant chloroplasts, likely being involved in plastid gene expression. This is the case for the ankyrin repeat protein Akrp in *A. thaliana* that blocks chloroplast differentiation [54] and an armadillo/β-catenine protein in spinach, found in the chloroplast nucleoids which might be involved in the structural maintenance of chromosomes [55]. Interestingly, a significant number of *pto* DTRF candidates of Chlorophyta, have Zinc finger MYND-type domains, which predominantly mediate protein-protein interactions and are found in transcriptional regulators as well as in DNA repair proteins [56]. RNA is not reported as the main interactant of all the cited domains, preferentially binding to DNA or protein. However, some mTERF proteins have been characterized as being involved in RNA splicing of introns and in one instance in transcript stability [57–59]. Note that there has not been yet conclusive experimental evidence that RNA phenotypes result from the loss of a direct interaction between the target messenger RNA and an MTERF protein [52, 57, 58], at variance with the experimental documentation of messenger RNA and OPR/PPR protein interactions, see for example [60, 61].

Also, many DTRF candidates contain intrinsically disordered regions, which interestingly have been found to contribute to single stranded RNA binding of Zinc finger domains, although of another type [62]. Many other *pto* DTRF candidates have no conserved PFAM domain, but are classified into the CATH Superfamily database, based on domains found in proteins with defined structure, which allows to find more remote similarity. This is the case of the cluster of 7 tandemly duplicated paralogs found in *C. reinhardtii* (Fig. 7), containing domains from

the superfamilies of ankyrin and pseudo-ankyrin repeats. This cluster might represent a reservoir of new candidates and its experimental characterization is ongoing, by looking for a possible photosynthetic phenotype of the corresponding mutants in *C. reinhardtii*.

The exploration of α-solenoid OTAFs candidates outside Archaeplastida made possible

Little is known as to how the expression of the organellar genomes is regulated outside of Opistokhonta and Viridiplantae (land plants and green algae). Our analysis shows the contrast in the global distribution of putative α-solenoid OTAF proteins, with far fewer OTAFs, either known (OPR/PPR) or unknown (*pto* DTRF) in Rhodophyta and Glaucophyta than in Viridiplantae, suggesting that modes of regulation may vary, possibly with a different balance between transcriptional and post-transcriptional steps in the corresponding organelles. In particular, the RF approach should be instrumental for studies of the largely unexplored chloroplast biology of Rhodophyta and Glaucophyta, but also in other types of photosynthetic eukaryotes, such diatoms or dinoflagellates, which result from secondary or even tertiary endosymbiosis, paving the way for elucidating the rules of the regulation of gene expression in organelle genomes in these complex plastids.

Materials and methods

Sequence data

The 43 proteomes from Archaeplastida species were retrieved either at the NCBI, Uniprot or JGI websites. The details of the proteomes and their versions are given in Additional file 2: Table S1. For proteomes downloaded at JGI, the file of the filtered models was chosen Files/Annotation/Filtered Models ("best")/Proteins. Quality and completeness of proteomes were assessed with BUSCO 5.3.2 [63] (Additional file 2: Table S1), with protein mode on the Eukaryota and Viridiplantae lineages (eukaryota_odb10, viridiplantae_odb10, 2020-09-10).

As starting points for the motif-based procedure (IPB), we used published and described repeat motifs from 12 known OPR proteins from *C. reinhardtii* (green microalga model species) and from 11 known PPR proteins from *Arabidopsis thaliana* (land plant model species), listed in Additional file 2: TableS2.

Annotations, structural predictions

Proteins annotations were retrieved at Uniprot for all proteins. For *C. reinhardtii*, the annotation v5.6 available at Phytozome, version 13 was also used, as well as the TAIR10 annotation for *A. thaliana*. 3D structure predictions were retrieved at the AlphaFold Protein Structure Database or estimated locally with AlphaFold2 [64]. We searched for PFAM domains using InterProScan (v5.75-106.0) [65] against PFAM 37.4 in InterPro 106.0. For each candidate, if a PFAM domain is detected, the IPR annotation is provided. Secondary structure assignment from PDB coordinates was made with STRIDE [66].

Sublocalization prediction

Deeploc v2.0 [67], LOCALIZER v1.0.4 [68], TargetP-2.0 [69] and WoLFPSort v0.2 [70] were used to predict the chloroplast or mitochondrial localization.

Reference phylogeny

The 43 species from Archaeplastida were selected based on proteome status as provided by the Published Plant Genomes (PPG) database (https://www.plabipd.de/) on June 2020. Reference phylogeny was inferred from PPG and references [71–79], with the help of the NCBI Taxonomy tool.

Motifs-based similarity procedure

The IPB procedure follows the four steps described below, that we ran iteratively either on OPR or on PPR motifs.

Step 1. Pairwise comparison of motifs

First, an all-against-all pairwise comparison of all motifs is performed by BLASTP v2.6.0 [80]. We kept only the pairs of motifs whose E-values were lower than a threshold t varying according to X the size of the data set, as follows: if $10^{n-1} \leq X \leq 10^n$, then $t = 10^{-n}$.

Step2. Motifs clustering

Second, motifs are clustered with MCL v.14–137 [81] based on the $-log(E\text{–}value)$ of the BLASTP hits. Clusters are computed for a series of inflation parameter I values, starting from 1.1 and by increasing it by steps of 0.1 until. For each I parameter tested, a multiple alignment of the motifs in a given cluster is performed with MAFFT v7.450 [82]. We choose the clustering with the I value providing a maximum number of clusters with at most 18 and 16 gaps in their multiple alignment, for OPR and PPR motifs respectively. These values correspond to half of an OPR/ PPR motif as we wished to detect motifs that were similar to the starting motifs over at least half of their sequence and avoid "motif slippage".

Step3. HMM profiles search against proteomes of interest

Third, profiles are built with *hmmbuild* from the HMMer suite, v3.1b2 [83]. Each HMM profile is then used to search against each of the 43 proteomes with *hmmsearch*. The length of an OPR/PPR motif being about 35/38 amino acids, only newly found sequences whose lengths are −6/+2 the length of an OPR motif, i.e. 32/40 amino acids long are kept. Two motifs overlapping by more than 80% are considered identical and the motif boundaries

are defined by the minimum overlapping region to reduce the tendency of the motifs from slipping in one side or the other. Two motifs overlapping by less than 20% are considered different and kept without changing their boundaries. Pairs of new overlapping motifs between 20% and 80% are both eliminated, being considered as no longer representative of the typical starting motif.

Step5. Selection of candidates on targeting properties

The procedure is iterated restarting from step 1 after adding the new motifs to the initial data set only if at least 1 new *pto* candidate, i.e. predicted to be addressed by at least two prediction softwares is retrieved. See "prediction of localization" paragraph for details of the prediction.

Step6. Selection of the candidate proteins

The final step consists in selecting only *pto* proteins, i.e. predicted to be addressed either to chloroplast or mitochondria by at least two prediction algorithms (see *Sub-localization prediction*).

Decision tree (DT) procedure

First of all, proteins having a predicted transmembrane helix after their N-terminal part (corresponding to the targeting peptide and that could erroneously be predicted as a transmembrane helix) were removed. TMHMM v2.0c [84] was used to predict transmembrane helix. The end of the N-terminal part was defined as the 78^{th} residue.

The DT procedure is given in Fig. 1, right panel. Parameters for selection along the decision tree have been defined based on a set of validated α-solenoid protein and a set of validated non α-solenoid proteins (see the paragraph *Training set composition*). Additional file 2: Table S10 recapitulates numbers of retrieved candidates, precision, accuracy and recall for all tested combinations of parameter values. We choose the global combination maximizing precision and recall.

Filtering of proteins according to the number of sequence repeats

We selected the proteins in which at least one repeat of at least 29 amino acids is detected by RADAR v1.3 [85]. The value of 29 amino acids gave the best precision score among a series of tested values, ranging from 20 to 40 (Additional file 2: Table S10).

Detection of α helix pairs

The presence of α-helix pairs was estimated with ARD2 [30], which uses a neural network procedure with a training set containing X-ray resolved α-helix pairs and gives a score for each amino acid position to be a linker between two αhelices.

In order to keep only proteins with multiple pairs of α-helices, we constructed and tested dozens of sets of criteria based on the distance between 2 α-helix pairs and the number of linkers between the helices forming a pair (Additional file 2: Table S10). We kept only the proteins with at least 4 linkers, with the distance between two consecutive linkers of between 32 and 400 amino acids. A given amino acid is considered as a linker if its ARD2 score is above 0.15, based on the results obtained on the training set.

Protein filtering according to 2D structure predictions

We kept only those proteins for which 65% of the amino acids were predicted in α-helix by S2D v2 [86], between the two most extreme linkers predicted by ARD2. This percentage value was obtained after testing from 50% to 100% for the best precision score (Additional file 2: Table S10). Finally, only *pto* candidates are selected, by using the same targeting prediction tool as in the IPB procedure.

Random forest classifier

As for DT, proteins with a predicted transmembrane helix after its N-terminal part are removed. The RF classifier uses the properties also used in the DT procedure: the median of the length of the detected repeats, the number of linkers predicted by ARD2 (see above), the median of the distance between two linkers and the probability for an amino acid to be a linker. The proportion of residues in the detected repeats is considered rather than the number of repeats in order to have a continuous distribution of values. In addition, the RF model uses the proportion of residues with disorder propensity predicted by IUPred3 [87] with score > 0.5 and default parameters, the 20 amino acid frequencies and 36 Auto-Cross Correlation (ACC) values of neighboring amino acids over a window of 4 residues computed for each protein based on the Z-scales amino acid descriptors Hellberg et al. [31] as described in [32]. All steps were performed with the scikit-learn Python package (version 0.21.2). The training set was divided in a validation set (10%) never used for learning or testing and the remaining 90%. We then selected the best model using the attribute best_estimator from the class GridSearchCV. We evaluated the performances of the best model using the *predict* function on the validation set and it gave an accuracy of 98,61%. The best hyperparameters were 500 estimators, a maximum tree depth of 10, a minimum number of 5 samples required to be at a leaf and all other parameters at default values. The confusion matrix comparing the expected classification results with those obtained from the model prediction (see Additional file 1: Figure S6) shows that only one protein out of 72 was misclassified. The RF classification was performed with the *predict* function on

the best model. We used the mean decrease in impurity procedure to determine the importance of each feature in the model. Finally, only *pto* candidates are selected as in IPB and DT.

Training set composition

The training set for DT and RF procedures comprises 426 known α-solenoids (positive set) and 286 known non α–solenoids (negative set) manually selected based on their 3D structure, resolved experimentally or predicted by AF2. The distributions of the properties used by DT and RF for the training set are given in Additional file 1: Figure S7. Note that in the negative controls, proteins that adopt a solenoid shape made of β-sheets and α–helices were included in an attempt to increase specificity towards "all-α" solenoids.

Motifs analysis

Multiple alignments of motifs were obtained by MAFFT v7.450. Positions with more than 80% of gap residues were filtered and used to create the logo with the standalone version of WebLogo 3.7.9 [88].

Clustering

Clustering was performed with MCL v.14–137 [81] based on the $-log(E\text{–}value)$ of the BLASTP hits. For the determination of the Inflation parameter in IPB, see IPB section above. For family reconstruction, tests using Inflation parameter I value varying from 20 to 80. I = 50 was selected as it is the one were DTRF and OPR families were stable. For the final clustering of OPR and PPR motifs, the Inflation parameter was set to 20 and the clustering was performed on the identity percentage rather than the E-value, on hits sharing similarity over at least 90% of their length. Network visualization was performed with Cytoscape (v3.9.1) [89].

Statistical analysis

Statistical tests and PCA were performed with the scikit-learn Python package version 0.21.2.

Abbreviations

ACC	Auto-Cross Correlation
AF2pred	Shape predicted by AlphaFold2
$cluster_{motifs}$	Motifs clustered according to their similarity
DT	Decision-tree procedure
DTRF candidates	Candidates detected by both DT and RF procedures
IPB	Iterative profile-based procedure
OTAF	Organelle trans-acting factor
OPR	Pentatricopeptide repeat protein
pto	Predicted to be targeted to organelles
PPR	Pentatricopeptide repeat protein
PPR UWA	PPR proteins listed in the PPR database at University of Western Australia with a UniProtKB identifier in the *A. thaliana* proteome
$repeats_{RADAR}$	Repeats predicted by RADAR
RF	Random forest procedure

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12754-3.

Supplementary Material 1: Figure S1. Importance of the variables in the RF classifier. Figure S2. A. Number of OPR/PPR motifs detected by IPB and RADAR repeats in IPB and DTRF candidates found in C. reinhardtii (left) and A. thaliana (right). B. Predicted Aligned Error plot of Fig. 3B. Approximate positions of the predicted OPR and PPR domains, and of the RAP domain are indicated as in Fig. 3 D. Figure S3. Similarity network of the 1551 IPB and pto DTRF candidates from the representative proteomes used in Fig. 3A. Figure S4. A. Similarity network of the 6058 motifs (813 OPR and 5245 PPR) found in the twelve representative proteomes. Figure S5. Correlation circles of the PCA of Fig. 8. Figure S6. Confusion matrix of the RF procedure on the validation set. Figure S6. Barplot of the properties used in the DT and RF procedures within the positive (grey) and negative (white) proteins used in the training set.

Supplementary Material 2: Table S1. List of proteomes used in this study. Table S2. OPR and PPR locus from which motifs were used to build the initial profiles for the IPB procedure, and associated references. Table S3. (a) List of the OPR/PPR from (Hillebrand et al. Nucleic Acids Research. 2018) according to the annotation Chlre4 of Chlamydomonas reinhardtii that were retrieved by IPB with the OPR motifs in this study performed with the annotation v5.6. (b) List of MTERF proteins in A. thaliana detected by DT and RF. Table S4. List of IPB candidates with at least two OPR/PPR repeats and pto DTRF candidates. Table S5. 3D Classification of the pto DTRF candidates. Table S6. InterProScan annotation of IPB candidates with at least two OPR/PPR repeats and pto DTRF candidates. Tables S7. Coordinates of the OPR and PPR motifs detected by IPB, and of the PFAM domains detected by InterProScan in 3 candidates identified in this study and the 2 orthologs found in other proteomes according to the Uniref50 clustering. Table S8. Phylostratigraphy analysis of candidate protein families with varying MCL Inflation value. Table S9. Phylostratigraphy analysis of OPR/PPR motifs clusters. Table S10. Parameterization of the DT procedure.

Acknowledgements

We thank Hédi Soula for fruitful discussions and advice for setting up the machine learning procedure; Richard Lavery for initial discussions on protein-RNA affinity; Yves Choquet for fruitful discussions on OPR; Olivier Vallon for providing access to the unreleased OPRdb database and for stimulating discussions with C.C.

Authors' contributions

CC designed and performed the experiments and wrote the first version of the manuscript. IL obtained funding, coordinated the project, conceived and designed experiments, performed the experiments, analyzed data and wrote the manuscript; RG and EC contributed to the creation of the detection pipelines and performed preliminary experiments; AA performed preliminary experiments; CHR obtained funding and contributed with valuable discussions; FAW conceived experiments, contributed with valuable discussions and to the writing of the manuscript. All authors have read and approved the final version of the manuscript.

Funding

This work was supported by funding from the Centre National de la Recherche Scientifique and Sorbonne University to UMR7141; the "LabEx Dynamo" (ANR-LABX-011), the DECRYPTOR grant (CNRS 80prime 2019) and the DecryProtARN grant (CNRS Infiniti 2018).

Data availability

All data are available in the supplementary files and scripts are available at https://github.com/UMR7141/AsolFind.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 October 2025 / Accepted: 11 March 2026
Published online: 20 March 2026

References

1. Archibald JM. Endosymbiosis and Eukaryotic Cell Evolution. Curr Biol. 2015;25(19):R911–21.
2. Rochaix JD. Post-Transcriptional steps in the expression of chloroplast genes. Ann Rev Cell Dev Biol. 1992;8:1–28.
3. Fox TD. Translational control of endogenous and recoded nuclear genes in yeast mitochondria: Regulation and membrane targeting. Experientia. 1996;52(12):1130–5.
4. Choquet Y, Wollman FA. Translational regulations as specific traits of chloroplast gene expression. FEBS Lett. 2002;529(1):39–42.
5. Small ID, Peeters N. The PPR motif – a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci. 2000;25(2):45–7.
6. Eberhard S, Loiselay C, Drapier D, Bujaldon S, Girard-Bascou J, Kuras R, et al. Dual functions of the nucleus-encoded factor TDA1 in trapping and translation activation of atpA transcripts in Chlamydomonas reinhardtii chloroplasts. Plant J. 2011;67(6):1055–66.
7. Barkan A, Small I. Pentatricopeptide repeat proteins in plants. Annu Rev Plant Biol. 2014;65(1):415–42.
8. Hammani K, Bonnard G, Bouchoucha A, Gobert A, Pinker F, Salinas T, et al. Helical repeats modular proteins are major players for organelle gene expression. Biochimie. 2014;100:141–50.
9. Andrade MA, Bork P. HEAT repeats in the Huntington's disease protein. Nat Genet. 1995;11(2):115–6.
10. Riggleman B, Wieschaus E, Schedl P. Molecular analysis of the armadillo locus: uniformly distributed transcripts and a protein with novel internal repeats are associated with a Drosophila segment polarity gene. Genes Dev. 1989;3(1):96–113.
11. Spassov DS, Jurecic R. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? IUBMB Life. 2003;55(7):359–66.
12. Mosavi LK, Cammett TJ, Desrosiers DC, Zyu P. The ankyrin repeat as molecular architecture for protein recognition. Protein Sci. 2004;13(6):1435–48.
13. Cheng S, Gutmann B, Zhong X, Ye Y, Fisher MF, Bai F, et al. Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. Plant J. 2016;85(4):532–47.
14. Sikorski RS, Boguski MS, Goebl M, Hieter P. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. Cell. 1990;60(2):307–17.
15. D'Andrea LD, Regan L. TPR proteins: the versatile helix. Trends Biochem Sci. 2003;28(12):655–62.
16. Preker PJ, Keller W. The HAT helix, a repetitive motif implicated in RNA processing. Trends Biochem Sci. 1998;23(1):15–6.
17. Gutmann B, Royan S, Schallenberg-Rüdinger M, Lenz H, Castleden IR, McDowell R, et al. The Expansion and diversification of pentatricopeptide repeat RNA-editing factors in plants. Mol Plant. 2020;13(2):215–30.
18. Tourasse NJ, Choquet Y, Vallon O. PPR proteins of green algae. RNA Biology 10;1526-1542
19. Rahire M, Laroche F, Cerutti L, Rochaix JD. Identification of an OPR protein involved in the translation initiation of the PsaB subunit of photosystem I. Plant J. 2012;72(4):652–61.
20. Merendino L, Perron K, Rahire M, Howald I, Rochaix JD, Goldschmidt-Clermont M. A novel multifunctional factor involved in trans -splicing of chloroplast introns in Chlamydomonas. Nucleic Acids Res. 2006;34(1):262–74.
21. Marx C, Wünsch C, Kück U. The octatricopeptide repeat (OPR) protein Raa8 is required for chloroplast trans-splicing. Eukaryot Cell. 2015;20: EC.00096 – 15.
22. Lefebvre-Legendre L, Choquet Y, Kuras R, Loubéry S, Douchi D, Goldschmidt-Clermont M. A nucleus-encoded chloroplast protein regulated by iron availability governs expression of the photosystem I subunit PsaA in chlamydomonas reinhardtii. Plant Physiol. 2015;167(4):1527.
23. Wang F, Johnson X, Cavaiuolo M, Bohne AV, Nickelsen J, Vallon O. Two chlamydomonas OPR proteins stabilize chloroplast mRNAs encoding small subunits of photosystem II and cytochrome b6 f. Plant J. 2015;82(5):861–73.
24. Manna S. An overview of pentatricopeptide repeat proteins and their applications. Biochimie. 2015;113:93–9.
25. Hayes ML, Santibanez PI. A plant pentatricopeptide repeat protein with a DYW-deaminase domain is sufficient for catalyzing C-to-U RNA editing in vitro. J Biol Chem. 2020;295(11):3497–505.
26. Boehm E, Zaganelli S, Maundrell K, Jourdain AA, Thore S, Martinou JC. FASTKD1 and FASTKD4 have opposite effects on expression of specific mitochondrial RNAs, depending upon their endonuclease-like RAP domain. Nucleic Acids Res. 2017;45(10):6135–46.
27. Boulouis A, Drapier D, Razafimanantsoa H, Wostrikoff K, Tourasse NJ, Pascal K, et al. Spontaneous dominant mutations in chlamydomonas highlight ongoing evolution by gene diversification. Plant Cell. 2015;27(4):984–1001.
28. Karpenahalli MR, Lupas AN, Söding J. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. BMC Bioinformatics. 2007;8(1):1–8.
29. Lipinski KA, Puchta O, Surendranath V, Kudla M, Golik P. Revisiting the yeast PPR proteins—application of an iterative hidden Markov model algorithm reveals new members of the rapidly evolving family. Mol Biol Evol. 2011;28(10):2935–48.
30. Fournier D, Palidwor GA, Shcherbinin S, Szengel A, Schaefer MH, Perez-Iratxeta C, et al. Functional and genomic analyses of alpha-solenoid proteins. PLoS ONE. 2013;8(11):e79894.
31. Hellberg S, Sjoestroem M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem. 1987;30(7):1126–35.
32. Garrido C, Caspari OD, Choquet Y, Wollman FA, Lafontaine I. Evidence supporting an antimicrobial origin of targeting peptides to endosymbiotic organelles. Cells. 2020;9(8):1795.
33. Kim MK, Kang YK. Positional preference of proline in alpha-helices. Protein Sci. 1999;8(7):1492–9.
34. Rivier C, Goldschmidt-Clermont M, Rochaix J. Identification of an RNA–protein complex involved in chloroplast group II intron trans-splicing in Chlamydomonas reinhardtii. EMBO J. 2001;20(7):1765–73.
35. Hillebrand A, Matz JM, Almendinger M, Müller K, Matuschewski K, Schmitz-Linneweber C. Identification of clustered organellar short (cos) RNAs and of a conserved family of organellar RNA-binding proteins, the heptatricopeptide repeat proteins, in the malaria parasite. Nucleic Acids Res. 2018;46(19):10417–31.
36. Cattelin C. Exploration de la diversité des protéines à solénoïdes alpha, régulatrices de l'expression des gènes des organites dans les lignées eucaryotes photosynthétiques et étude de la dynamique conformationnelle des protéines à PentatricoPeptide Repeats [phdthesis]. Sorbonne Université; 2023. Available from: https://theses.hal.science/tel-04164696
37. Lefebvre-Legendre L, Reifschneider O, Kollipara L, Sickmann A, Wolters D, Kück U, Goldschmidt-Clermont M. 2016. A pioneer protein is part of a large complex involved in trans-splicing of a group II intron in the chloroplast of Chlamydomonas reinhardtii. Plant J 85: 57–69.
38. Kück U, Schmitt O. 2021. The Chloroplast Trans-Splicing RNA–Protein Supercomplex from the Green Alga Chlamydomonas reinhardtii. Cells 10: 290.
39. Auchincloss AH, Zerges W, Perron K, Girard-Bascou J, Rochaix JD. Characterization of Tbc2, a nucleus-encoded factor specifically required for translation of the chloroplast psbC mRNA in Chlamydomonas reinhardtii. J Cell Biol. 2002;157(6):953–62.
40. Jarrige D. Deciphering the OPR code to further assess the physiological role of OPR proteins [phdthesis]. Sorbonne Université; 2019. Available from: https://theses.hal.science/tel-03349187
41. Yin P, Li Q, Yan C, Liu Y, Liu J, Yu F, et al. Structural basis for the modular recognition of single-stranded RNA by PPR proteins. Nature. 2013;504(7478):168–71.
42. Craig RJ, Gallaher SD, Shu S, Salomé PA, Jenkins JW, Blaby-Haas CE, et al. The Chlamydomonas Genome Project, version 6: Reference assemblies for mating-type plus and minus strains reveal extensive structural mutation in the laboratory. Plant Cell. 2023;35(2):644–72.
43. Mirarab S, Warnow T. FASTSP: linear time calculation of alignment accuracy. Bioinformatics. 2011;27(23):3250–8.
44. Høie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, et al. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. Nucleic Acids Res. 2022;50:gkac439.

45. Qu K, Wei L, Yu J, Wang C. Identifying plant pentatricopeptide repeat coding gene/protein using mixed feature extraction methods. Front Plant Sci. 2019;9. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2018.01961/full. Cited 2020 Aug 21.
46. Feng C, Zou Q, Wang D. Using a low correlation high orthogonality feature set and machine learning methods to identify plant pentatricopeptide repeat coding gene/protein. Neurocomputing. 2021;424:246–54.
47. Zhao X, Wang H, Li H, Wu Y, Wang G. Identifying plant pentatricopeptide repeat proteins using a variable selection method. Front Plant Sci. 2021;12. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2021.506681. Cited 2023 Jan 9.
48. Nikov GI, Pretorius D, Murray JW. SOLeNNoID: a deep learning pipeline for solenoid residue detection in protein structures. Bioinformatics. 2025;41(8):btaf415.
49. Gallaher SD, Fitz-Gibbon ST, Strenkert D, Purvine SO, Pellegrini M, Merchant SS. High-throughput sequencing of the chloroplast and mitochondrion of Chlamydomonas reinhardtii to generate improved de novo assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. Plant J. 2018;93(3):545–65.
50. Iyer LM, Zhang D, Rogozin IB, Aravind L. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. Nucleic Acids Res. 2011;39(22):9473–97.
51. Leister D, Kleine T. Extending the Repertoire of mTERF Proteins with Functions in Organellar Gene Expression. Mol Plant. 2020;1(6):817–9.
52. Robles P, Quesada V, Robles P, Quesada V. Research progress in the molecular functions of plant mTERF proteins. Cells. 2021;10(2). Available from: https://www.mdpi.com/2073-4409/10/2/205. Cited 2026 Jan 6.
53. Small I, Melonek J, Bohne AV, Nickelsen J, Schmitz-Linneweber C. Plant organellar RNA maturation. Plant Cell. 2023;35(6):1727–51.
54. Zhang H, Scheirer DC, Fowle WH, Goodman HM. Expression of antisense or sense RNA of an ankyrin repeat-containing gene blocks chloroplast differentiation in arabidopsis. Plant Cell. 1992;4(12):1575–88.
55. Melonek J, Matros A, Trösch M, Mock HP, Krupinska K. The core of chloroplast nucleoids contains architectural SWIB domain proteins. Plant Cell. 2012;24(7):3060–73.
56. Kamaliyan Z, Clarke TL. Zinc finger proteins: guardians of genome stability. Front Cell Dev Biol. 2024;12:1448789.
57. Méteignier LV, Ghandour R, Meierhoff K, Zimmerman A, Chicher J, Baumberger N, et al. The Arabidopsis mTERF-repeat MDA1 protein plays a dual function in transcription and stabilization of specific chloroplast transcripts within the psbE and ndhH operons. New Phytol. 2020;227(5):1376–91.
58. Méteignier LV, Ghandour R, Zimmerman A, Kuhn L, Meurer J, Zoschke R, et al. Arabidopsis mTERF9 protein promotes chloroplast ribosomal assembly and translation by establishing ribonucleoprotein interactions in vivo. Nucleic Acids Res. 2021;49(2):1114–32.
59. Lee K, Leister D, Kleine T, Lee K, Leister D, Kleine T. Arabidopsis mitochondrial transcription termination factor mTERF2 promotes splicing of group IIB introns. Cells. 2021;10(2). Available from: https://www.mdpi.com/. Cited 2026 Jan 12.
60. Cavaiuolo M, Kuras R, Wollman F, Choquet Y, Vallon O. Small RNA profiling in Chlamydomonas: insights into chloroplast RNA metabolism. Nucleic Acids Res. 2017;45(18):10783–99.
61. Pfalz J, Bayraktar OA, Prikryl J, Barkan A. Site-specific binding of a PPR protein defines and stabilizes 5′ and 3′ mRNA termini in chloroplasts. EMBO J. 2009;28(14):2042–52.
62. Kijima S, Kitao A. RNA binding mechanism of the FUS Zinc finger in concert with its flanking intrinsically disordered region. J Chem Inf Model. 2025;65(15):8262–75.
63. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–54.
64. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
65. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
66. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res. 2004;32(Web Server issue):W500–502.
67. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33: 3387–3395.
68. Sperschneider J, Catanzariti AM, DeBoer K, Petre B, Gardiner DM, Singh KB, et al. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci Rep. 2017;7(1):44598.
69. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2(4):953–71.
70. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35(Web Server issue):W585–587.
71. Nakamura Y, Sasaki N, Kobayashi M, Ojima N, Yasuike M, Shigenobu Y, et al. The first symbiont-free genome sequence of marine red alga, susabi-nori (Pyropia yezoensis). PLoS ONE. 2013;8(3):e57122.
72. Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci. 2014;111(45):E4859–68.
73. Lemieux C, Vincent AT, Labarre A, Otis C, Turmel M. Chloroplast phylogenomic analysis of chlorophyte green algae identifies a novel lineage sister to the Sphaeropleales (Chlorophyceae). BMC Evol Biol. 2015;15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4665822/. Cited 2017 Nov 27.
74. Leliaert F, Tronholm A, Lemieux C, Turmel M, DePriest MS, Bhattacharya D, et al. Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. Sci Rep. 2016;6:25367.
75. Liu Y, Johnson MG, Cox CJ, Medina R, Devos N, Vanderpoorten A, et al. Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. Nat Commun. 2019;10(1):1485.
76. Nakada T, Tsuchida Y, Tomita M. Improved taxon sampling and multigene phylogeny of unicellular chlamydomonads closely related to the colonial volvocalean lineage Tetrabaenaceae-Goniaceae-Volvocaceae (Volvocales, Chlorophyceae). Mol Phylogenet Evol. 2019;130:1–8.
77. Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, et al. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. Cell. 2019;179(5):1057–e106714.
78. Hanschen ER, Starkenburg SR. The state of algal genome quality and diversity. Algal Res. 2020;50:101968.
79. Zhao T, Zwaenepoel A, Xue JY, Kao SM, Li Z, Schranz ME, et al. Whole-genome microsynteny-based phylogeny of angiosperms. Nat Commun. 2021;12(1):3498.
80. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
81. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucl Acids Res. 2002;30:1575–84.
82. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
83. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucl Acids Res. 2013;41(12):e121–121.
84. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305(3):567–80.
85. Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. Proteins. 2000;41(2):224–37.
86. Sormanni P, Camilloni C, Fariselli P, Vendruscolo M. The s2D method: simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. J Mol Biol. 2015;427(4):982–96.
87. Erdős G, Pajkos M, Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res. 2021;49(2):W297–303.
88. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.
89. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.